



Otros métodos de clasificación avanzados



Otros métodos de clasificación avanzados

UNIVERSIDAD DE CÓRDOBA

1. Random Forest

Random Forest es uno de los métodos específicos basado en ensembles más extendidos. Se puede considerar una implementación específica de Bagging, donde los clasificadores base son árboles de decisión; sin embargo, Random Forest va más allá. Por un lado, Random Forest entrena cada uno de sus clasificadores base con muestreo del conjunto de datos original, y utiliza un voto mayoritario para producir la predicción final.

Por otro lado, Random Forest incrementa la diversidad en cada uno de los árboles por medio de incluir procesos aleatorios para la selección de atributos en los distintos nodos. En cada nodo de decisión, no considera todos los posibles atributos (m) para realizar la partición, sino que considera las posibles particiones de un subconjunto aleatorio de los atributos disponibles de tamaño $m' < m$. En la literatura, es bastante extendido utilizar un valor de $m' = \log(m)$, aunque el valor podría modificarse dependiendo de cada problema específico. Nótese que, pese a que en estos nodos se utilice un subconjunto aleatorio de m' atributos, de entre ellos se selecciona la mejor partición posible de forma determinista (como se hace en los árboles de decisión usuales).

La idea de Random Forest es simple, pero aun así muy potente, y ha demostrado muy buenos resultados en la literatura. Este buen rendimiento se basa en la diversidad que se genera entre los distintos clasificadores base (es decir, árboles) del ensemble. Al estar basado en Bagging, y ya que todos los árboles del ensemble son independientes entre sí,

2. Redes neuronales artificiales

Las redes neuronales artificiales son otro de los métodos más extendidos en la literatura en clasificación. Más aún, con el reciente auge de las llamadas redes neuronales profundas (*Deep neural networks*). Dado que existe otro curso en el máster para el estudio de dichas redes, en esta sección solo se pretende introducir lo que son las redes neuronales para su conocimiento, sin entrar en mucho más detalle.

Las redes neuronales se conforman por tres elementos principales: capas, neuronas, y conexiones. En la Figura 1 se puede observar un ejemplo de red neuronal artificial. En cuanto a las capas, destacamos principalmente tres tipos o conjuntos capa de entrada, capa (o capas) oculta, y capa de salida.

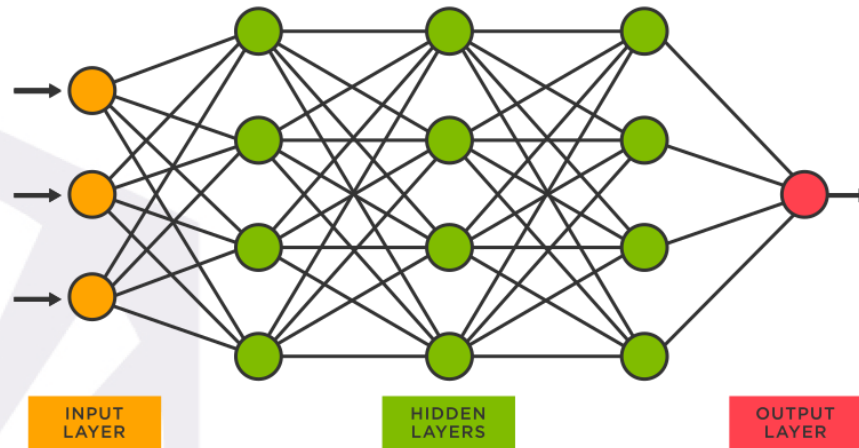


Figura 1. Ejemplo de red neuronal artificial

Cada capa está compuesta por un conjunto de neuronas. En la capa de entrada, existirán tantas neuronas como atributos de entrada; dado un patrón, en cada una de las neuronas de entrada se colocaría cada uno de los valores para cada uno de los atributos. En problemas binarios, en la capa de salida existirá, por lo general, una única neurona, cuya salida representa la probabilidad de pertenencia a una de las clases; en problemas multi-clase, sería suficiente con tener $k-1$ neuronas, siendo k el número de posibles clases.

