Clasificación bayesiana

UCO ONLINE

Clasificación bayesiana

Introducción

Los métodos de clasificación probabilísticos usan inferencia estadística para encontrar la mejor clase para una instancia dada. Además, estos métodos no solo ofrecen como salida la clase asignada, sino que proporcionan una probabilidad *a posteriori* $p(C_k|\mathbf{x}_i)$ de pertenencia a una clase C_k dadas las características de un patrón $\mathbf{x}_i$. La probabilidad a posteriori se define como la probabilidad tras observar las características de la instancia. Por otro lado, la probabilidad *a priori* corresponde simplemente con la fracción de instancias de entrenamiento que pertenecen a cada una de las clases, sin tener aún ninguna información de la instancia de test. Por lo general, la clase con mayor probabilidad a posteriori es la escogida como la mejor.

Existen dos métodos principales para estimar dichas probabilidades a posteriori. En primer lugar, se basa en determinar las probabilidades condicionales de la clase $p(\mathbf{x}_i|C_k)$ para cada una de las clases individualmente, e inferir la probabilidad a priori $p(C_k)$ por separado. Posteriormente, utilizando el Teorema de Bayes, se pueden obtener las probabilidades a posteriori $p(C_k|\mathbf{x}_i)$. Como las probabilidades condicionales de clase definen el proceso estadístico que generan las características medidas, estos enfoques que modelan la distribución de la entrada y la salida se llaman *modelos generativos*. Si los datos observados estuvieran realmente obtenidos de dicho modelo generativo, el aprendizaje de los parámetros del modelo maximizando la verosimilitud de los datos es un método común. El clasificador Naïve Bayes, que se verá más adelante en esta sección, es un ejemplo de este tipo de métodos.

Por otro lado, también se pueden modelar directamente las probabilidades a posteriori aprendiendo una función discriminante que mapee la entrada $\mathbf{x}_i$ en las etiquetas de clase C_k. Estos métodos se conocen comúnmente como *modelos discriminativos*. Uno de estos métodos es la Regresión logística, vista en lecciones anteriores.

Teorema de Bayes

El método de Naïve Bayes se basa en el Teorema de Bayes, que se introduce en esta sección. Antes de presentar dicho teorema, se revisan dos reglas fundamentales de teoría de la probabilidad, considerando que $p(X,Y)$ es una probabilidad conjunta (ambas cosas ocurren a la vez), $p(Y|X)$ es la probabilidad condicional de que ocurra Y si ocurre X, y $p(X)$ es una probabilidad marginal.

- Regla de la suma: $p(X) = \sum_Y p(X,Y)$
- Regla del producto: $p(X,Y) = p(Y|X)p(X)$

Basándonos en ambas reglas, y conociendo la propiedad simétrica $p(X,Y) = p(Y,X)$, es fácil obtener la siguiente regla del teorema de Bayes:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$

Consideremos un ejemplo básico para entender los conceptos básicos de la teoría de probabilidad y el teorema de Bayes. Supongamos dos cajas, una roja y una blanca; en la caja roja hay dos manzanas, cuatro limones, y seis naranjas; mientras que en la caja blanca hay tres manzanas, seis limones, y una naranja. Supongamos que escogemos al azar una de las cajas, y de su interior, seleccionamos un ítem al azar, lo observamos, y volvemos a dejarlo en su lugar. Supongamos que repetimos varias veces el experimento, y que al azar se ha seleccionado la caja roja el 40% de las veces y la blanca el 60%, y que todos los elementos de la caja tienen la misma probabilidad de ser seleccionados.

Se define una variable aleatoria Y para indicar la caja escogida, por lo que tendríamos $p(Y = r) = {}^{4}/_{10}$ para la caja roja, y $p(Y = b) = {}^{6}/_{10}$ para la caja blanca.

