0pt 2pt

Maestría en Ciencia de los Datos y Analítica
Universidad EAFIT
Estadstica en Analítica
Medellín, abril 23 - mayo 15 de 2021

Universidad EAFIT

Jose Antonio Solano Atehortúa[a], [b]

Director científico, as research, Medellín, Colombia

MÓDULO DE CLASIFICACIÓN

Es notable que una ciencia que comenzó con consideraciones sobre juegos de azar haya llegado a ser el objeto más importante del conocimiento humano.

Comprender y estudiar el azar es indispensable, porque la probabilidad es un soporte necesario para tomar decisiones en cualquier ámbito Pierre-Simon Laplace.

1. Apuntes de clase

El reconocimiento de patrones o la discriminación consiste en predecir la naturaleza desconocida de una observación, en términos de una cantidad discreta como blanco o negro, uno o cero, enfermo o sano, real o falso.

Una observación es una colección de medidas numéricas, como una imagen, un vector de datos meteorológicos, o un electrocardiograma. Más formalmente, una observación es un vector n−dimensional X. La naturaleza desconocida de la observación se llama clase. Se denota con Y y toma valores en un conjunto finito $\{1, 2, \ldots, M\}$. En el reconocimiento de patrones, uno crea una función $f(X) : \mathbb{R}^n \to \{1, \ldots, M\}$ que representa la conjetura de Y dado X. El mapeo f se llama clasificador. El clasificador se equivoca en X si $f(X) \neq Y$.

Especificar f para un X particular necesita del conocimiento propio y de la experiencia en la rama en donde fueron tomados los datos de las medidas (medicina, imágenes, etc). Un marco teórico aceptado es introducir un entorno probabilístico de la situación: sea (X, Y) un par aleatorio $\mathbb{R}^n \times \{1, 2, \ldots, M\}-$valuado.

La distribución de (X, Y) en la práctica describe la frecuencia con la que se encuentran parejas particulares. Un error ocurre cuando $f(X) \neq Y$, y la *probabilidad de error* para un clasificador f es

$$L(f) = \mathbb{P}\{f(X) \neq Y\} \tag{1}$$

Existe el mejor clasificador posible, f^*, el cual se define como

$$f^* = \arg \min_{f:\mathbb{R}^n \to \{1,\ldots,M\}} \mathbb{P}\{f(X) \neq Y\} \tag{2}$$

[a]E-mail: jasolanoa@eafit.edu.co
[b]E-mail: info@asresearch.co

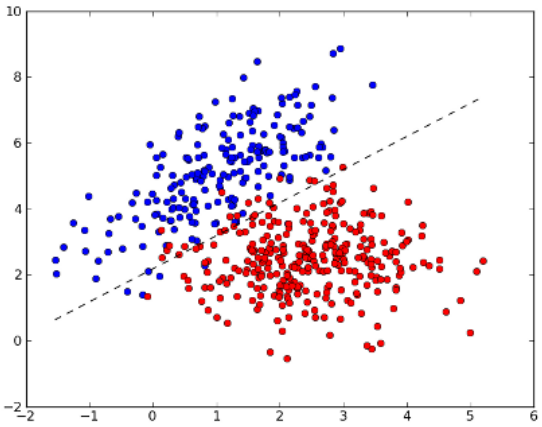

FIGURA 1: Clasificación

Observe que f^* depende de la distribución de (X, Y). Si se conoce esta distribución, se puede calcular f^*. El problema de encontrar f^* se llama el *problema de Bayes*, y el clasificador f^* se llama el *clasificador de Bayes* (o la *regla de Bayes*). La mínima probabilidad de error se llama el *error de Bayes* y se denota por $L^* = L(f^*)$. Por lo general la distribución de (X, Y) es desconocida, asi que f^* también es desconocida.

Aunque $f*$ es desconocida, se tiene acceso a una base de datos de parejas (X, Y), $1 \leq i \leq m$, observadas en el pasado. Esta base de datos puede ser el resultado de observaciones en la realización de experimentos (como por ejemplo datos metereológicos, ...). También pueden se pueden obtener por medio de un experto o profesor quien completó los Y_is depués de haber visto los X_is. Para encontrar un clasificador f con una probabilidad de error baja es necesario asumir algún tipo de comportamiento conjunto del par (X_i, Y_i).