Por otro lado, cada una de las neuronas de capa oculta representan una función de activación en base a las entradas recibidas. En los casos más sencillos, se suele utilizar una única capa oculta, aunque es posible tener varias capas (como se ve en la figura), cada una de ellas con distinto número de neuronas. Dadas las entradas de una neurona (que a continuación veremos), la función de esa neurona proporcionará una salida, que se irá propagando hacia adelante en siguientes capas.

Por último, las conexiones representan pesos entre las distintas neuronas. Así, la entrada de una neurona vendrá dada por el peso de cada conexión por el valor de la neurona de la capa anterior que la conectaba. Supongamos la red simple en la Figura 2. Las neuronas x_1 y x_2 representan los atributos de entrada del problema; h_1 y h_2 son dos neuronas de la única capa oculta existente; y o_1 es la neurona de salida. Las entradas a la neurona h_1 serían $x_1 * w_{1,1}^1$ y $x_2 * w_{2,1}^1$. En base a estas entradas, la función de activación de la neurona h_1 proporcionará una salida. De similar manera, la función de activación de la neurona de salida o_1 vendrá dada por las entradas $h_1 * w_1^2$ y $h_2 * w_2^2$, proporcionando la salida final para un patrón dado.

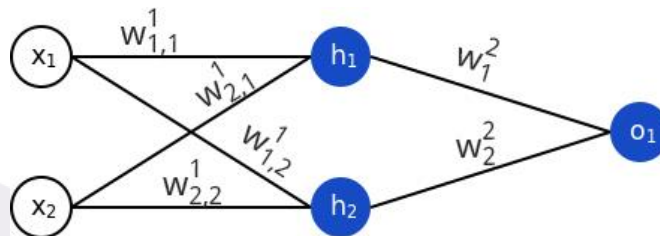


Figura 2. Ejemplo simple de red neuronal con dos neuronas de entrada y dos en capa oculta.

Como se comentaba anteriormente, el objetivo de esta sección es conocer los métodos de redes neuronales artificiales y su procedimiento de inferencia. Más detalles acerca de estos métodos, sus diferentes arquitecturas, y su construcción o aprendizaje se estudiarán en otro curso de este máster.

3. Otros métodos

En esta sección se mencionan (sin entrar en mayor detalle), otros métodos avanzados para clasificación, para conocimiento del estudiante. No se espera que el estudiante entre tampoco en mucho mayor nivel de detalle en ellos, pero sí al menos es interesante conocerlos por si en algún momento futuro tiene que enfrentarse a ellos. Además de estos métodos, existen muchos más; incluso se anima al estudiante a realizar una búsqueda para encontrar otros métodos avanzados no tratados en la asignatura.

- XGBoost¹. Se trata de una implementación específica de un método basado en *boosting* utilizando árboles de decisión internamente, y que también utiliza gradiente descendiente en su proceso de aprendizaje.
- Modelos ocultos de Markov²³. Estos modelos se basan en estructuras de autómatas abstractos de estados finitos que permiten modelar procesos estocásticos, basándose en la transición entre distintos estados.
- Modelos de aprendizaje profundo (*Deep learning*)⁴. Los métodos de aprendizaje profundo son unos de los métodos del estado del arte actualmente. Estos modelos están basados en las redes neuronales artificiales, pero su estructura ha avanzado hacia modelos con gran número de capas ocultas, y que pueden incluir procesos tanto de preprocesado como de aprendizaje. Se estudiarán estos métodos en más detalle en otro de los cursos de este máster.

¹ <https://xgboost.readthedocs.io/en/latest/index.html>

² Más información: <https://www.cs.us.es/cursos/aia-2019/temas/tema-Markov.pdf>

³ Más información: <https://www.redalyc.org/pdf/993/99324907003.pdf>

⁴ Más información: <https://es.mathworks.com/discovery/deep-learning.html>

Referencias

- [Agg15] Aggarwal, C. C. (2015). Data Classification. Algorithms and Applications. *Chapman and Hall/CRC*.
- [Mai10] Maimon, O., & Rokach, L. (Eds.). (2010). Data mining and knowledge discovery handbook, 2nd edition. *Springer*.
- [Wit11] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: practical machine learning tools and techniques, 3rd edition. *Morgan Kaufmann*.