Las probabilidades condicionales de seleccionar cada uno de los elementos, dada una de las cajas se pueden describir como a continuación, donde por ejemplo $p(X = m|Y = r)$ indica la probabilidad de escoger una manzana ($X = m$) dado que la caja es la roja. Dicha probabilidad es de $2/12$ porque de los 12 elementos de la caja, 2 son manzanas, y los conocemos. Nótese que las probabilidades están normalizadas de modo que la suma de las probabilidades condicionales para cada caja es 1.

- $p(X = m|Y = r) = 2/12$
- $p(X = l|Y = r) = 4/12$
- $p(X = n|Y = r) = 6/12$
- $p(X = m|Y = b) = 3/10$
- $p(X = l|Y = b) = 6/10$
- $p(X = n|Y = b) = 1/10$

Ahora supongamos que se selecciona un ítem y es una naranja, y queremos saber de qué caja se obtuvo. Para ello, es necesario evaluar la distribución de probabilidad de las cajas condicionado a la identidad del ítem obtenido, al contrario de las ecuaciones inmediatamente anteriores, que

describen la distribución del ítem condicionado a la caja. Basado en el teorema de Bayes, podemos obtener la probabilidad a posteriori como:

$$p(Y = r|X = n) = \frac{p(X = n|Y = r)p(Y = r)}{p(X = n|Y = r)p(Y = r) + p(X = n|Y = b)p(Y = b)}$$
$$= \frac{(^6/_{12}) * (^4/_{10})}{(^6/_{12}) * (^4/_{10}) + (^1/_{10}) * (^6/_{10})} = 0.769$$

$$p(Y = b|X = n) = \frac{(^1/_{10}) * (^6/_{10})}{(^6/_{12}) * (^4/_{10}) + (^1/_{10}) * (^6/_{10})} = 0.231$$

Por lo general, estamos interesados en obtener la probabilidad de pertenencia a una clase dados los datos de entrada. Por tanto, supongamos que utilizamos una variable aleatoria Y para determinar la etiqueta de clase para las instancias de los datos, y una variable aleatoria X para representar las características de dichas instancias. Podemos interpretar $p(Y = C_k)$ como la probabilidad a priori para la clase C_k, que representa la probabilidad de que un patrón pertenezca a dicha clase antes de conocer sus características. En este sentido, el valor de $p(X|Y)$ expresa cuán probable es el dato observado X para la clase Y, también llamado verosimilitud (*likelihood*). Nótese que dicha verosimilitud no es una distribución de probabilidad sobre Y, ni la suma (o integral) de sus valores sobre Y suma 1.

Naïve Bayes

El clasificador Naïve Bayes se conoce como el clasificador Bayesiano más simple, convirtiéndose en un importante modelo probabilístico, alcanzando gran éxito en la práctica pese a sus fuertes asunciones de independencia.

En primer lugar, definamos el problema como a continuación. Supongamos que tenemos un conjunto de entrenamiento $\{(x^{(i)}, y^{(i)})\}$ formado por N instancias, siendo cada $x^{(i)}$ un vector de características d-dimensional, y cada $y^{(i)}$ denota la etiqueta de clase para dicho ejemplo. Asumimos variables aleatorias Y y X con componentes $X_1, \dots, X_d$ correspondientes a la etiqueta de clase y y el vector de características $x = \langle x_1, \dots, x_d \rangle$. Nótese que se utilizan superíndices para indicar las instancias, y los subíndices para referirse a cada característica o variable aleatoria de un vector. Además, Y es una variable discreta que puede tomar un valor de entre $\{C_1, \dots, C_k\}$.

La tarea es construir un clasificador cuya salida sea la probabilidad a posteriori $p(Y|X)$ para los distintos posibles valores de Y. De acuerdo con el teorema de Bayes, la probabilidad a posteriori $p(Y = C_k|X = x)$ se puede representar como:

$$p(Y = C_k|X = x) = \frac{p(X = x|Y = C_k)p(Y = C_k)}{p(X = x)} = \frac{p(X_1 = x_1, \dots, X_d = x_d|Y = C_k)p(Y = C_k)}{p(X_1 = x_1, \dots, X_d = x_d = x)}$$

Un modo de aprender las probabilidades a posteriori es usar los datos de entrenamiento para estimar tanto $p(X|Y)$ como $p(Y)$. Así, se podrán utilizar dichas estimaciones junto con el teorema de Bayes para determinar $p(Y|X = x^{(i)})$ para una nueva instancia.