Una supuesto en este capitulo es que $(X_1, Y_1), \ldots, (X_m, Y_m)$, Los *datos*, forman una sucesión de parejas aleatorias independientes identicamente distribuidas (*i.i.d.*) que tienen la misma distribución de la pareja (X, Y).

Un clasificador se construye con base a $X_1, Y_1, \ldots, X_m, Y_m$ y se denota por f_n. Y se 'adivina' con la expresión $f_n(X; X_1, Y_1, \ldots, X_n, Y_n)$. El proceso de construir f_n se llama *aprendizaje (learning)*, o *aprendizaje supervisado (supervised learning)*, o *aprendizaje con un profesor (learning with a teacher)*.

El desempeño de f_n se mide con la *probabilidad de error* condicional.

$$L_m = L(f_m) = \mathbb{P}\{f_n(X; X_1, Y_1, \ldots, X_m, Y_m) \neq Y | X_1, Y_1, \ldots, X_m, Y_m\}.$$

Esta es una variable aleatoria porque depende de los datos. O sea, L_m pondera sobre la distribución de (X, Y), pero los datos se mantienen fijos. Es decir, en una aplicación particular a uno le toca trabajar con los datos realizados. Y en este sentido sería útil saber el número $\mathbb{E}[L_m]$ porque este número indica la calidad del promedio de una sucesión de datos, no de la sucesión particular de los datos.

Es así que el foco de L_m, se centra en la probabilidad condicional del error.

Un mapeo individual

$$f_m : \mathbb{R}^n \times \{\mathbb{R}^n \times \{1, 2, \ldots, M\}\}^m \to \{1, 2, \ldots, M\}$$

se sigue llamando un *clasificador*. Una sucesión $\{f_m, m \geq 1\}$ se llama una *regla* (*discriminación*). Por lo tanto, los clasificadores son funciones, y las reglas son sucesiones de funciones.

Una regla es consistente si

$$\lim_{m \to \infty} \mathbb{E}[L_n] = L^*.$$

Una regla de clasificación consistente garantiza que se pueden usar cada vez más muestras para reconstruir la distribución desconocida de (X, Y) porque L_m se puede tomar tan cercano de L^* como se necesite.

En otras palabras, se puede obtener una cantidad infinita de información a partir de muestras finitas.

La consistencia de una regla no impone condiciones sobre la distribución de (X, Y) ya que estas condiciones no se pueden verificar. Si una regla es consistente para todas las distribuciones de (X, Y), se dice que es *universalmente consistente*.

Hasta el año 1977 no se sabia si existia una regla universalmente consistente. En este año Stone mostró que la regla de clasificaión construida tomando cualesquiera $k-$vecinos más cercanos con $k = k(n) \to \infty$ y $k/n \to \infty$. Y construyendo el clasificador $f_m(x)$ tomando la mayoría de votos sobre los Y_i's en el subconjunto de los k pares (X_i, Y_i) de $(X_1, Y_1), \ldots, (X_m, Y_m)$ que tienen los valores más pequeños para $||X_i - x||$ (o sea, los X_i más cercanos a x). Desde la prueba de la consistencia universal de la regla $k-$vecinos más cercanos de Stone han surgido varias reglas de clasificación para las que se ha probado que son universalmente consistentes también.

El concepto de consistencia recae sobre la convergencia y se puede preguntar que tipo de convergencia se tiene: convergencia con probabilidad uno? convergencia en probabilidad? Esta es la tarea del probabilista. Por otro lado, existen la desigualdades de concentración del tipo de McDiarmid, las cuales proporcionan elementos necesarios y suficientes que pueden probar la equivalencia entre varios tipos de convergencia en espacios de medida.

Por ejemplo, para la regla de $k-$vecinos más cercanos, existe un número $c > 0$, tal que para todo $\epsilon > 0$, existe $N(\epsilon) > 0$ dependiente de la distribución de (X, Y), tal que

$$\mathbb{P}\{L_m - L^* > \epsilon\} \leq e^{-cn\epsilon^2} \quad , \quad n \geq N(\epsilon).$$

Como puede observarse esta desigualdad proporciona una cota para el desvio de L_m con respecto a L^*. Este tipo de desigualdades son conocidas como Desigualdades de concentración.

Esto sugiere uno de los objetivos principales en el estudio teórico de los problemas de clasificación: estimar L_m con los datos proporcionados.

2. Calentamiento

Para comenzar te sugiero que hagas los cálculos en detalle para comprender el funcionamiento del método. La actividad consiste en construir un clasificador usando la técnica de k vecinos más cercanos usando como conjunto de entrenamiento los patrones descritos en la siguiente matriz X con etiquetas dadas en el vector Y.

$$X = \begin{bmatrix} 0.8 & 0.8 \\ 0.8 & 1.2 \\ 3.8 & 2.8 \\ 4.2 & 3.2 \\ 1.0 & 1.0 \\ 1.2 & 1.2 \\ 4.2 & 2.8 \\ 4.4 & 2.8 \\ 3.5 & 1.0 \\ 4.0 & 1.0 \\ 3.8 & 0.5 \\ 4.0 & 0.7 \end{bmatrix} \quad , \quad Y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix}$$

1. Calcule la precisión en el conjunto de entrenamiento.

2. Calcule la predicción para el punto de testeo $P = (3.0, 2.0)$.

3. Experimenta con qué valor de k se obtinen la mejor precisión.

4. Vecinos más cercanos con distancia ponderada: este es un algoritmo similar a kNN, la única diferencia es que estos k vecinos más cercanos se ponderan de acuerdo con su distancia desde el punto de prueba. Cada uno de los vecinos está asociado con el peso w que se define como

$$w_j = \begin{cases} \frac{d_k - d_j}{d_k - d_1} & \text{si } d_k \neq d_1 \\ 1 & \text{si } d_k = d_1 \end{cases}$$

donde $j = 1, 2, ..., k$. Después de calcular los w_j, el algoritmo asigna al patrón de testeo P la clase para la cual la suma de las ponderaciones de los k vecinos más próximos sea la mayor.

Testear con el punto $P(3, 2)$ tomando $k = 1$ y $k = 5$.

3. Ejercicio de implementación

1. Implementar el algoritmo de clasificación KNN. Valide con los datos del problema 1 (el de calentamiento).

2. Para este ejercicio use la base de datos *DatosFisher.xlsx* disponible en la página del curso.

3. Seleccione el k para el cual el error de entrenamiento sea mínimo. $k =$?

4. Ejercicio de implementación

1. Construir un script para usar el algoritmo de árboles de decisión y ajustar un modelo de clasificación para los datos *DatosFisher.xlsx*.

2. Construya la gráfica donde se aprecie el error de entrenamiento y el error de testeo para diferentes tamaños del conjunto de entrenamiento.

3. Puede plantear alguna hipótesis sobre el comportamiento de estos dos errores a medida que el tamaño del conjunto de entrenamiento aumenta?

5. Descripción de datos multivariantes

Tener en cuenta que la matriz de datos $X_{m \times n}$ tiene m filas y n columnas. Las filas de la matriz se denotan por $X^{(1)}, X^{(2)}, ..., X^{(m)}$, y las columnas se denotan por $X_1, X_2, ..., X_n$. Las filas represnetan los ejemplos y las columnas representan las variables medidas en cada ejemplo.

Para los datos del archivo *ELE.xlsx* hacer un análisis descriptivo de datos multivariantes usando como guía la siguiente lista de actividades:

En cada caso se puede usar una función de cualquier paquete de software pero también se puede construir la función con la fórmula dada en clase y en este documento.

1. Estudiar el tipo de variable de manera univariante. Es decir, determine los descriptivos principales tales como media, curtosis, asimetría, etc . . . , de cada variable.

2. Calcular las medidas de centralización: el vector de medias.

   Usar un paquete estadístico o usar la fórmula:

$$\overline{X} = \frac{1}{m} X^t \mathbb{1}$$

donde $\mathbb{1}$ es un vector de unos de la dimensión adecuada y X^t es la matriz transpuesta de X.

3. Calcular la matriz de varianzas y covarianzas.

   Usar un paquete estadístico o usar la fórmula:

   $$S = \frac{1}{m} X^t P X$$

   donde la matriz cuadrada P está definida como $P = I - \frac{1}{m}\mathbb{1}\mathbb{1}^t$

   Otra expresión para la matriz de varianzas y covarianzas es a partir de la matriz de datos centrados $\widetilde{X}$, así:

   $$S = \frac{1}{m} \widetilde{X}^t \widetilde{X}$$

   donde $\widetilde{X} = X - \mathbb{1}\widetilde{X}^t$ es la matriz de datos centrados que se obtiene de restar a cada dato su media.

   Calcular la matriz de varianzas corregida.

   Consiste en dividir la matriz de covarianzas por $(m-1)$ en lugar de m para tener un estimador insesgado de la matriz de la población.

   $$\widehat{S} = \frac{1}{(m-1)} \widetilde{X}^t \widetilde{X}$$

4. Centralizar los datos y calcular la matriz de varianzas y covarianzas.

5. Calcular la variabilidad total y la varianza promedio.

   La variabilidad total es la trasa de la matriz de varianzas y covarianzas.

   Para la varianza total puede usar la fórmula:

   $$T = tr(S) = \sum_{i=1}^{n} s_{ii}^2$$

   y para la varianza promedio

   $$\bar{s}^2 = \frac{n}{n} \sum_{i=1}^{n}$$

6. Calcule la varianza generalizada y la variabilidad promedio.

   La varianza generalizada es el determinante de la matriz de covarianzas. Es decir,

   $$VG = |S|$$

   Su raíz cuadrada se denomina *desviación típica generalizada.*

   Para el cáculo de la variabilidad promedio se usa la fórmula:

   $$VP = |S|^{\frac{1}{n}}$$

7. Distancia de Mahalanobis Se define la distancia de Mahalanobis entre un punto y su vector de medias como

   $$d_i = [(X_i - \overline{X})^t S^{-1} (X_i - \overline{X})]^{1/2}.$$

6. Análisis de dependencias lineales

Se denomina matriz de precisión a la inversa de la matriz de varianzas y covarianzas. Un resultado importante es que la matriz de precisión contiene la información sobre la relación multivariante entre cada una de las variables y el resto.

Uno de los objetivos más importantes de la descripción de los datos multivariantes es comprender la estructura de dependencias entre las variables.

Para esta actividad tome el archivo de datos *ELE.xlsx* disponible en la página web del curso.

1. Analizar la correlación entre pares de variables.
2. Análizar dependecia entre una variable y todas las demás. En este caso encontrar las betas de la regresión múltiple de la variable que mejor se puede explicar a partir de las demás.
3. Analizar la correlacion entre pares de variables pero eliminando el efecto de las demás variables.
4. Analizar el conjunto completo de todas las variables.

7. Problema de clasificación

Use el método de su escogencia para entrenar un clasificador con los datos del archivo *ELE2.xlsx*.

Considere eliminar variables de acuerdo al ejercicio anterior. Evalue la precisión y exponga el error de entrenamiento y de testeo.

8. Relaciones teóricas

Formalmente la regla k-NN se define como

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^{n} w_{n_i} I_{\{Y_i=1\}} > \sum_{i=1}^{n} w_{n_i} I_{\{Y_i=0\}} \\ 0 & \text{en otro caso,} \end{cases}$$

donde $w_{n_i} = 1/k$ si X_i está entre los k vecinos más próximos de x, y $w_{n_i} = 0$ en otro caso. Se dice que X_i es el k-ésimo vecino más próximo de x si la distancia $||x - X_i||$ es la k-ésima más pequeña entre $||x - X_1||, ..., ||x - X_n||$. La decisión sobre la clase se basa en la votación de la mayoría.

En esta actividad te puedes guiar por experimentos numéricos para comprobar que los resultados se cumplen. Cuando te hayas convencido intenta generalizar por medio de la demostración.

1. Demuestre que
$$L_{NN} \leq 2L^*$$
2. Demuestre que
$$L_{kNN} \leq L^* \left(1 + \sqrt{\frac{2}{k}}\right)$$