Subject: Probability and Statistics

Class XV: Discriminant Analysis

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

UdeA

Análisis Discriminante

Key Terms for Discriminant Analysis

Covariance

A measure of the extent to which one variable varies in concert with another (i.e., similar magnitude and direction).

Discriminant function

The function that, when applied to the predictor variables, maximizes the separation of the classes.

Discriminant weights

The scores that result from the application of the discriminant function, and are used to estimate probabilities of belonging to one class or another.

UdeA

Covarianza

La covarianza es una medida de la relación de dos variables, a través de su variabilidad.

Al igual que la correlación, valores positivos indican una relación directa, y negativos una relación inversa, pero la covarianza no está restringida a un valor de -1 a 1 como la correlación.

$$s_{x,z} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})}{n-1}$$

UdeA

Matriz de Covarianza

Al igual que en correlación, cuando se tienen varias variables es de utilidad construir una matriz de covarianza. En este caso la diagonal no sería 1, sino que sería la varianza la variable y los demás puntos serían la covarianza entre las variables. Igualmente es una matriz simétrica.

$$\hat{\Sigma} = \begin{bmatrix} s_x^2 & s_{x,z} \\ s_{x,z} & s_z^2 \end{bmatrix}$$

UdeA

Discriminante lineal de Fisher

Si bien hay variantes de los modelos discriminantes, el más conocido es este, aunque ha dejado de ser empelado con los años dada la aparición de modelos más sofisticados de clasificación. Sin embargo se revisa este caso porque es la base de otro métodos.

Tomemos el caso más simple, en donde queremos determinar una salida binaria a partir de dos variables continuas.
Técnicamente, se debe cumplir que los predictores sean normales y continuas, pero en la practica que observa que se desempeña bien incluso con variable con datos atípicos o variables binarias.

UdeA

Discriminante lineal de Fisher

Aquí se tiene dos conceptos importantes

Variabilidad entre grupos

¿Qué tanto varia un grupo de otro?

Variabilidad dentro de grupos

¿Qué tanto varían los datos dentro del grupo?

UdeA

Discriminante lineal de Fisher

Aquí se tiene dos conceptos importantes

La variabilidad entre grupos se estima como la suma de las distancias al cuadrado de las medias de los grupos. (SS_{entre})

La variabilidad dentro de grupos se determina como la dispersión de los datos del grupo respecto de su media, ponderado respecto de la matriz de covarianza. (SS_{dentro})

UdeA

Discriminante lineal de Fisher

Entonces por ejemplo, si se quiere hacer cierta clasificación usando dos variables, x y z, entonces el método del discriminante lineal de Fisher buscan un modelo lineal de la forma $w_x x + w_z z$, tal que se maximice la relación de las sumas de cuadrados dentro y entre grupos:

$$\frac{SS_{entre}}{SS_{dentro}}$$

UdeA

Regresión Logística

Este es un tipo de regresión en la cual la variable dependiente es binaria. Por este motivo este tipo de regresión es un método de clasificación más que de regresión.

Veamos la deducción de los modelos de regresión logística.

Lo primero es que la salida del modelo (y) no solo como una variable binaria, sino también como la probabilidad p de que la etiqueta sea “1”, y esto podría modelarse linealmente:

$$p = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

UdeA

$$p = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Pero este modelo podría arrojar valores por fuera del rango de una probabilidad, que es entre 0 y 1.

Entonces se elije modelar p de la siguiente forma, usando lo que se conoce como una función logística, y así se aseguran valores entre 0 y 1:

$$p = \frac{1}{1 + e^{-(b_0 + b_1X_1 + \cdots + b_nX_n)}}$$

Ahora, sería de utilidad que la exponencial no estuviera en el denominador. Para esto se hace uso de un concepto conocido como la probabilidad relativa (*odds* en inglés)

$$odds(Y=1) = \frac{P(Y=1)}{P(Y=0)} = \frac{p}{1-p}$$

Entonces, con esta relación y la anterior:

$$odds(Y=1) = e^{b_0+b_1X_1+\cdots+b_nX_n}$$

UdeA

Ahora si tomamos el logaritmo natural a ambos lados:

$$Ln(odds(Y=1)) = Ln(e^{b_0+b_1X_1+\cdots+b_nX_n})$$

$$Ln(odds(Y=1)) = b_0 + b_1X_1 + \cdots + b_nX_n$$

Y esto se conoce como función Log-odds o función logit, y permite mapear probabilidades entre $(0,1)$ a valores entre $(-\infty, \infty)$.

Ahora para la selección de la clase entre 1 y 0, se establece un umbral de probabilidad para la clase 1, y cualquier caso con probabilidad de ser 1 por encima del umbral se le asigna la clase 1, de lo contrario la clase 0.

Ahora veamos como se utiliza este tipo de clasificación en Sklearn.

Usemos la base de datos Iris de sklearn:

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering