CAPÍTULO VI

Interpretabilidad con redes bayesianas*

Pedro Larrañaga

La creciente difusión y popularidad de los sistemas inteligentes basados en aprendizaje automático plantea la disyuntiva de su uso indiscriminado frente a la necesidad de comprender su funcionamiento interno. La explicación *post hoc*, a la que habitualmente recurren los modelos de caja negra, no resulta suficiente en el ámbito científico ni tampoco en situaciones de alto riesgo. En tales casos se requiere que el humano sea capaz de interpretar tanto la salida del sistema como el proceso interno seguido por el mismo hasta alcanzar dicho resultado. En este capítulo analizaremos las potencialidades en interpretabilidad del paradigma conocido como red bayesiana enfatizando su transparencia y versatilidad para llevar a cabo distintos tipos de razonamiento, como por ejemplo el predictivo, diagnóstico, intercausal, contrafáctico, o abductivo. Se mostrará un caso de uso real en neurociencia computacional.

Palabras clave: interpretabilidad, redes bayesianas, neurociencia computacional.

* Agradecimientos: este trabajo ha sido financiado en parte por el Ministerio de Ciencia e Innovación de España a través de los Proyectos AEI/10.13039/501100011033-PID2022-139977NB-I00, PLEC2023-010252 y TED2021-131310B-I00, y por la Comunidad Autónoma de Madrid en el marco de la *ELLIS Unit Madrid* y del proyecto IDEA-CM (TEC-2024/COM-89), junto con el proyecto del CDTI, PLEC2023-010252.

1. INTRODUCCIÓN

La inteligencia artificial se ha convertido en una tecnología disruptiva capaz de transformar aspectos relevantes de nuestras vidas y con una gran repercusión en el desarrollo económico de las naciones. Los medios de comunicación se hacen eco a diario de avances en inteligencia artificial al tiempo que las grandes empresas tecnológicas, la mayoría de las cuales tienen sus sedes en Estados Unidos o en China, llevan años compitiendo por un mercado cada vez más numeroso y que les proporciona ganancias importantes. Los avances conseguidos en estos últimos años en el desarrollo de aplicaciones relativas al tratamiento de imágenes y al procesamiento de lenguaje natural han sido sorprendentes.

A pesar de dichos avances tecnológicos, la realidad en cuanto a su implementación y utilización deja mucho que desear. Así por ejemplo, la Agencia de Administración de Alimentos y Medicamentos del Gobierno de los Estados Unidos norteamericanos (FDA, en su acrónico en inglés) ha ido incrementando anualmente de manera exponencial el número de sistemas inteligentes médicos aprobados, hasta situarse en cifras cercanas a los 120 en estos últimos años, la mayoría de los cuales han sido en el ámbito de la radiología (Muehlematter *et al.,* 2020). Sin embargo la utilización de dichos sistemas inteligentes en el día a día hospitalario tiene una tasa de implantación muy baja. Los motivos aducidos son muy variados, si bien destaca el hecho de que los médicos se resisten a utilizar sistemas que no les explican el porqué de las decisiones aconsejadas. Esta necesidad de entender todo el proceso de razonamiento (tanto la algoritmia en la que se fundamenta la inducción del modelo como la propia cadena de pasos que conduce a una determinada salida) está ocasionando que los humanos requieran de explicaciones ante las decisiones que les proponen los sistemas inteligentes cuando estos se aplican a situaciones que pueden considerarse críticas o de especial relevancia. Dichas situaciones pueden producirse en la generación de nuevo conocimiento científico, en decisiones judiciales acerca de la concesión de libertad provisional a presos, en el diagnóstico y el pronóstico de enfermedades en medicina, en el mantenimiento predictivo de plantas industriales o en la monitorización de puentes en carreteras por las que diariamente pasan miles de coches.

Esta necesidad de explicar tanto los modelos de aprendizaje automático, como los procesos de inducción de los mismos y de los diferentes tipos de inferencia que pueden llevarse a cabo con ellos, ha dado origen a la denominada explicabilidad en inteligencia artificial (XAI) (Gunning *et al.,* 2019). Esta XAI es también recogida por la normativa de regulación sobre la inteligencia artificial (EU AI Act) aprobada recientemente por el Parlamento Europeo. EU AI Act se fundamenta en una jerarquía de riesgos, que van desde mínimo a limitado, alto o inaceptable, y que desde el segundo nivel ya requieren de una transparencia en las decisiones propuestas. Si bien el término XAI se extiende a toda la inteligencia artificial lo cierto es que la inmensa mayoría de su aplicabilidad es a modelos inducidos a partir de métodos de aprendizaje automático. Cualquiera de las tareas que se llevan a cabo dentro del mismo, es decir clasificación supervisada, regresión, *clustering,* selección de variables, aprendizaje por refuerzo u optimización heurística, por citar las más habituales, han ido desarrollando sus propios métodos de explicabilidad, siendo la clasificación supervisada la que ha recibido más atención en la literatura especializada (Molnar, 2022).

La necesidad de explicar modelos de aprendizaje automático calificados como cajas negras (por ejemplo los modelos entrenados con redes neuronales profundas) ha llevado a los investigadores a desarrollar explicaciones basadas en modelos subrrogados que sean transparentes, que si bien proporcionan aproximaciones al desempeño de los modelos originales, muchas veces se encuentran alejados en sus prestaciones (Nakka *et al.,* 2023). Además la dificultad de explicar adecuadamente dichas cajas negras ha ocasionado la existencia de una línea de investigación que desaconseja este tipo de explicaciones en situaciones que sean de vital importancia para el ser humano y que defiende el uso de modelos interpretables en su lugar (Rudin, 2019). Esta es también la aproximación propuesta en este capítulo.

La organización del capítulo es como sigue. La sección 2 describe los conceptos de interpretabilidad y explicabilidad en sistemas inteligentes recalcando sus diferencias fundamentales, enfatizando las tres propiedades que un sistema inteligente debe de cumplir para ser considerado interpretable y describiendo las ventajas de contar con dichos sistemas interpretables. La sección 3 introduce el paradigma de redes bayesianas a partir del concepto de independencia condicional entre tripletas de variables, proporcionando algoritmos de aprendizaje a partir de datos, así como distintos métodos de inferencia tanto exacta como aproximada. En la sección 4 se mostrará cómo las redes bayesianas verifican las tres propiedades necesarias para su consideración como paradigma interpretable, focalizándose a continuación en las potencialidades de interpretabilidad intrínsecas a las redes bayesianas. La sección 5 muestra un ejemplo de utilización de las redes bayesianas en la interpretación de modelos de clasificación morfológica de interneuronas GABA-érgicas. La sección 6 recoje las conclusiones del capítulo junto con propuestas para seguir avanzando en la utilización de las redes bayesianas como modelos interpretables.

2. INTERPRETABILIDAD VERSUS EXPLICABILIDAD

En 2016 la agencia norteamericana de investigación DARPA lanzó un programa donde el foco principal era el desarrollo de modelos de inteligencia artificial que fueran explicables. Los responsables de dicho programa (Gunning *et al.,* 2029) lo motivan por la necesidad de que los usuarios sean capaces de comprender la solución proporcionada por el sistema, así como todas y cada una de las fases que el mismo ha efectuado hasta alcanzar la solución. Se pretendía que el ser humano se involucrase de manera activa en las decisiones sugeridas por el sistema, llegando a aceptarlas incluso en casos en los que no coincida su juicio con el sugerido por el sistema, siempre y cuando la explicación propuesta fuese comprensible y admisible para el experto. Esto generaría la posibilidad de crear un nuevo conocimiento basado en aquellas sugerencias que no coincidiesen con el estado del arte, pero que el experto una vez analizadas y comprendido todo el proceso de decisión estuviese de acuerdo en admitirlas.

Si bien los sistemas inteligentes de caja negra, han sido (están siendo) explicados con la ayuda de otros paradigmas (subrogados) de caja blanca que son inducidos a partir del comportamiento de dichas cajas negras, y por tanto no llegan a explicarlas en realidad, esta

práctica está siendo admitida y dada por buena por una parte importante de la comunidad científica. Sin embargo, en problemas que puedan ser considerados como críticos, tanto porque afectan a aspectos relevantes de la naturaleza humana, como por su relevancia económica, judicial o científica, la necesidad de contar con sistemas inteligentes interpretables es innegable (Rudin, 2019).

Según Lipton, 2019, la interpretabilidad se refiere a la capacidad de un ser humano para comprender el funcionamiento interno del modelo. Es decir, permite que las personas comprendan cómo el modelo llega a sus conclusiones o predicciones. Dicha interpretabilidad descansa en tres propiedades que todo sistema inteligente debe de cumplir para ser considerado como interpretable.

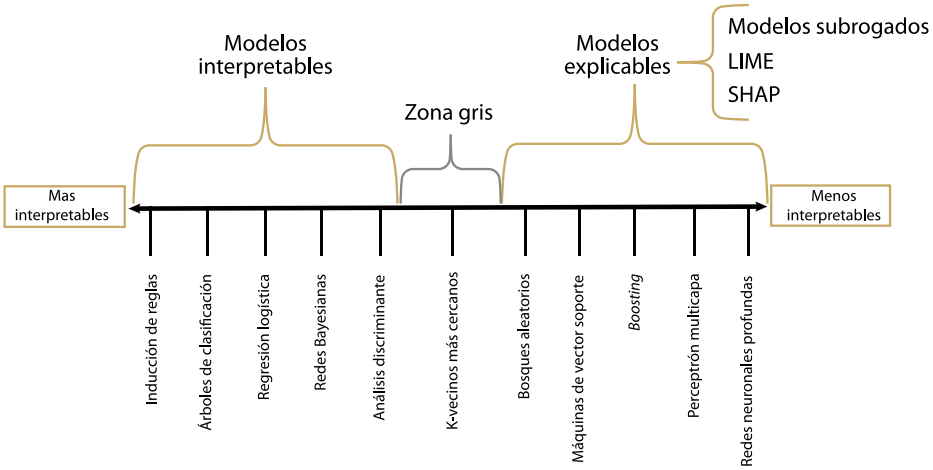
- *Simulabilidad*, la cual hace referencia a la capacidad de un modelo para ser simulado por un ser humano. No se trata solo de que sea sencillo, sino de que una persona pueda reproducir el proceso llevado a cabo por el modelo para sugerir una respuesta. Este concepto se aplica al nivel del modelo completo.
- *Descomponibilidad*, denota la capacidad de descomponer un modelo en partes, para de esta manera facilitar la comprensión del modelo global. Esto se aplica a nivel de los componentes individuales del modelo.
- *Transparencia algorítmica*, expresa la capacidad de comprender el procedimiento que el modelo sigue para generar su salida. Esto se aplica a nivel del algoritmo de entrenamiento del modelo, permitiendo entender cómo el modelo aprende y procesa los datos, así como a nivel del algoritmo de inferencia para entender cómo se ha llegado a la decisión.

La interpretabilidad del modelo nos va a permitir que el humano sea el centro de las decisiones y participe de manera activa en las posibles mejoras del sistema. Por otra parte, un modelo interpretable va a posibilitar la justificación de las decisiones sugeridas por el mismo, la mejora de las prestaciones a nivel cuantitativo, la compresión de sus debilidades, el descubrimiento de nuevo conocimiento, el testeo de su robustez ante situaciones extremas, el análisis de la existencia de sesgos que puedan mediatizar su comportamiento, y la transferencia de los modelos inducidos a problemas en otros dominios de características similares.

La figura 1 muestra una ordenación de algunos tipos de modelos de clasificación supervisada. Los interpretables (inducción de reglas, árboles de clasificación, regresión logística, redes bayesianas, análisis discriminante) se posicionan a la izquierda, mientras que los k-vecinos más cercanos se encuentran en una posición intermedia. En la parte de la derecha nos encontramos con paradigmas que sin ser interpretables pueden ser explicados. Los bosques aleatorios, las máquinas de vector soporte, el *boosting*, el perceptrón multicapa y las redes neuronales profundas forman parte de este grupo. La explicabilidad en estos paradigmas se lleva a cabo usando distintos modelos subrogados, así como por medio de técnicas como LIME, generando aproximaciones locales a las predicciones del modelo, o SHAP que busca calcular la contribución de cada variable predictora al modelo.

Figura 1.

Paradigmas de clasificación supervisada en función de su interpretabilidad



Fuente: Elaboración propia.

Una crítica habitual sobre los modelos interpretables es que sus prestaciones cuantitativas en cuanto a la clasificación no llegan a ser competitivas al compararlas con las que proporcionan los modelos explicativos. Este asunto ha quedado rebatido por varios trabajos empíricos (Lukas-Valentin *et al.,* 2023; Shueli, 2010).

Las redes bayesianas al compararlas con los otros paradigmas interpretables de la figura 1 destacan por su gran versatilidad en cuanto a los diferentes tipos de interpretabilidad que pueden efectuarse con las mismas, algunos exclusivos de este paradigma (véase sección 4).

3. REDES BAYESIANAS

Las redes bayesianas representan explícitamente el conocimiento incierto inherente a un dominio por medio de una representación gráfica que resulta ser muy intuitiva, facilitando la interpretabilidad del modelo. Son capaces de acomodar tanto variables discretas como continuas, e incluso variables que varían en el tiempo. Son también aplicables a situaciones con datos *missing*. El modelo de red bayesiana puede ser obtenido con la ayuda de un experto, automáticamente a partir de datos, o combinando ambas aproximaciones. La aproximación automática resulta de especial interés en la actualidad donde la accesibilidad a grandes bases de datos es cada vez más sencilla, y donde la modelización basada en el experto puede acarrear un gran coste en tiempo a la vez que está sometida a errores derivados del gran número de variables a tratar al unísono.

Una gran ventaja de las redes bayesianas en comparación con otros paradigmas del aprendizaje automático es que con ellas se pueden llevar a cabo la gran mayoría de las tareas habituales en aprendizaje automático: clasificación supervisada tanto en la predicción de una única variable clase (Varando *et al.,* 2015), como cuando nuestro interés radica en predecir varias variables clase a la vez (Bielza y Larrañaga, 2011); regresión de una única variable a predecir como de varias variables (Borchani *et al.,* 2015); clustering probabilista con sus diferentes variantes de multivista o multipartición (Rodriguez-Sanchez *et al.,* 2021); descubrimiento de asociaciones entre variables (Larrañaga *et al.,* 1996); aprendizaje por refuerzo (Valverde *et al.,* 2023); detección de anomalías (Puerto-Santana *et al.,* 2022); selección de variables (Inza *et al.,* 2000) u optimización heurística (Larrañaga y Bielza, 2024).

3.1. Elementos básicos

Supongamos que partimos de un conjunto de datos con N observaciones y n variables, denotadas como $X_1,...,X_n$. Sea $\mathcal{D}=\{\boldsymbol{x}^1,...,\boldsymbol{x}^N\}$ el conjunto de datos, donde $\boldsymbol{x}^h=\left(x_1^h,...,x_n^h\right)$, $h=1,...,N$, $\boldsymbol{X}=(X_1,...,X_n)$ y $x_i\in\Omega_{X_i}=\{1,2,...,R_i\}, i=1,...,n$.

Una *red bayesiana* (Pearl, 1988; Koller y Friedman, 2009) es una representación de una distribución de probabilidad conjunta (JPD) $p\left(X_1,...,X_n\right)$ sobre un conjunto de variables aleatorias discretas, $X_1,...,X_n$. La *regla de la cadena* permite escribir la JPD como

$$p\left(X_1,...,X_n\right)=p(X_1)p(X_2\mid X_1)p(X_3\mid X_1,X_2)...p(X_n\mid X_1,...,X_{n-1}) \qquad [1]$$

La JPD contiene toda la información probabilística sobre el dominio y puede ser usada para responder a cualquier tipo de pregunta en términos probabilistas. El problema de usar la JPD es que su tamaño crece exponencialmente con el número de variables *n*. Las redes bayesianas evitan la necesidad de contar con un número exponencial de parámetros gracias a la consideración del concepto de independencia condicional que involucra a tripletas de variables.

Dos variables aleatorias X e Y son *condicionalmente independentes* (c.i.) dada una tercera variable aleatoria Z si $p(x,y\mid z)=p(x\mid z)\ \ \forall x,y,z$. Es decir, para cualquier $Z=z$, la información $Y=y$ no influencia la probabilidad de $X=x$. La definición puede escribirse de manera alternativa como $p(x,y\mid z)=p(x\mid z)p(y\mid z)\ \ \forall x,y,z$, y también puede extenderse al caso en que X, Y, Z sean vectores disjuntos de variables aleatorias.

El concepto de independencia condicional va a permitir reducir el número de parámetros necesarios para especificar la JPD. Supongamos que para cada X_i somos capaces de encontrar un subconjunto de variables $\boldsymbol{Pa}(X_i)\subseteq\{X_1,...,X_{i-1}\}$ tal que dado $Pa(Xi)$, Xi es c.i. de todas las variables en $\{X_1,...,X_{i-1}\}\setminus\boldsymbol{Pa}(X_i)$, *i. e.,* $p(X_i\mid X_1,...,X_{i-1})=p(X_i\mid\boldsymbol{Pa}(X_i))$. La JPD de la ecuación [1] es ahora:

$$p\left(X_1,...,X_n\right)=p(X_i\mid\boldsymbol{Pa}(X_i))...p(X_n\mid\boldsymbol{Pa}(X_n)) \qquad [2]$$

en la que es de esperar que se reduzca sustancialmente el número de parámetros. Esta modularidad va a facilitar el mantenimiento del modelo, así como el desarrollo de procedimientos de razonamiento eficientes que posibiliten su interpretabilidad.

Una red bayesiana representa esta factorización de la JPD con un grafo acíclico dirigido (DAG), el cual constituye la estructura de la red bayesiana. Un grafo G se representa por el par (V,E), donde V es el conjunto de nodos y E es el conjunto de *arcos* entre los nodos en V. Los nodos del DAG representan las variables aleatorias del dominio, $X_1,...,X_n$. Los arcos representan las dependencias probabilísticas entre las variables. Dichas dependencias se cuantifican por medio de distribuciones de probabilidad condicionadas. Los *padres* de un nodo X_i, $\boldsymbol{Pa}(X_i)$, X_i, están constituidos por todos los nodos que apuntan a X_i. Diremos que X_i es un nodo *hijo* de dichos nodos padres. El término *acíclico* significa que el grafo no contiene ciclos, es decir, que no existe una secuencia de nodos que siguiendo la dirección de los arcos comience y termine en el mismo nodo.

Por tanto, la red bayesiana tiene dos componentes: un DAG y un conjunto de distribuciones de probabilidad condicionales asociadas a cada nodo X_i dados sus padres en el DAG, $p(X_i \mid \boldsymbol{Pa}(X_i))$, que determinan una única JPD tal y como se muestra en la ecuación [2]. El primer componente cualitativo se refiere a la *estructura de la red bayesiana*, mientras que el segundo componente cuantitativo hace referencia a los *parámetros de la red bayesiana*. Cuando todas las variables son discretas estos parámetros están tabulados y constituyen la *tabla de probabilidad condicional* (CPT).

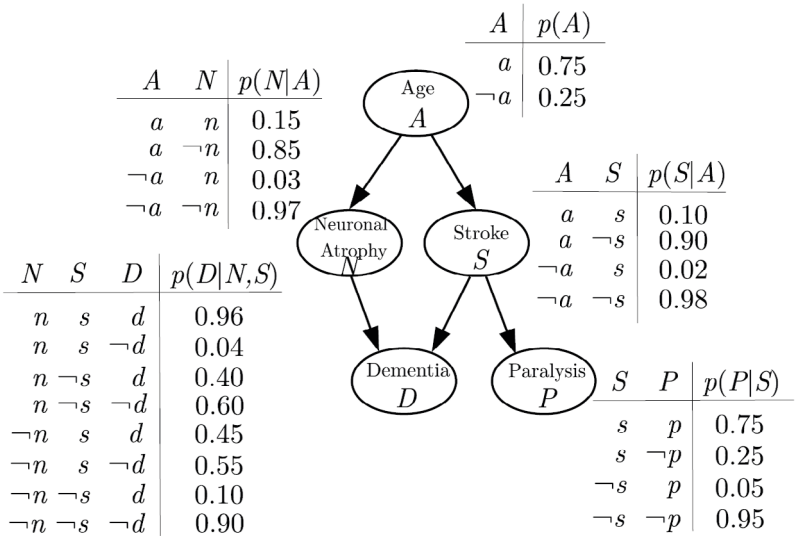
Ejemplo. Riesgo de demencia

La figura 2 muestra un ejemplo hipotético de una red bayesiana modelando el riesgo de demencia. Todas las variables son binarias, con *x* denotando la "presencia" y ¬*x* la "ausencia", para `Dementia` *D*, `Neuronal Atrophy` *N*, `Stroke` *S* y `Paralysis` *P*. Para `Age` *A*, *a* significa "edad +65", en caso contrario el estado es ¬*a*. De la estructura de la red bayesiana se desprende que tanto `Stroke` como `Neuronal Atrophy` están influenciados por `Age` (padre de ambos en el DAG). Ambas variables tienen influencia en `Dementia` (su hijo). `Paralysis` está directamente asociada con `Stroke`. Las CPT indican las probabilidades condicionales asociadas a cada nodo. Por ejemplo, un paciente que tiene "atrofia neuronal" y ha sufrido un "accidente cerebro vascular" tiene una probabilidad de 0,96 de tener "demencia": $p(d \mid n,s) = 0,96$. Sin embargo, en ausencia de "atrofia neuronal" y "accidente cerebro vascular", esta probabilidad es sólo de 0,10, *i. e.*, $p(d \mid \neg n,\neg s) = 0.10$.

La JPD factoriza como $p(A,N,S,D,p) = p(A)p(N \mid A)p(S \mid A)p(D \mid N,S)p(P \mid S)$, requiriendo 11 probabilidades en lugar de los 2^5−1=31 parámetros necesarios en el caso de no usar la red bayesiana.

Figura 2.

Ejemplo hipotético de una red bayesiana modelando el riesgo de demencia



Fuente: Elaboración propia.

La propiedad de independencia condicional entre tripletas de variables puede ser chequeada gráficamete a partir del criterio de *u-separation* (Lauritzen *et al.,* 1990). El criterio para chequear si ***X*** e ***Y*** están u-separados por ***Z*** requiere de tres pasos. En el primer paso se obtiene el menor subgrafo que contiene a ***X***, ***Y*** y ***Z*** y a sus ancestros. En el segundo paso se moraliza el subgrafo resultante, *i. e.,* se añade una arista entre padres con hijos en común y, a continuación, se transforman los arcos en aristas. En el tercer paso, diremos que ***Z*** u-separa ***X*** e ***Y*** si cualquier camino que se pueda establecer desde una variable en ***X*** a una variable en ***Y*** contiene una variable en ***Z***.

3.2. Inferencia

Las redes bayesianas además de proporcionarnos visualizaciones de las relaciones entre las variables del dominio considerado y de darnos la posibilidad de chequear gráficamente independencias condicionales entre tripletas de variables, son modelos muy útiles para llevar a cabo predicciones, diagnósticos y explicaciones. Esto se efectúa por medio del cálculo de la distribución de probabilidad de una variable (o de un conjunto de variables) de interés dados los valores de otras variables. Las variables observadas se denominan *evidencia observada* ***E=e***. A tales efectos existen tres tipos de variables en ***X***: una variable sobre la que se efectúa la pregunta, X_i, (habitualmente una única variable, aunque un vector de variables es también posible), las variables evidencia ***E*** y el resto de variables, que son variables no observadas ***U***.

El término *inferencia* se refiere a encontrar la probabilidad de una variable X_i (o de un vector de variables) condicionado a ***e***, *i. e.,* $p(X_i \mid \boldsymbol{e})$. La teoría de la probabilidad nos proporciona herramientas para computar $p(X_i \mid \boldsymbol{e})$. Esta computación se denomina *razonamiento probabilista* bajo incertidumbre, donde la evidencia se propaga a través de la estructura probabilística actualizando el resto de probabilidades. Si no hay evidencia, las probabilidades de interés son distribuciones *a priori*, $p(X_i)$. La inferencia en redes bayesianas puede combinar evidencia existente en cualquier parte de la red y responder a diferentes tipos de preguntas. Bajo causalidad, podemos predecir los efectos dadas las causas (*razonamiento predictivo*), diagnosticar las causas dados los efectos (*razonamiento diagnóstico*) o explicar una causa como responsable de un efecto (*razonamiento intercausal*). Dicho razonamiento intercausal es único en redes bayesianas. En la estructura $C_1 \rightarrow X \leftarrow C_2, C_1 \; y \; C_2$ son independientes, pero una vez que su hijo en común es observado se convierten en dependientes. Además, cuando el efecto *X* es conocido, la presencia de una causa explicativa hace que la causa alternativa sea menos probable.

Inferencia tambien se refiere a encontrar los valores de un conjunto de variables que mejor explican la evidencia observada. Esto se denomina *inferencia abductiva*. En *abducción total*, se busca el $\arg \max_{\mathbf{U}} p(\mathbf{U}|\mathbf{e})$, *i. e.*, el objetivo es encontrar la *explicación más probable* (MPE), mientras que en la *abducción parcial* el objetivo es similar, en este caso referido a un subconjunto de variables de ***U*** (el conjunto de explicación), y se conoce también bajo el nombre de *máximo a posteriori parcial* (MAP).

Ejemplo. Riesgo de demencia (diferentes tipos de razonamiento probabilístico)

El primer tipo de probabilidades que podemos examinar se refieren a las distribuciones *a priori* de cada variable univariante $p(X_i)$, *i. e.*, sin haber observado ningún tipo de evidencia. La figura 3(a) muestra estas probabilidades usando GeNIe[1] . Por ejemplo, la probabilidad de padecer "demencia" es 0,17.

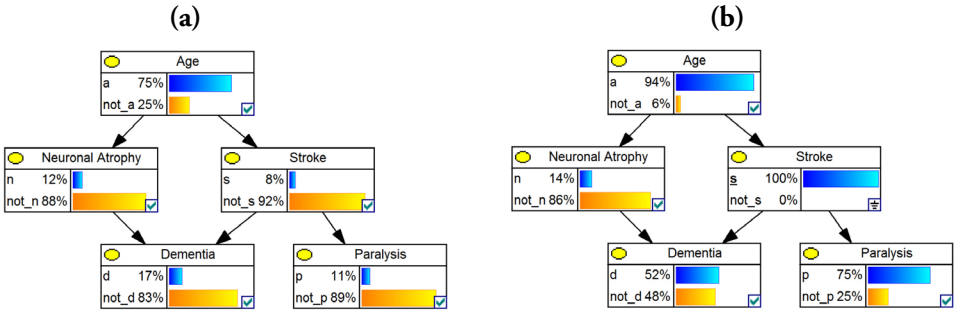
Supongamos que tenemos un paciente que ha sufrido un "accidente cerebro vascular" (*E*=*S*=*s*), es decir el estado *s* de la variable `Stroke` está fijado con la barra al 100 %. Las probabilidades actualizadas dada esta evidencia, *i. e.,* $p(x_i \mid s)$ para los nodos *A,N,D* o *P*, se muestran en la figura 3(b). La probabilidad de sufrir "demencia" se ha incrementado a $p(d \mid s) = 0,52$. Para el estado opuesto –no tener "accidente cerebro vascular"– se obtiene $p(d \mid \neg s) = 0,14$ (no mostrado). Este es un ejemplo de razonamiento predictivo.

Como un ejemplo de razonamiento diagnóstico, dado el efecto de "parálisis" (*P*=*p*), la probabilidad de que la causa sea un "accidente cerebro vascular" es alta, $p(s \mid p) = 0,57$. Para llevar a cabo el razonamiento intercausal, podemos considerar *N, S* y *D*. `Neuronal atrophy` (*N*) y `Stroke` (*S*) son independientes y tenemos $p(n) = 0,12, p(s) = 0,08$ (figura 3(a)), pero una vez que `Dementia` (*D*) es observada, *e. g., D*=*d*, las probabilidades de

[1] https://www.bayesfusion.com/

Figura 3.

(a) Las distribuciones *a priori* p(X_i) se muestran como gráficos de barras, para cada nodo X_i. (b) Después de observar que el paciente ha sufrido un "accidente cerebro vascular" (S=s), las distribuciones se actualizan como p($X_i|s$).



las posibles causas, *n* y *s*, aumentan: $p(n \mid d) = 0,33$ y $p(s \mid d) = 0,25$. La presencia de "atrofia neuronal" (N=n), explicaría la "demencia" observada *d*, lo cual hace decrecer la probabilidad de que `Stroke` sea la causa, *i. e.*, $p(s \mid d, n) = 0,20 < p(s \mid d) = 0,25$. Por otra parte, la presencia de un "accidente cerebro vascular" hace que la "atrofia neuronal" sea menos probable, $p(n \mid d, s) = 0,26 < p(n \mid d) = 0,33$.

Además, podemos encontrar la explicación mas probable para un paciente con parálisis, si resolvemos $\arg\max_{\{A,N,S,D\}} p(A,N,S,D \mid p)$. Obtenemos ($a$,¬$n$,$s$,¬$d$), con probabilidad 0,25. Es decir, la parálisis se explica por: tener una "edad de +65 años", haber tenido un "accidente cerebro vascular", y no tener ni "atrofia neuronal" ni "demencia". Las otras posibles configuraciones son menos probables. Finalmente, en la abducción parcial buscaríamos la explicación (reducida) del conjunto de variables (A, S) dada la misma evidencia de parálisis. En otras palabras tratamos de resolver el problema $\arg\max_{\{A,S\}} p(A,S \mid p)$ y encontramos la configuración (a, s), con probabilidad 0,53. Esto nos dice que "parálisis" está (parcialmente) explicada por tener una "edad de +65 años" y haber sufrido un "accidente cerebro vascular".

3.2.1. Inferencia exacta

La *inferencia exacta* en redes bayesianas es un problema NP-difícil[2] (Cooper, 1990). En la práctica muchas de las inferencias exactas que se llevan a cabo no se relacionan con el peor

[2] La clase de problemas NP-difíciles (*NP-hard*) es un concepto fundamental en teoría de la complejidad computacional y describe problemas que son al menos tan difíciles como los problemas de la clase NP (*Nondeterministic Polynomial time*), problemas cuya solución puede ser verificada en tiempo polinómico por una máquina determinista.

caso y se puede dar una respuesta no muy costosa computacionalmente usando los algoritmos que se muestran en lo que sigue.

Existen varios algoritmos exactos de propagación de la evidencia en redes bayesianas. La *aproximación por fuerza bruta* usa la factorización proporcionada por la red bayesiana para obtener la JPD y posteriormente por marginalización da respuesta a las preguntas de inferencia, siendo sin embargo, incapaz de organizar eficientemente las operaciones necesarias para llevar a cabo dicha inferencia, aspecto en el que sí que se focaliza el *algoritmo de eliminación de variables*. Este algoritmo se ilustra en el ejemplo que aparece, a continuación. El *algoritmo de paso de mensajes* (Lauritzen y Spiegelhalter, 1988) opera enviando mensajes entre los nodos de la red bayesiana. Un nodo actúa como un procesador autónomo que colecta los mensajes que le llegan (factores) de sus nodos vecinos, efectúa ciertas operaciones (sumas y multiplicaciones) y envía un mensaje de salida a sus nodos vecinos. El proceso termina cuando el nodo raíz ha recibido mensajes de todos sus nodos adjuntos.

Ejemplo. Riesgo de demencia (algoritmo de eliminación de variables)

Supongamos que estamos interesados en la probabilidad de `Stroke` para un paciente que no sufre de demencia, $p(S \mid \neg d)$. Comenzamos con una lista $\mathcal{L}$ que contiene todas las distribuciones locales dadas en la red bayesiana instanciadas por $D = \neg d$: $\mathcal{L} = \{f_A(A), f_N(N,A), f_S(S,A), f_P(P,S), f_D(\neg d, S, N)\}$.

Consideremos el orden de eliminación *P-A-N*. En primer lugar eliminamos P computando $\sum_P f_P(P,S) \equiv 1$. La nueva lista no contiene $f_P(P,S)$. En segundo lugar eliminamos A calculando $\sum_A f_A(A) f_N(N,A) f_S(S,A) = g_1(N,S)$, que nos proporciona $g_1(n,s) = 0,0114, g_1(\neg n, s) = 0,0686, g_1(n, \neg s) = 0,1086$ y $g_1(n, \neg s) = 0,8114$. Ahora la lista es $\mathcal{L} = \{g_1(N,S), f_D(\neg d, S, N)\}$. Finalmente, eliminamos N calculando $\sum_N f_D(\neg d, S, N) g_1(N,S) = g_2(S)$, con $g_2(s) = 0,0382$ y $g_2(\neg s) = 0,7954$. Ahora $\mathcal{L} = \{g_2(S)\}$ y $g_2(S)$ es $p(S, \neg d)$. Para obtener $p(S \mid \neg d)$, una normalización es necesaria:

$$p(s \mid \neg d) = \frac{0,0382}{0,0382 + 0,7954} = 0.0458 \text{ y } p(\neg s \mid \neg d) = \frac{0,7954}{0,0382 + 0,7954} = 0,9542.$$

La expresión computada ha sido $p(S \mid \neg d) \propto \sum_N p(\neg d \mid N, S) \sum_A p(N \mid A) p(S \mid A) p(A) \sum_P p(P \mid S)$.

3.2.2. Inferencia aproximada

La inferencia aproximada se plantea como una alternativa a la exacta para redes bayesianas extremadamente complejas y de gran tamaño. Sigue siendo un problema NP-difícil (Dagum y Luby, 1993). La inferencia aproximada se basa en la simulación de la red bayesiana para generar un gran número de casos (instancias completas) de la JPD, y posteriormente estimar la probabilidad solicitada a partir de las frecuencias observadas en las muestras.

Consideremos que nuestro objetivo es estimar la probabilidad $p(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{e})$ para cualquier $\boldsymbol{Y} \subseteq \{X_1,...,X_n\}$. El *muestreo lógico probabilístico* (Henrion, 1988) parte de un orden ancestral (padres anteceden a los hijos) entre los nodos. Se muestrea un nodo X después de muestrear de todos sus padres, $\boldsymbol{Pa}(X)$, lo cual proporciona un valor fijo, $\boldsymbol{pa}(X)$. Después de haber muestreado $p(X|\boldsymbol{pa}(X))$ de todos los nodos, obtenemos una muestra de $p(\boldsymbol{X})$, la distribución de probabilidad conjunta. Este proceso se repite M veces. Podemos estimar $p(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{p(\boldsymbol{y},\boldsymbol{e})}{p(\boldsymbol{e})}$ a partir de la estimación del numerador, como la fracción de muestras (del total de M) en las que hemos visto ($\boldsymbol{y}$, $\boldsymbol{e}$) y del denominador como la fracción de muestras (del total de M) donde hemos visto $\boldsymbol{e}$.

Otros métodos de inferencia aproximada incluyen el denominado *pesado por verosimilitud* (Shachter y Peot, 1989) y el *muestreo de Gibbs* (Pearl, 1987).

3.3. Aprendizaje a partir de datos

Los modelos de redes bayesianas pueden obtenerse a partir de la interacción con un experto en el dominio a modelar, de manera automática induciendo el modelo a partir de un algoritmo o como una combinación de ambos.

3.3.1. Descubrimiento de relaciones entre variables

R_i denota la cardinalidad de Ω_{Xi}, es decir el número de valores posibles de la variable X_i. Sea $q_i = |\Omega_{\mathbf{pa}(X_i)}|$ el número de combinaciones posibles, cada una denotada $\mathbf{pa}_i^j$, de los valores de los padres de X_i, es decir, $\Omega_{\mathbf{pa}(X_i)} = \left\{\mathbf{pa}_i^1,...,\mathbf{pa}_i^{q_i}\right\}$. Entonces, la CPT de X_i contiene los parámetros $\theta_{ijk} = p(X_i = k \mid \mathbf{Pa}(X_i) = \mathbf{pa}_i^j)$, la probabilidad condicional de que X_i tome su k-ésimo valor dado que sus padres toman su j-ésimo valor. Por lo tanto, la CPT de X_i requiere la estimación de los parámetros $\boldsymbol{\theta}_i$, un vector de $R_i q_i$ componentes. $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,..., \boldsymbol{\theta}_n)$ que incluye todos los parámetros en la red bayesiana, es decir, $\theta_{ijk}, \forall i = 1,...,n, j = 1,...,q_i, k = 1,...,R_i$, siendo un vector con $\Sigma_{i=1}^{n} R_i q_i$ componentes.

Aprendizaje de parámetros. Una vez que se ha encontrado la estructura de la red bayesiana $\mathcal{G}$, los parámetros θ_{ijk} se estiman a partir del conjunto de datos $\mathcal{D}$. Sean N_{ij} el número de casos en $\mathcal{D}$ en los que se ha observado la configuración $\mathbf{Pa}(X_i) = \mathbf{pa}_i^j$, y N_{ijk} el número de casos en $\mathcal{D}$ donde se ha observado simultáneamente que $X_i = k$ y $\mathbf{Pa}(X_i) = \mathbf{pa}_i^j$ $(N_{ij} = \Sigma_{k=1}^{R_i} N_{ijk})$.

La *estimación máximo verosimil* busca los valores $\hat{\theta}^{ML}$ de los parámetros que maximizan la verosimilitud del conjunto de datos dado el modelo:

$$\hat{\boldsymbol{\theta}}^{ML} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}, \mathcal{G}) = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{h=1}^{N} p(x^h \mid \mathcal{G}, \boldsymbol{\theta}) \qquad [3]$$

Usando las premisas de *independencia global de los parámetros* y de *independencia local de los parámetros* (Spiegelhalter y Lauritzen, 1990), la anterior expresión puede calcularse como $\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}, \mathcal{G}) = \Pi_{i=1}^{n} \Pi_{j=1}^{q_i} \Pi_{k=1}^{R_i} \theta_{ijk}^{N_{ijk}}$. Por lo tanto, los estimadores máximo verosímiles de θ_{ijk} pueden estimarse mediante conteos de frecuencias en $\mathcal{D}: \hat{\theta}_{ijk}^{ML} = \frac{N_{ijk}}{N_{ij}}$, donde $N_{ij} = \Sigma_{k=1}^{R_i} N_{ijk}$.

En el *enfoque de estimación bayesiana* los parámetros $\boldsymbol{\theta}$ se modelan con una variable aleatoria con distribución de probabilidad $f(\boldsymbol{\theta} \mid \mathcal{G})$, codificando el conocimiento previo sobre los posibles valores de $\boldsymbol{\theta}$. La distribución *a posteriori* dados los datos $\mathcal{D}$ y el grafo $\mathcal{G}$ se calcula aplicando la regla de Bayes: $f(\boldsymbol{\theta} \mid \mathcal{D}, \mathcal{G}) \propto p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathcal{G})$. Esta distribución puede resumirse usando alguna medida de tendencia central, típicamente la media *a posteriori* $\hat{\boldsymbol{\theta}}^{Ba} = \int \boldsymbol{\theta} f(\boldsymbol{\theta} \mid \mathcal{D}, \mathcal{G}) d\boldsymbol{\theta}$.

La distribución de Dirichlet es una familia conjugada para distribuciones categóricas (como las que aparecen en las redes bayesianas discretas). Por lo tanto, asumiendo una distribución de Dirichlet para la distribución *a priori* (Spiegelhalter y Lauritzen, 1990), la *a posteriori* también seguirá una distribución de Dirichlet. Así, tenemos que para la distribución *a posteriori* de los parámetros $\boldsymbol{\theta}_{ij} = (\theta_{ij}, ..., \theta_{ijRi})$ en la variable condicionada $Xi \mid \mathbf{pa}_i^j$, si $(\boldsymbol{\theta}_{ij} \mid \mathcal{G}) \sim Dir(\alpha_{ij1}, ..., \alpha_{ijR_i})$, entonces $(\boldsymbol{\theta}_{ij} \mid \mathcal{D}, \mathcal{G}) \sim Dir(\alpha_{ij1} + N_{ij1}, ..., \alpha_{ijRi} + N_{ijR_i})$, y por lo tanto la media *a posteriori* de cada $\boldsymbol{\theta}_{ijk} (k = 1, ..., R_i)$ da lugar a la estimación bayesiana: $\hat{\theta}_{ijk}^{Ba} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}}$, donde $\alpha_{ij} = \Sigma_{k'=1}^{R_i} \alpha_{ijk'}$. Los hiperparámetros α_{ijk} pueden interpretarse como conteos imaginarios basados en nuestra experiencia previa. La interpretación de la distribución de Dirichlet sería que antes de obtener $\mathcal{D}$ hemos observamos virtualmente una muestra de tamaño α_{ij}, donde $X_i = k \mid \mathbf{pa}_i^j$, fue visto α_{ijk} veces.

Aprendizaje de estructuras. Existen dos maneras diferentes de abordar el aprendizaje de la estructura: (a) Testando las independencias condicionales de las variables por medio de tests de hipótesis; (b) Puntuando la bondad de cada estructura candidata la cual es buscada heurísticamente.

(a). Los métodos basados en restricciones testan estadísticamente las independencias condicionales entre tripletas de variables a partir de los datos, para a continuación encontrar un DAG que representa un gran porcentaje de (y siempre que sea posible todas) las restricciones de independencia condicional identificadas por los tests de hipótesis.

El método más representativo es el *algoritmo PC* (Spirtes y Glymour, 1991). PC comienza con un grafo completo no dirigido y consta de tres etapas. La etapa 1 produce las adyacencias en el grafo (el esqueleto de la estructura aprendida) mediante la eliminación de aristas a través de pruebas de hipótesis. La etapa 2 identifica colisionadores que sirven para orientar algunas aristas. La etapa 3 tiene como objetivo orientar el resto (dentro de lo posible) de las aristas.

La etapa 1 actúa de manera iterativa. La primera iteración de PC verifica la independencia marginal para todas las parejas de nodos (X_i, X_j), es decir, sin condicionar en otras variables $(\boldsymbol{S} = \emptyset)$. Si son independientes, se elimina la conexión y el conjunto vacío se guarda como conjunto de separación en $\boldsymbol{S}_{ij} = \boldsymbol{S}_{ji}$. En la siguiente iteración, el tamaño t de $\boldsymbol{S}$ aumenta en una unidad. Así, el algoritmo verifica si para cada pareja ordenada (X_i, X_j) que aún están adyacentes en $\mathcal{G}$, X_i y X_j son c.i. dado $\boldsymbol{S}$, para cualquier $\boldsymbol{S}$ de tamaño t=1 de $\text{Adj}_i \backslash \{X_j\}$. Se consideran todos los posibles conjuntos $\boldsymbol{S}$ según el orden. La arista X_i–X_j se elimina si y solo si se encuentra un conjunto $\boldsymbol{S}$ que haga a $\boldsymbol{X}_i$ y $\boldsymbol{X}_j$ c.i. Este $\boldsymbol{S}$ se guarda como conjunto de separación en $\boldsymbol{S}_{ij} = \boldsymbol{S}_{ji}$. Si se han considerado todas las parejas ordenadas de nodos adyacentes para testar la independencia condicional dados todos los subconjuntos $\boldsymbol{S}$ de tamaño t de sus conjuntos de adyacencia, el algoritmo aumenta t en una unidad. Por lo tanto, el proceso se repite para conjuntos de condicionamiento $\boldsymbol{S}$ con dos nodos, luego tres nodos, etc., afinando el grafo hasta que se agoten todas las posibilidades (hasta n–2 variables), y no queden más conjuntos de adyacencia por verificar. El número de pruebas de independencia condicional para una pareja dada (X_i, X_j) se va reduciendo en cada iteración. A medida que PC avanza, la fiabilidad de las pruebas estadísticas de independencia condicional disminuye porque los conjuntos $\boldsymbol{S}$ incluyen un número creciente de variables (reduciendo así el tamaño de las muestras sobre las que se aplican las pruebas).

Basado en el esqueleto y los conjuntos de separación, la etapa 2 identifica colisionadores para orientar algunas aristas. En las conexiones divergentes ($X_i \leftarrow X_j \rightarrow X_k$) y en serie ($X_i \rightarrow X_j \rightarrow X_k$) entre tres nodos, X_i y X_k son independientes dado X_j. Para una conexión convergente ($X_i \rightarrow X_j \leftarrow X_k$), X_i y X_k son dependientes dado X_j. Por lo tanto, si consideramos X_i–X_j–X_k (donde X_i y X_k no son adyacentes) en $\mathcal{G}$, primero identificamos un colisionador en X_j probando si X_i y X_k son dependientes dado X_j. Si lo son, dibujamos $X_i \rightarrow X_j \leftarrow X_k$. Esto se puede inferir a partir de las pruebas ya realizadas para eliminar la arista entre X_i y X_k y los conjuntos de separación guardados. X_j es un colisionador si y solo si $X_j \notin S_{ik}$, es decir, si X_j no se incluyó en el conjunto de condicionamiento que hizo a X_i y X_k c.i.

Finalmente, la etapa 3 orienta tantas aristas no dirigidas restantes como sea posible. Este proceso se realiza por consistencia con las aristas ya orientadas y de modo que no se formen nuevos colisionadores ni ciclos. Se lleva a cabo de manera recursiva hasta que no se puedan orientar más aristas. Las siguientes tres reglas se aplican repetidamente:

- Orientar X_j–X_k como $X_j \rightarrow X_k$ siempre que haya una arista dirigida $X_i \rightarrow X_j$ tal que X_i y X_k no sean adyacentes (de lo contrario se crearía un nuevo colisionador);
- Orientar X_i–X_j como $X_i \rightarrow X_j$ siempre que se forme una cadena $X_i \rightarrow X_k \rightarrow X_j$ (de lo contrario se crearía un ciclo dirigido);
- Orientar X_i–X_j como $X_i \rightarrow X_j$ siempre que haya dos cadenas X_i-$X_k \rightarrow X_j$ y X_i–$X_h \rightarrow X_j$ tales que X_k y X_h no sean adyacentes (de lo contrario se crearía un nuevo colisionador o un ciclo dirigido).

(b). Los *métodos basados en puntuación y búsqueda* utilizan una función de puntuación relativa a los datos para medir la calidad de cada estructura de red bayesiana candidata. El objetivo es encontrar una estructura de red que maximice la función de puntuación. Los métodos generalmente comienzan a partir de una estructura inicial (generada aleatoriamente o a partir del conocimiento ya existente sobre el dominio). Un método de búsqueda, habitualmente basado en un heurístico, es el encargado de realizar movimientos inteligentes en el espacio de posibles estructuras de red. En esta sección nos centraremos en el espacio de los DAG.

La cardinalidad del *espacio de los DAG* viene dada por la fórmula recursiva de Robinson (Robinson, 1977) y es superexponencial en el número de nodos. En concreto, el número $f(n)$ de posibles DAG que contienen n nodos está dado por la recurrencia $f(n)=\Sigma_{i=1}^{n}(-1)^{i+1}\binom{n}{i}2^{i(n-i)}f(n-i)$, para $n>2$, la cual se inicializa con $f(0)=f(1)=1$. La tarea de encontrar una estructura de red que optimice la puntuación es un problema de optimización combinatoria y se sabe que es NP-difícil (Chickering, 1996). Esto motiva el uso de heurísticos para llevar a cabo dicha búsqueda.

La calidad del ajuste de un DAG $\mathcal{G}$ al conjunto de datos $\mathcal{D}$ se mide por una función $Q(\mathcal{D},\mathcal{G})$, de tal manera que cuanto mejor sea el ajuste, mayor será la puntuación. El problema del aprendizaje de la estructura es encontrar arg máx$_{\mathcal{G}}$ $Q(\mathcal{D},\mathcal{G})$. Un criterio simple es la *log-verosimilitud estimada de los datos dada la red bayesiana*:

$$\log \mathcal{L}(\hat{\theta} \mid \mathcal{D},\mathcal{G}) = \log p(\mathcal{D} \mid \mathcal{G},\hat{\theta}) = \log \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{R_i}\hat{\theta}_{ijk}^{N_{ijk}} = \Sigma_{i=1}^{n}\Sigma_{j=1}^{q_i}\Sigma_{k=1}^{R_i}N_{ijk}\log\hat{\theta}_{ijk}, \qquad [4]$$

donde $\hat{\theta}_{ijk}=\hat{\theta}_{ijk}^{ML}=\dfrac{N_{ijk}}{N_{ij}}$ es el estimador máximo verosimil de θ_{ijk}.

El inconveniente de usar la verosimilitud como puntuación es que aumenta de manera monótona con la complejidad del modelo (sobreajuste estructural). Por lo tanto, la estructura que maximiza la verosimilitud coincide con el grafo completo. Para evitar este problema se ha propuesto una familia de puntuaciones de *log-verosimilitud penalizada* que penalizan la complejidad de la red. Su expresión general es $Q^{pen}(\mathcal{D},\mathcal{G})=\Sigma_{i=1}^{n}\Sigma_{j=1}^{q_i}\Sigma_{k=1}^{R_i}N_{ijk}\log\dfrac{N_{ijk}}{N_{ij}}-dim(\mathcal{G})pen(N)$, donde $dim(\mathcal{G})=\Sigma_{i=1}^{n}(R_i-1)q_i$ denota la dimensión del modelo (número de parámetros necesarios en la red bayesiana), y $pen(N)$ es una función de penalización no negativa. Si $pen(N)=1$, tenemos el *criterio de información de Akaike* (AIC) (Akaike, 1974) y si $pen(N)=\dfrac{1}{2}\log N$, el *criterio de información bayesiano* (BIC) (Schwarz, 1978).

El objetivo de un enfoque bayesiano es encontrar la estructura con la máxima probabilidad *a posteriori* dado el conjunto de datos, es decir, arg máx$_{\mathcal{G}}$ $p(\mathcal{G}|\mathcal{D})$. Usando la fórmula de Bayes, $p(\mathcal{G} \mid \mathcal{D}) \propto p(\mathcal{D},\mathcal{G}) = p(\mathcal{D} \mid \mathcal{G})p(\mathcal{G})$. El segundo factor, $p(\mathcal{G})$, denota la distribución *a priori* sobre las estructuras. El primer factor, $p(\mathcal{D}|\mathcal{G})$, es la *verosimilitud marginal* de los datos, definida como $p(\mathcal{D} \mid \mathcal{G}) = \int p(\mathcal{D} \mid \mathcal{G},\boldsymbol{\theta}) = f(\boldsymbol{\theta} \mid \mathcal{G})d\boldsymbol{\theta}$, donde $p(\mathcal{D}|\mathcal{G},\boldsymbol{\theta})$ es la verosimilitud de los datos dada la red bayesiana (estructura $\mathcal{G}$ y parámetros $\boldsymbol{\theta}$), y $f(\boldsymbol{\theta}|\mathcal{G})$ es la distribución *a priori* sobre los parámetros.

Supongamos que el vector de parámetros de la red bayesiana tiene una distribución de Dirichlet $(\boldsymbol{\theta}_{ij} \mid \mathcal{G}) \sim Dir(\alpha_{ij1},...,\alpha_{ijR_i})$. En tal caso (Cooper y Herskovits, 1992) obtuvieron una expresión cerrada para la verosimilitud marginal que al tomar logaritmos se transforma en la puntuación *bayesiana de Dirichlet*

$$Q^{BD}(\mathcal{D},\mathcal{G}) = \log p(\mathcal{G}) + \Sigma_{i=1}^{n}\Sigma_{j=1}^{q_i}\left(\log\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} + \Sigma_{k=1}^{R_i}\log\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}\right).$$

Esta puntuación tiene poco interés práctico ya que requiere la especificación de todos los hiperparámetros α_{ijk} para todos los *i, j, k*. Por esta razón se han propuesto algunos casos particulares más simples. La *puntuación K2* (Cooper y Herskovits, 1992) utiliza la asignación no informativa $\alpha_{ijk} = 1$, para todos los *i, j, k*, obteniéndose

$$Q^{K2}(\mathcal{D},\mathcal{G}) = \log p(\mathcal{G}) + \Sigma_{i=1}^{n}\Sigma_{j=1}^{q_i}\left(\log\frac{(R_i-1)!}{(N_{ij}+R_i-1)!} + \Sigma_{k=1}^{R_i}\log N_{ijk}!\right).$$

En cuanto a procedimientos de búsqueda, el más conocido es el *algoritmo K2* (Cooper y Herskovits, 1992). K2 utiliza un método de búsqueda voraz y la *puntuación K2*. Comienza con una estructura sin padres, un orden sobre los nodos y un límite superior sobre el número máximo de padres que se permite tener a cualquier nodo. El algoritmo agrega de manera incremental, desde el conjunto de nodos que preceden a cada nodo X_i, y siguiendo el orden establecido, el padre cuya adición incremente más la función $g(X_i, \mathbf{Pa}(X_i)) = \prod_{j=1}^{q_i}\frac{(R_i-1)!}{(N_{ij}+R_i-1)!}\prod_{k=1}^{R_i}N_{ijk}!$. Cuando la adición de ningún padre individual no mejora la puntuación, no se añaden más padres al nodo X_i, y pasamos al siguiente nodo en el orden.

Otros algoritmos que se han usado para la búsqueda de la mejor estructura de red bayesiana han sido: búsqueda voraz (Buntine, 1991), enfriamiento estocástico (Heckerman *et al.,* 1995) métodos MCMC (Giudici y Green, 1999), algoritmos genéticos (Larrañaga *et al.,* 1996), y algoritmos de estimación de distribuciones (Larrañaga *et al.,* 2000).

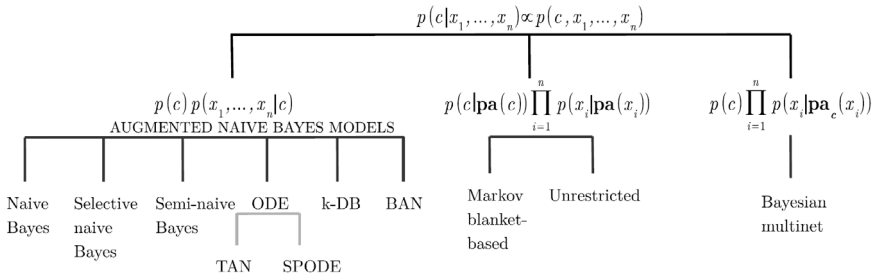
3.3.2. Clasificadores bayesianos

Los clasificadores bayesianos con variables predictoras discretas aproximan $p(\boldsymbol{x}, c)$ con la factorización proporcionada por la red bayesiana para las variables aleatorias $\boldsymbol{X_1},...,\boldsymbol{X_n}$, C. Es decir $p(\boldsymbol{x},c) = p(c \mid \mathbf{pa}(c))\prod_{i=1}^{n}p(x_i \mid \mathbf{pa}(x_i))$. Con una función de pérdida 0-1, la regla de decisión de Bayes consiste en encontrar c* tal que $c^* = \arg\max_c p(c \mid \boldsymbol{x}) = \arg\max_c p(\boldsymbol{x} \mid c)$.

La figura 4 muestra un esquema de distintos clasificadores bayesianos en función de si la variable C tenga o no padres y de si se permiten independencias condicionales dependientes del contexto.

Figura 4.

Clasificaciones bayesianos discretos en base a la factorización de $p(x,c)$



Fuente: Elaboración propia.

Para el caso en que $\mathbf{Pa}(C) = \emptyset$, el problema consiste en máx$_c$ $p(c)p(\boldsymbol{x}|c)$. En tal caso, los diferentes clasificadores bayesianos se agrupan bajo el nombre de *modelos de naive Bayes aumentado* donde se recogen modelos clasificatorios que van aumentando en complejidad. Si $\mathbf{Pa}(C) \neq \emptyset$, $p(\boldsymbol{x},c)$ se puede factorizar de diversas maneras, bien por medio del *manto de Markov* de C o usando *clasificadores basados en redes bayesianas no restringidas*. Finalmente, se pueden considerar relaciones de independencia condicional específicas para cada valor de c, lo cual da origen a *clasificadores bayesianos basados en multirredes*.

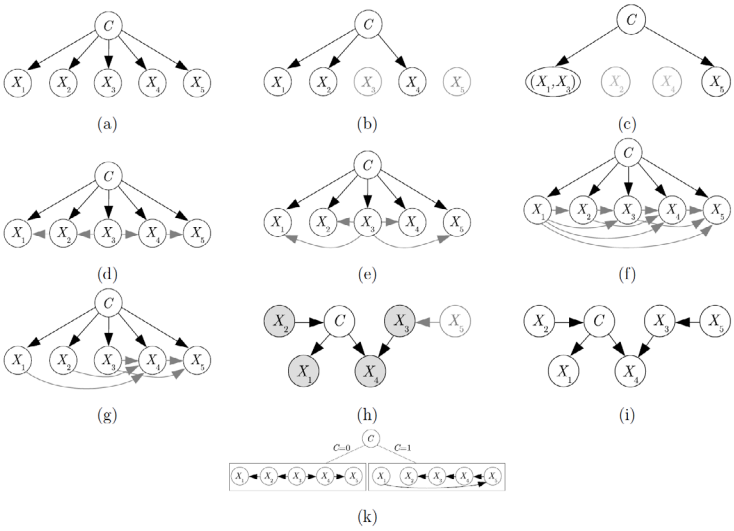
La figura 5 muestra las estructuras de los diferentes clasificadores bayesianos que se presentarán a continuación. *Naive Bayes* (Maron y Kuhns, 1960) es el clasificador bayesiano más simple. Las variables predictoras se asumen que son condicionalmente independientes dada la clase. Esta premisa, si bien está muy alejada de lo que realmente ocurre en un problema real, resulta ser muy útil cuando n es grande y/o N es pequeño. La premisa conduce a que $p(c \mid \boldsymbol{x}) \propto p(c)\prod_{i=1}^{n} p(x_i \mid c)$ (Minsky, 1961) demostró que la frontera de decisión asociada a un clasificador naive Bayes es un hiperplano.

Las prestaciones en cuanto a resultados clasificatorios del naive Bayes mejoran si tan sólo se consideran aquellas variables predictoras que son relevantes y no redundantes. Este proceso de selección de variables conduce al modelo *naive Bayes selectivo*, donde $p(c \mid \boldsymbol{x}) \propto p(c \mid \boldsymbol{x}_F) = p(c)\prod_{i\in F} p(x_i \mid c)$, y $\boldsymbol{X}_F$ denota la proyección de $\boldsymbol{X}$ en el subconjunto de variables seleccionadas $F \subseteq \{1, 2, ..., n\}$. La selección de variables puede llevarse a cabo por medio de una aproximación por filtrado (Pazzani y Billsus, 1997), o a partir de una aproximación basada en envoltura (Inza *et al.,* 2000).

El modelo *semi naive Bayes* relaja la asunción de independencia condicional introduciendo nuevas variables obtenidas como el producto cartesiano de dos o más variables predictoras originales. Estas nuevas variables predictoras siguen siendo condicionalmente independientes dado el valor de la clase. Se tiene $p(c \mid \boldsymbol{x}) \propto p(c)\prod_{j=1}^{K} p(x_{S_j} \mid c)$, donde $S_j \subseteq \{1,2,\ldots,n\}$ denota los índices en la j-ésima variable (original o producto cartesiano), $j = 1, ..., K$, $S_j \cap S_l = \emptyset$, para $j \neq l$. El algoritmo *de selección secuencial hacia adelante y unión*

Figura 5.

Estructuras de clasificación bayesianos



Notas: (a) Naive Bayes. (b) Selective naive Bayes. (c) Seminaive Bayes. (d) Naive Bayes aumentado a árbol. (e) Clasificador basado en superpadres con dependencias de orden 1. (f) Clasificador bayesiano *k*-dependiente. (g) Naive Bayes aumentado a red bayesiana. (h) Clasificador bayesiano basado en el manto de Markov. (i) Clasificador bayesiano no restringido. (k) Clasificador bayesiano basado en multirredes. Fuente: Elaboración propia.

(Pazzani, 1996) se guía por la precisión en la clasificación, comenzando con una estructura vacía. A continuación, el algoritmo considera la mejor opción entre (a) agregar una variable no utilizada por el clasificador actual como condicionalmente independiente de las variables (originales o productos cartesianos) utilizadas en el clasificador, y (b) unir una variable no utilizada por el clasificador actual con cada variable (original o producto cartesiano) presente en el clasificador. La mejor de estas dos opciones se aplica de manera iterativa hasta que el modelo deje de mejorar en sus prestaciones clasificatorias.

El subgrafo formado por las variables predictoras de un modelo *naive Bayes aumentado con árbol* (TAN) (Friedman *et al.,* 1997) es un árbol. Por lo tanto, todas las variables predictoras tienen exactamente un padre, excepto la variable raíz del árbol, que no tiene padres. El algoritmo para aprender una estructura TAN calcula la información mutua de cualquier par de variables predictoras condicionada a C, $I(X_i, X_j|C)$. A continuación, las aristas de un grafo no dirigido completo con nodos X_1, ..., X_n se anotan con los números de información mutua condicional calculados previamente. El algoritmo de Kruskal (Kruskal, 1956) se usa para encontrar las $n-1$ aristas del grafo, de tal manera que formen un árbol y cuya suma de pesos sea máxima. El árbol no dirigido se convierte entonces en un árbol dirigido seleccionando una variable al azar como nodo raíz y reemplazando las aristas por arcos. Final-

mente, se superpone una estructura naive Bayes para formar la estructura TAN. Se tiene: $p(c \mid \boldsymbol{x}) \propto p(c)p(x_r \mid c)\prod_{i=1,i\neq r}^{n} p(x_i \mid c, x_{j(i)})$, donde X_r denota el nodo raíz y $\{X_{j(i)}\} = \mathbf{Pa}(X_i) \setminus C$, para cualquier $i \neq r$.

Los *clasificadores bayesianos basados en superpadres con dependencias de orden* 1 (SPODEs) (Keogh y Pazzani, 2002) son modelos en los que todas las variables predictoras dependen de una única variable predictora (la misma para todas) llamada superpadre, además de la clase. En tal caso, $p(c \mid \boldsymbol{x}) \propto p(c)p(x_{sp} \mid c)\prod_{i=1,i\neq sp}^{n} p(x_i \mid c, x_{sp})$, donde X_{sp} denota el nodo superpadre. La gran ventaja de esta aproximación es que no necesita de un algoritmo de aprendizaje para fijar la estructura.

El *clasificador bayesiano con k dependencias* (*k*-DB) (Sahami, 1996) permite que cada variable predictora tenga un máximo de *k* variables padre, además de la variable clase. Se tiene $p(c \mid \boldsymbol{x}) \propto p(c)\prod_{i=1}^{n} p(x_i \mid c, x_{i_1}, ..., x_{i_k})$ donde X_{i1}, ...,X_{ik} son los padres de X_i en la estructura. El orden de inclusión de las variables predictoras X_i en el modelo se basa en I(X_i, C), empezando con la más alta. Una vez que X_i entra en el modelo, sus padres se seleccionan eligiendo las *k* variables X_j en el modelo con los valores más altos de I(X_i, $X_j|C$).

El clasificador *naive Bayes aumentado con una red bayesiana* (BAN), tiene cualquier estructura de red bayesiana como el subgrafo de variables predictoras (sección 3.3.1). Se tiene $p(c \mid \boldsymbol{x}) \propto p(c)\prod_{i=1}^{n} p(x_i \mid \mathbf{pa}(x_i))$. Ezawa y Norton (1996) primero, calculan I(X_i, C) para las *n* variables predictoras y luego seleccionan el número mínimo de variables predictoras *k* que satisfacen $\Sigma_{j=1}^{k}\mathrm{I}(X_j, C) \geq t_{CX}\Sigma_{j=1}^{n}\mathrm{I}(X_j, C)$, donde $0 < t_{CX} < 1$. En segundo lugar, se calcula I(X_i, $X_j|C$) para todos los pares de las variables seleccionadas. Se seleccionan las aristas correspondientes a los valores más altos hasta que se supera un porcentaje t_{XX} de la información mutua condicional total $\Sigma_{i<j}^{k}\mathrm{I}(X_i, X_j \mid C)$. La direccionalidad de las aristas se basa en la ordenación de las variables predictoras obtenida en el primer paso.

El manto de Markov de C está formado por aquellas variables (padres, hijos y padres de los hijos de C) cuyo conocimiento es lo único que afecta a la distribución de probabilidad de la variable clase. Podemos inducir un *clasificador bayesiano basado en el manto de Markov* (Koller y Sahami, 1996), partiendo de todas las variables predictoras, y en cada paso eliminando del manto de Markov considerado en esa iteración, aquella variable que menos información aporta con respecto de C. En concreto, proponen eliminar la variable que permite mantener $p(C|\boldsymbol{MB}^{(t)}(C))$ lo más cercana posible a $p(C|\boldsymbol{X})$.

Los clasificadores bayesianos generales no restringidos no consideran a C como una variable especial en el proceso de inducción. Podemos usar cualquier algoritmo de la sección 3.3.1 para inducir el DAG con las n+1 variables. El correspondiente manto de Markov de C se usa con fines de clasificación.

Las *multirredes bayesianas* (Geiger y Heckerman, 1996) permiten codificar independencias condicionales asimétricas, que solo se cumplen para algunos, pero no para todos los valores de las variables involucradas. Para clasificación supervisada, consisten en varias redes

bayesianas (locales), cada una de las cuales está asociada con un subconjunto del dominio de una variable *C*. Así, condicionado a cada c, las variables predictoras pueden formar diferentes redes locales con diferentes estructuras. Por lo tanto, las relaciones entre las variables predictoras no tienen que ser las mismas para todos los valores de *c*. Se tiene $p(c \mid \boldsymbol{x}) \propto p(c)\prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_c(x_i))$, donde $\mathbf{Pa}_c(X_i)$ es el conjunto de padres de X_i en la red bayesiana local asociado con $C = c$.

4. INTERPRETABILIDAD CON REDES BAYESIANAS

Veamos cómo las redes bayesianas verifican las tres condiciones impuestas en [6] para que un paradigma pueda ser considerado como interpretable (sección 2).

Por lo que respecta a la *simulabilidad*, un ser humano puede reproducir el proceso que lleva a cabo una red bayesiana para sugerir una respuesta, ya que todos los procedimientos tanto exactos como aproximados para llevar a cabo cualquier tipo de inferencia consisten en efectuar sumas y multiplicaciones. Obviamente si el número de nodos del DAG es elevado y/o la densidad de las conexiones es alta, el seguimiento por parte del humano de todas las operaciones necesarias para que el modelo proporcione una respuesta puede ser tedioso.

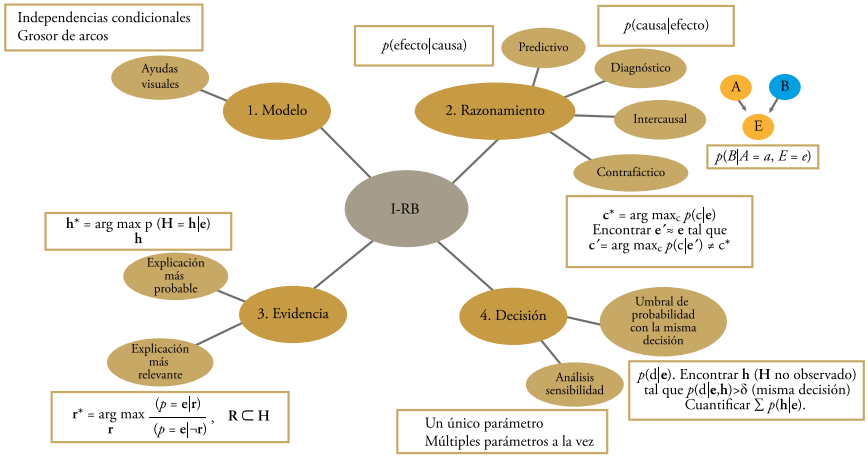
La *descomponibilidad* del modelo es una propiedad que es intrínseca a las redes bayesianas. La factorización de la distribución de probabilidad conjunta como producto de las distribuciones de probabilidad de cada variable dados sus padres en el DAG es la esencia de la modularidad que va a facilitar la interpretación del modelo global a partir de la interpretación de cada uno de los submodelos asociados a cada una de las variables en el DAG. Por otra parte, la descomponibilidad de las métricas habitualmente usadas para inducir el modelo a partir de datos (sección 3.3) hace que el aprendizaje de las estructuras de redes bayesianas pueda llevarse a cabo de manera modular.

Finalmente, la *transparencia algorítmica* tanto en el proceso de inducción del modelo, como de su posterior uso, son evidentes en las redes bayesianas. Tan sólo se requiere comprender conceptos básicos de Probabilidad, Estadística, y de Optimización Heurística, disciplinas en las que descansan ambos procesos.

La figura 6 muestra las posibilidades que proporciona una red bayesiana para poder interpretar un modelo desde distintos puntos de vista. Por lo que se refiere al *modelo* es factible testar en el grafo cualquier independencia condicional entre tripletas de variables a partir del criterio de u-separación. Adicionalmente podemos representar los arcos del DAG con grosores que sean proporcionales a la importancia de estos, cuantificando la importancia de cada arco a partir de su aportación a la métrica que el algoritmo de inducción está tratando de maximizar. Las redes bayesianas posibilitan distintos tipos de *razonamiento* que van desde un razonamiento predictivo, en el cual dadas unas causas nos preguntamos por las probabilidades de determinados efectos, o alternativamente por un razonamiento diagnóstico, en el cual se conocen las causas y se trata de calcular las probabilidades de los distintos efectos. El razonamiento intercausal (exclusivo de las redes bayesianas) asume la existencia de dos (o más causas) comunes para un mismo efecto. Si se conoce que una de las causas se ha pro-

Figura 6.

Interpretando una red bayesiana a partir del estudio del modelo, de diferentes tipos de razonamiento, de distintas opciones de propagación de la evidencia y de posibilidades de entender varias maneras de decidir



Fuente: Elaboración propia.

ducido, el modelo de red bayesiana va a reducir la probabilidad de que la otra causa haya ocurrido. En el denominado razonamiento contrafáctico se busca variar lo menos posible una determinada evidencia con el objetivo de que el valor de la variable clase predicha por el modelo cambie hacia una situación que resulte más satisfactoria (por ejemplo "vivo" en lugar de "muerto").

Dada una *evidencia*, dentro del denominado razonamiento abductivo, podemos preguntarnos por la instanciación del resto de las variables (inferencia abductiva global) que tenga asignada la máxima probabilidad tratando de encontrar la denominada explicación más probable. Si nuestro interés radicase exclusivamente en un subconjunto de las variables no instanciadas estaríamos tratando de resolver un problema de inferencia abductiva local. En la explicación más relevante, dada una evidencia se trata de encontrar la instanciación de un subconjunto de variables (a determinar por el algoritmo) no evidenciadas que tengan asociadas el máximo valor del factor de Bayes generalizado.

A nivel de *decisión*, dada una evidencia y una decisión adoptada al constatar que su probabilidad *a posteriori* dada la evidencia supera un determinado umbral preestablecido, la red bayesiana nos va a posibilitar la búsqueda de instanciaciones de las variables no observadas que actuando de condicionantes (junto a la evidencia) hagan que la probabilidad *a posteriori* de la decisión previa siga superando el umbral. La cuantificación de las probabilidades de todas las instanciaciones en las que se verifica esta condición nos permite determinar la robustez de la decisión adoptada. Las redes bayesianas también nos permiten llevar a cabo un análisis de sensibilidad bien de un único parámetro cada vez, o de varios parámetros al unísono.

5. INTERPRETANDO LA CLASIFICACIÓN MORFOLÓGICA DE INTERNEURONAS

El sistema nervioso humano es el sistema biológico más complejo. Para detectar y responder eficazmente a los cambios en el entorno, es capaz de aprender, tener autoconciencia y dar lugar al intelecto. Aunque muchos aspectos fundamentales de la estructura y función neuronal se entienden bien, quedan muchas preguntas sin respuesta. Responder a estas preguntas se está volviendo cada vez más urgente, principalmente debido al enorme costo social y económico de los trastornos del sistema nervioso. Los trastornos cerebrales, como la demencia, la depresión y la adicción, representan el 36 % de la carga total de enfermedades en los países de altos ingresos (Silberberg, 2015), con ocho millones de muertes atribuibles por año (Walker *et al.,* 2015).

Progresar hacia la comprensión del cerebro es un esfuerzo monumental. Con este fin, se han lanzado proyectos ambiciosos de neurociencia a nivel mundial en la última década. Estos incluyen el Proyecto Cerebro Humano (HBP) (Markram, 2012) de la Unión Europea, la iniciativa de Investigación del Cerebro a través del Avance de Neurotecnologías Innovadoras (BRAIN) (Insel *et al.,* 2013), el Instituto Allen para la Ciencia del Cerebro en los Estados Unidos, existiendo también iniciativas en Canadá, China, Japón, Corea e Israel. La mayoría son proyectos extremadamente grandes que requieren de esfuerzos interdisciplinarios, lo que refleja la complejidad de la tarea.

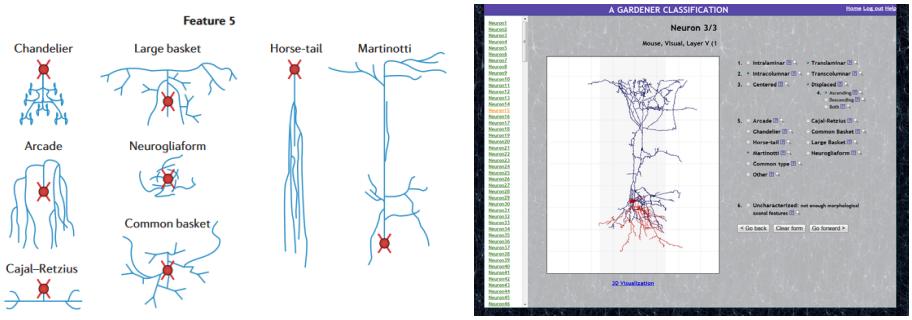
Las redes bayesianas han sido ampliamente aplicadas en la investigación en neurociencia. En Bielza y Larrañaga, 2014 se revisan muchas de estas aplicaciones. Un desafío clave en la neurociencia es la clasificación de interneuronas GABA-érgicas (Ascoli *et al.,* 2008). Estas neuronas constituyen alrededor del 20 %-30 % de las neuronas en la corteza cerebral y son el componente principal de los circuitos corticales inhibitorios, los cuales, a su vez, están asociados con trastornos como la epilepsia, el autismo y la esquizofrenia. Aunque la generación de datos a gran escala puede permitir el aprendizaje de una taxonomía sistemática a partir de datos en un futuro cercano mediante la agrupación de características moleculares, morfológicas y electrofisiológicas (Yuste *et al.,* 2020; Mihaljević *et al.,* 2018), los investigadores actualmente utilizan y se refieren a tipos morfológicos establecidos, como *chandelier, Martinotti, neurogliaforme* y *basket* (Markram *et al.,* 2004). Tener un modelo simple y preciso para clasificar automáticamente las interneuronas en estos tipos morfológicos podrían aportar conocimiento y ser útil para los profesionales.

Varios estudios han abordado el problema de la clasificación de interneuronas con métodos basados en redes bayesianas. Todos ellos se basan en un estudio fundamental sobre el consenso entre la comunidad científica en cuanto a la clasificación de interneuronas (DeFelipe *et al.,* 2013), en el cual 42 neurocientíficos expertos clasificaron 320 interneuronas de acuerdo con una taxonomía predefinida (figura 7). El estudio requería clasificar cinco variables morfológicas, además del tipo de inter- neurona, de ahí que por cada uno de los 42 neurocientíficos se obtuvo una tabla con 320 filas y 6 variables clase. Además, las morfologías de 240 de las 320 interneuronas fueron reconstruidas digitalmente, lo que permitió estudiar la clasificación supervisada de morfologías de interneuronas (Mihaljević *et al.,* 2014; Mihaljević *et al.,* 2015).

Figura 7.

Tipos de interneuronas y características morfológicas en el esquema de clasificación

(a) (b)



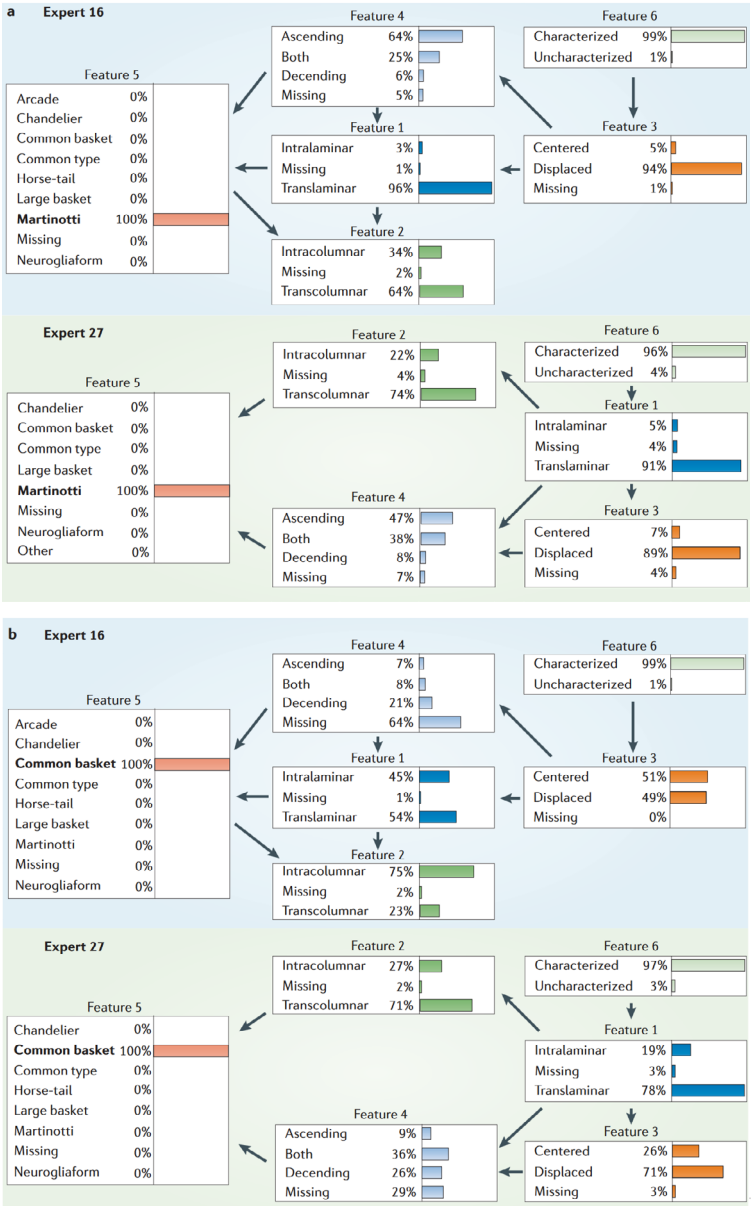
Nota: El esquema contempla diversos tipos de interneuronas: (a) Arcade, chandelier, horse-tail, Martinotti, common basket, Cajal-Retzius, large basket y neurogliaform (los tipos "común" y "otro" no se muestran en la gráfica). El tipo "otro" se utiliza cuando el neurocientífico considera que ninguno de los tipos restantes es adecuado. Además del tipo de interneurona, el esquema de clasificación contempla cinco variables morfológicas de alto nivel. Estas características, denominadas F1, F2, F3, F4 y F6 (F5 siendo el tipo de interneurona previamente discutido), tienen las siguientes categorías: (F1) intralaminar y translaminar; (F2) intracolumnar y transcolumnar; (F3) centrado y desplazado; (F4) ascendente, descendente y ambos; (F6) caracterizado y no caracterizado. La categoría no caracterizado de F6 significa que la reconstrucción de una célula no es lo suficientemente buena como para clasificarla de manera confiable. Cuando una célula se etiqueta como no caracterizada en la característica F6, el neurocientífico no puede anotarla según ninguna de las cinco características restantes, F1-F5. F4 solo se aplica a las células que se etiquetan como translaminares y desplazadas en F1 y F3, respectivamente. (b) La aplicación web utilizada para recopilar las elecciones de clasificación de los neurocientíficos para el conjunto de 320 interneuronas.
Fuente: DeFelipe *et al.* (2013).

En el estudio original (DeFelipe *et al.*, 2013), los autores utilizaron redes bayesianas para analizar las elecciones de clasificación de los neurocientíficos (figura 8). En particular, aprendieron una red bayesiana para cada neurocientífico con el fin de modelar su razonamiento en términos de las seis variables. Gracias a esta modelización se pudo estudiar, entre otras cosas, cómo cada experto relacionaba las características morfológicas, por ejemplo, si el axón era intra- o translaminar, con el tipo de interneurona. Estas redes permitieron identificar similitudes y diferencias en el razonamiento entre los expertos. Así, por ejemplo, vemos que los expertos 16 y 27 coinciden en muchas apreciaciones en cuanto al concepto de *Martinotti* (ver figura 8). Para ambos expertos se trata de un tipo de célula con una alta probabilidad de ser Translaminar, Caracterizada y Desplazada. Sin embargo, para el concepto de *common basket*, ambos expertos no se ponen de acuerdo. Mientras que para el experto 16 se trata de un tipo de interneurona que es mayormente Intracolumnar, para el experto 27 sería fundamentalmente Transcolumnar. También presentan mucha discrepancia en sus valoraciones de Desplazado y de Translaminar.

Un segundo objetivo fue predecir el tipo de interneurona y cuatro características morfológicas a partir de las morfologías reconstruidas digitalmente. En (Mihaljević *et al.*, 2015)

Figura 8.

Redes bayesianas para dos de los 42 neurocientíficos

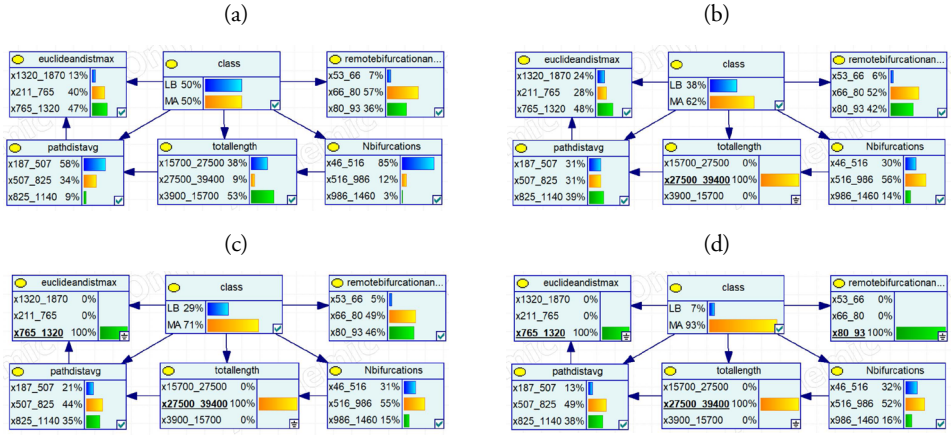


Nota: Los gráficos de barras muestran las probabilidades propagadas de las características restantes, condicionadas al tipo *Martinotti* (arriba) y al tipo *common basket* (abajo).
Fuente: DeFelipe *et al.* (2013).

se predijeron cada una de estas cinco variables por separado utilizando clasificadores bayesianos discretos. A diferencia de un escenario típico de clasificación supervisada, en este caso había hasta 42 etiquetas para cada instancia, proporcionadas por los diferentes neurocientíficos. El nivel de acuerdo entre los expertos variaba según las células: mientras que en 29 neuronas al menos 35 neurocientíficos coincidieron en su tipo de interneurona, hubo 67 células en las que no más de 15 expertos coincidieron en un tipo único. Los autores etiquetaron cada célula con la etiqueta más común entre las proporcionadas por los 42 neurocientíficos, pero repitieron la clasificación en diferentes subconjuntos de neuronas, formados al filtrar las células por debajo de un cierto umbral de fiabilidad de la etiqueta, definido como el número mínimo de neurocientíficos que coincidían en el tipo mayoritario. Los modelos alcanzaron una precisión de hasta el 89,52 % para el tipo de interneurona, siendo dicha precisión incluso mayor para las características morfológicas. La figura 9 ilustra cómo un clasificador naive Bayes aumentado a árbol puede utilizarse para interpretar el razonamiento detrás de la clasificación de una célula, proporcionando información sobre las características cuantitativas de dos tipos de interneuronas.

Figura 9.

Ilustración de la clasificación de una neurona con un clasificador bayesiano discreto naive Bayes aumentado a árbol, que discrimina entre los tipos de interneurona *Martinotti* (MA) y *large basket* (LB), aprendido a partir de 101 de las 240 reconstrucciones digitales utilizadas en DeFelipe *et al.* (2013)



Notas: El nodo de clase es el tipo de interneurona, mientras que los demás nodos corresponden a variables predictoras. Inicialmente, sin evidencia en las predictoras, una célula dada tiene la misma probabilidad de pertenecer a cualquiera de las dos clases (a). Si conocemos que la neurona tiene una longitud total media (en el rango de 27500–39400 µm) y establecemos eso como evidencia en la red (b), la probabilidad de que la neurona sea una célula Martinotti aumenta a 0,62. Observaciones posteriores sobre la máxima distancia euclidiana al soma (c) y el ángulo de bifurcación remota (d) de la neurona aumentan aún más esta probabilidad, hasta llegar a 0,93 en (d). Así, el neurocientífico puede entender la predicción del modelo y obtener información sobre las características cuantitativas de los dos tipos de interneuronas (Mihaljević *et al.*, 2021).
Fuente: DeFelipe *et al.* (2013).

6. CONCLUSIONES

Ante la necesidad cada vez más imperiosa de contar con sistemas inteligentes que sean interpretables por el ser humano (más allá de la explicabilidad a veces insuficiente y habitual en los paradigmas de caja negra) este capítulo ha mostrado la manera en la que las redes bayesianas dan respuesta a esta problemática. Por una parte, se ha mostrado cómo este paradigma verifica las tres condiciones necesarias para que el sistema sea considerado como interpretable: capacidad humana para simular el modelo, descomponibilidad del modelo en submodelos comprensibles, y transparencia algorítmica que posibilite al humano el entendimiento de los procesos de inducción y posterior utilización del modelo.

Adicionalmente hemos visto cómo las redes bayesianas posibilitan la interpretabilidad a distintos niveles. En primer lugar, a nivel de modelo con el testeo de independencias condicionales o con la asignación de pesos de importancia a cada uno de los arcos. En segundo lugar, posibilitando distintos tipos de razonamiento, desde el predictivo al diagnóstico, intercausal o contracfáctico. En tercer lugar, la evidencia introducida en la red bayesiana permite llevar a cabo distintos tipos de inferencia abductiva (global, parcial) como de explicación más probable. Finalmente, las decisiones tomadas con el modelo posibilitan dar soluciones al problema de determinar el umbral de decisión con la misma probabilidad, así como al análisis de sensibilidad del propio modelo.

En el capítulo nos hemos restringido a redes bayesianas discretas en dominios estáticos. Sin embargo, la extensión de los conceptos aquí mostrados a dominios continuos, con premisas de Gaussianidad o con estimaciones de densidades vía *kernels*, o incluso a dominios en los que los datos llegan temporalmente, a partir de las redes bayesianas dinámicas, las redes bayesianas en tiempo continuo, o los modelos ocultos de Markov, es factible existiendo un gran número de trabajos en la literatura. Finalmente, no sólo la clasificación supervisada puede beneficiarse de la interpretabilidad de las redes bayesianas, sino que problemas de regresión, *clustering*, aprendizaje por refuerzo u optimización heurística (vía algoritmos de estimación de distribuciones) pueden también ser interpretados bajo dicho prisma.

Referencias

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Ascoli, G. A. *et al.* (2008). Petilla terminology: Nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nature Reviews* Neuroscience, 9, 557–568.

Bielza, C., y Larrañaga, P. (2014). Bayesian networks in neuroscience: A survey. *Frontiers in Computational Neuroscience*, 8, 131.

Bielza, C., Li, G., y Larrañaga, P. (2011). Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning,* 52, 705-727.

Borchani, H., Varando, G., Bielza, C., y Larrañaga, P. (2015). A survey on multi-output regression. *WIREs Data Mining and Knowledge Discovery*, 5, 216–233.

Buntine, W. L. (1991). Theory refinement on Bayesian networks. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence* (52–609). Morgan Kaufmann.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. *Learning from Data: Artificial Intelligence and Statistics V,* 121–130. Springer.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence, 42*(23), 393–405.

Cooper, G. F., y Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 309–347.

Dagum, P., y Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence, 6*(1), 141–153.

Ezawa, K. J., y Norton, S. W. (1996). Constructing Bayesian networks to predict uncollectible telecommunications accounts. *IEEE Expert, 11*(5), 45–51.

Friedman, N., Geiger, D., y Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29, 131-163.

Geiger, D., y Heckerman, D. (1996). Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82, 45–74.

Giudici, P., y Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika, 86*(4), 785–801.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., y Yang, G. Z. (2019). XAI-Explainable artificial intelligence. *Scientific Robotics*, 4, 37.

Heckerman, D., Geiger, D., y Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20, 197–243.

Henrion, M. (1988). Propagating uncertainty in Bayesian networks by probabilistic logic sampling. *Uncertainty in Artificial Intelligence,* 2, 149–163. Elsevier.

Insel, T. R., Landis, S. C., y Collins, F. S. (2013). The NIH BRAIN initiative. *Science,* 340, 687–688.

Inza, I., Larrañaga, P., Etxeberria, R., y Sierra, B. (2000). Feature subset selection by Bayesian network–based optimization. *Artificial Intelligence,* 123, 157-184.

Keogh, E. J., y Pazzani, M. J. (2002). Learning the structure of augmented Bayesian classifiers. *International Journal on Artificial Intelligence Tools, 11*(4), 587–601.

Koller, D., y Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* The MIT Press.

Koller, D., y Sahami, M. (1996). Toward optimal feature selection. In *Proceedings of the 13th International Conference on Machine Learning,* 284–292.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society, 7*(1), 48-50.

Larrañaga, y P., Bielza, C. (2024). Estimation of distribution algorithms in machine learning: A survey. *IEEE Transactions on Evolutionary Computation, 28*(5).

Larrañaga, P., Etxeberria, R., Lozano, J. A., y Peña, J. M. (2000). Optimization in continuous domains by learning and simulation of Gaussian networks. In *Proceedings of the Genetic and Evolutionary Computation Conference* (201–204). Morgan Kaufmann.

Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R. H., y Kuijpers, C. M. H. (1996). Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(9), 912-926.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., y Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks, 20*(5), 491–505.

Lauritzen, S. L., y Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B (Methodological), 50*(2), 157–224.

Lipton, Z. C. (2019). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31-57.

Lukas-Valentin, H., Heinrich, K., Wanner, J., y Janiesc, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, 69, 102538.

Markram, H. (2012). The Human Brain Project. *Scientific American*, 306, 50–55.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, y G., Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience* 5, 793–807.

Maron, M., y Kuhns, J. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery,* 7, 216–244.

Mihaljević, B., Benavides-Piccione, R., Bielza, C., DeFelipe, J., y Larrañaga, P. (2015). Bayesian network classifiers for categorizing cortical GABAergic interneurons. *Neuroinformatics*, 13, 192–208.

Mihaljević, B., Bielza, C., Benavides-Piccione, R., DeFelipe, J., y Larrañaga, P. (2014). Multidimensional classification of GABAergic interneurons with Bayesian network-modeled label uncertainty. *Frontiers in Computational Neuroscience*, 8, 150.

Mihaljević, B., Bielza, C., y Larrañaga, P. (2021). Bayesian networks for interpretable machine learning and optimization. *Neurocomputing*, 456, 648-665.

Mihaljević, B., Larrañaga, P., Benavides-Piccione, R., Hill, S., DeFelipe, J., y Bielza, C. (2018). Towards a supervised classification of neocortical interneuron morphologies. *BMC Bioinformatics,* 19, 511.

Minsky, M. (1961). Steps toward artificial intelligence. *Transactions on Institute of Radio Engineers,* 49, 8–30.

Molnar, C. (2022). *Interpretable Machine Learning.* Lilu.com.

Muehlematter, U. J., Daniore, P., y Vokinger, K. N. (2020). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): A comparative analysis. *Lancet Digit Health,* 3, e195-203.

Nakka, R., Harursampath, D., y Ponnusami, S. A. (2023). A generalised deep learning-based surrogate model for homogenisation utilising material property encoding and physics-based bounds. *Scientific Reports*, 13, 9079.

Pazzani, M. (1996). Constructive induction of Cartesian product attributes. In *Proceedings of the Information, Statistics and Induction in Science Conference*, 66–77.

Pazzani, M., Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27, 313–331.

Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence, 32*(2), 245–257.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann.

Puerto-Santana, C., Larrañaga, P., y Bielza, C. (2022). Autoregressive asymmetric linear Gaussian hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4642-4658.

Robinson, R. (1977). Counting unlabeled acyclic digraphs. *Lecture Notes in Mathematics,* 622, 28-43. Springer.

Rodriguez-Sanchez, F., Rodriguez-Blazquez, C., Bielza, C., y Larrañaga, P. (2021). Identifying Parkinson´s disease subtypes with motor and non-motor symptoms via model-based multipartition clustering. *Scientific Reports, 11*(1), 1-10.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 335–338.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Shachter, R. C., y Peot, M. A. (1989). Simulation approaches to general probabilistic inference on belief networks. In *Proceedings of the 5th Annual Conference on Uncertainty in Artificial Intelligence* (221–234). Elsevier.

Shueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289-310.

Silberberg, D., Anand, N. P., Michels, y K., Kalaria, R. N. (2015). Brain and other nervous system disorders across the lifespan – global challenges and opportunities. *Nature*, 527, S151–S154.

Spiegelhalter, D. J., y Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20, 579–605.

Spirtes, P., y Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review, 90*(1), 62–72.

Valverde, G., Quesada, D., Larrañaga, P., y Bielza, C. (2023). Causal reinforcement learning based on Bayesian networks applied to industrial settings. *Engineering Applications of Artificial Intelligence*, 125, 106657.

Varando, G., Bielza, C., y Larrañaga, P. (2015). Decision boundary for discrete Bayesian network classifiers. *Journal of Machine Learning Research*, 16, 2725-2749.

Walker, E. R., McGee, R. E., y Druss, B. G. (2015). Mortality in mental disorders and global disease burden implications: A systematic review and meta-analysis. *JAMA Psychiatry*, 72, 334–341.

Yuste, R. *et al.* (2020). A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nature Neuroscience*, 1–13.