Por lo general, aprender un modelo Bayesiano exacto es intratable. Considerando una variable objetivo binaria, y un conjunto de d atributos de entrada, se necesitarían estimar aproximadamente 2^d parámetros. Para tratar el problema de una manera factible, el clasificador Naïve Bayes reduce la complejidad haciendo una asunción de independencia condicional de que las características $X_1, \dots, X_d$ son todas condicionalmente independientes unas de otras, dada Y. Así, se reduce el número de parámetros a aprender para modelar $p(X|Y)$, de $2(2^d - 1)$ a solamente $2d$. Considerando la verosimilitud $p(X|Y)$, tendríamos:

$$p(X_1 = x_1, \dots, X_d = x\_d|Y = C_k) = \prod_{j=1}^{d} p\left(X_j = x_j|X_1 = x_1, \dots, X_{j-1} = x_{j-1}, Y = C_k\right)$$
$$= \prod_{j=1}^{d} p\left(X_j = x_j|Y = C_k\right)$$

En la anterior fórmula, la forma final se obtiene tomando la sunción de Naïve Bayes, es decir, que el valor de la variable aleatoria X_j es independiente del resto de características, pero que sigue condicionado al valor de la etiqueta de clase. Esta asunción es relativamente fuerte pero muy útil. De este modo, la probabilidad a posteriori dada por el clasificador Naïve Bayes se puede obtener como:

$$p(Y = C_k|X_1, \dots, X_d) = \frac{p(Y = C_k)\prod_j p\left(X_j|Y = C_k\right)}{\sum_i p(Y = C_i)\prod_j p\left(X_j|Y = C_i\right)}$$

Estimador de máxima verosimilitud para Naïve Bayes

En la mayoría de las aplicaciones prácticas, para la estimación de parámetros de Naïve Bayes se utiliza el método de máxima verosimilitud. En resumen, un modelo Naïve Bayes necesita estimar dos tipos de parámetros. El primero es:

$$\pi_k \equiv p(Y = C_k)$$

para cada posible clase o valor de C_k. Este parámetro se puede interpretar como la probabilidad de observar la clase C_k, teniendo las restricciones $\pi_k \geq 0$ y $\sum_{k=1}^{K} \pi_K = 1$. Nótese que hay K de estos parámetros a aprender, siendo $K - 1$ independientes (y el restante puede calcularse a partir de los demás).

Para los d atributos de entrada X_i, supongamos que cada uno puede tomar J posibles valores discretos, y utilizamos $X_i = x_{ij}$ para indicar cada posible valor. El segundo tipo de parámetro a estimar es:

$$\theta_{ijk} \equiv p(X_i = x_{ij} | Y = C_k)$$

para cada atributo de entrada X_i, cada posible valor x_{ij}, y cada posible valor C_k. El valor θ_{ijk} puede interpretarse como la probabilidad de que el atributo X_i tome el valor x_{ij}, condicionado a que la instancia sea de la clase C_k. Nótese que se debe satisfacer la condición $\sum_j \theta_{ijk} = 1$ para cada par de valores i, k, y debe hacer $d \cdot J \cdot K$ de estos parámetros, siendo independientes únicamente $d \cdot (J - 1) \cdot K$ de ellos.

Todos estos parámetros pueden ser estimados mediante el método de máxima verosimilitud, calculando las frecuencias relativas de los diferentes eventos en los datos. Las estimaciones de máxima verosimilitud para θ_{ijk} dado un conjunto de datos de entrenamiento son:

$$\theta_{ijk} = \hat{p}(X_i = x_{ij} | Y = C_k) = \frac{contar(X_i = x_{ij} \wedge Y = C_k)}{contar(Y = C_k)}$$

que no es más que el ratio de instancias de la clase C_k cuya característica X_i coincide con el valor x_{ij}. Para evitar aquellos casos en los que la condición propuesta en el numerador no aparece en ningún dato de entrenamiento, se suele utilizar un estimador suavizado como se propone a continuación, donde J es el número de valores distintos que puede tomar X_i, y l la fuerza del suavizado. Si se utiliza $l = 1$, se llama estimador suavizado de Laplace.

$$\theta_{ijk} = \hat{p}(X_i = x_{ij} | Y = C_k) = \frac{contar(X_i = x_{ij} \wedge Y = C_k) + l}{contar(Y = C_k) + l \cdot J}$$

Por último, el estimador de máxima verosimilitud para π_k se muestra a continuación, incluyendo también el estimador suavizado. Nótese que el estimador calcula el ratio de instancias de

entrenamiento que pertenecen a una clase concreta (siendo N el número total de instancias de entrenamiento).

$$\pi_k = p(Y = C_k) = \frac{contar(Y = C_k)}{N}$$

$$\pi_k = p(Y = C_k) = \frac{contar(Y = C_k) + l}{N + l \cdot K}$$

Redes Bayesianas

Las redes bayesianas son modelos estadísticos fundamentados teoréticamente para representar las distribuciones de probabilidad de manera concisa y comprensible de manera gráfica. Se dibujan como una red de nodos, uno por cada atributo, conectadas por aristas dirigidas de tal manera que no existan ciclos; es decir, formando un grafo dirigido acíclico.

En los ejemplos expuestos a continuación, se asumirá que los atributos son siempre nominales, y que no existen valores perdidos en ellos. Aunque se puede aprender redes bayesianas sin la necesidad de realizar estas asunciones, no se va a entrar en tal nivel de detalle en esta lección.

En la Figura 1 se muestra una red bayesiana para el conjunto de datos del tiempo atmosférico. Como se observa, tiene un nodo para cada uno de sus 4 atributos: *Outlook*, *temperature*, *humidity*, y *windy*, además de un atributos para la clase, *play*. La red tiene una arista que va desde el nodo *play* hasta cada uno de los otros nodos, además de otras ciertas dependencias entre nodos. Sin embargo, la estructura del grafo es solo la mitad de la red. La información en las tablas dentro de cada nodo define la distribución de probabilidad que se utiliza para predecir la probabilidad de las clases para cada instancia.

En general, estas tablas tienen a la izquierda una columna con los valores de cada nodo que apunta al nodo actual, mostrando las posibles combinaciones de valores de dichas columnas (el nodo *play*, al no tener "padres", no tiene valores a la izquierda). Cada fila de la tabla corresponde a una posible combinación de valores para las columnas de la izquierda, y cada celda de la tabla indica la probabilidad de que se de cada valor para los atributos del nodo actual, dada la combinación de valores a la izquierda. Cada fila define una distribución de probabilidad, por lo que la suma de todos los valores de una fila es siempre 1.

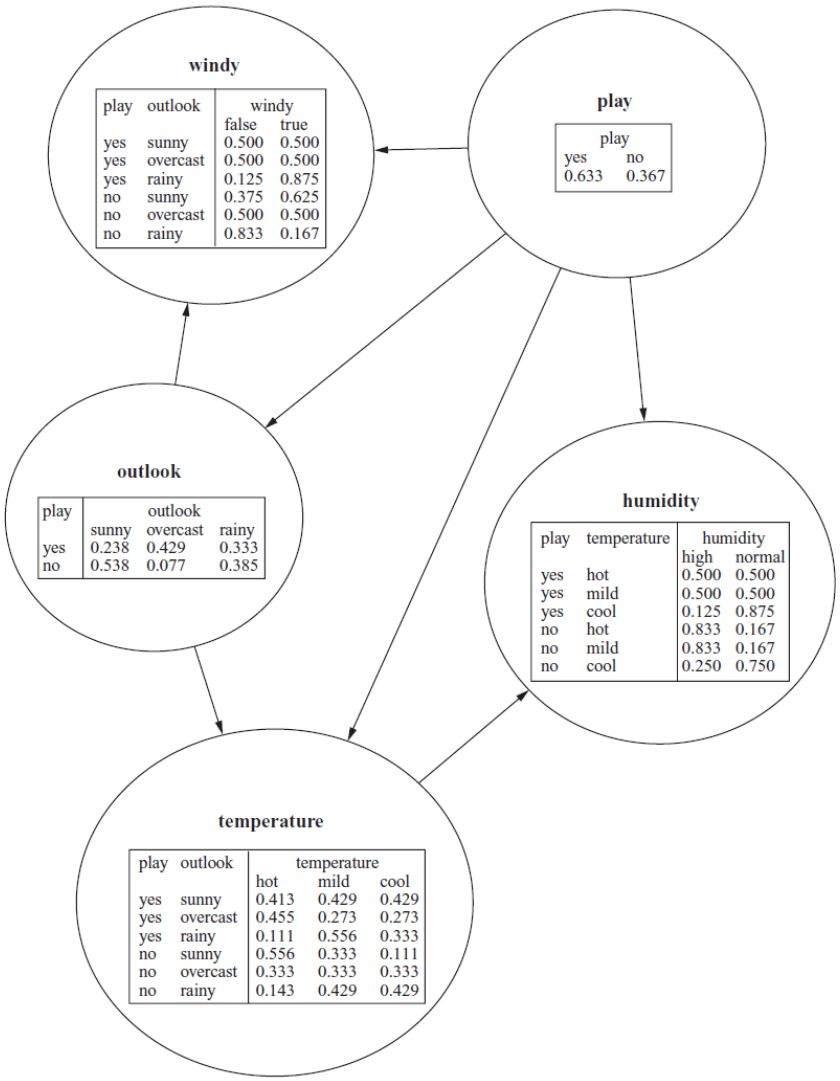

Figura 1. Ejemplo de red bayesiana.

Dado el ejemplo en la Figura 1, centrémonos en la primera fila de la tabla asociada al nodo *temperature*. En la parte izquierda tiene valores para cada atributo padre: *play* y *Outlook*; la parte derecha da la probabilidad para cada valor de temperatura. Por ejemplo, el primer valor (0.143) es la probabilidad de que la temperatura tome el valor *hot*, sabiendo que los atributos *play* y *Outlook* tienen los valores *yes* y *sunny*, respectivamente.

Utilizar una red bayesiana para proporcionar un valor de probabilidad para cada clase, dada una nueva instancia es muy sencillo. La nueva instancia especifica un valor para cada atributo. Para cada nodo en la red, se mira la probabilidad de dicho valor del atributo basado en los valores de los atributos de los padres. Después, simplemente se multiplican todas estas probabilidades.

Por ejemplo, consideremos una instancia con los valores: *outlook = rainy*, *temperature = cool*, *humidity = high*, y *windy = true*. Para calcular la probabilidad de *play = no*, observamos en la red que da la probabilidad 0.367 en el nodo *play*, 0.385 en *outlook*, 0.429 en *temperature*, 0.250 en *humidity*, y 0.167 en *windy*. El resultado de multiplicar todos estos valores es 0.0025. El mismo cálculo, para la probabilidad de *play = yes* devuelve 0.0077. Si normalizamos estos valores (dividiendo cada uno por la suma de ambos), tenemos las siguientes probabilidades, considerando X_i la instancia a predecir:

$$P(play = no|X_i) = \frac{0.0025}{0.0025 + 0.0077} = 0.245$$

$$P(play = yes|X_i) = \frac{0.0077}{0.0025 + 0.0077} = 0.755$$

Las redes bayesianas pueden construirse por expertos, o aplicando métodos de *machine learning* que estudian las correlaciones entre atributos para obtener la estructura final de la red. No se entrará en tanto detalle en esta lección, pero puede consultarse la sección 6.7 de [Wit11].

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*.