Minería de Datos para el Análisis de Big Data

Por: Carlos Carreño

ccarrenovi@Gmail.com

Abril,2021

Modulo 8 Análisis Multivariante

- Análisis de Componentes Principales – PCA
- Análisis Discriminante Lineal - LDA

PCA

- El análisis de componentes principales (***principal component analysis***) o PCA es una de las técnicas de aprendizaje no supervisado.
- Una de las aplicaciones de PCA es la **reducción de dimensionalidad** (variables), perdiendo la menor cantidad de información (varianza) posible.
- cuando contamos con un ***gran número de variables*** cuantitativas posiblemente correlacionadas (indicativo de existencia de información redundante), ***PCA permite reducirlas a un número menor*** de variables transformadas (**componentes principales**) que expliquen gran parte de la variabilidad en los datos.
- Cada dimensión o **componente principal** generada por PCA será una **combinación lineal de las variables originales**, y serán además independientes o no correlacionadas entre sí.
- Se utiliza para enfatizar la variación y sacar a relucir **patrones fuertes** en un conjunto de datos. A menudo se utiliza para hacer que los datos sean fáciles de **explorar y visualizar**.

Ejemplo: PCA

- Realizar un PCA de los resultados obtenidos en la competición de heptatlón femenino de los Juegos Olímpicos de Seúl (1988).

LDA

- El análisis discriminante lineal (LDA - ***linear discriminant analysis***) y el discriminante lineal relacionado de Fisher son métodos utilizados en la estadística, el reconocimiento de patrones y el aprendizaje automático para encontrar una ***combinación lineal de características*** que caracteriza o separa dos o más clases de objetos o eventos.
- La combinación resultante puede utilizarse como clasificador lineal o, más comúnmente, para la reducción de la dimensionalidad antes de su posterior clasificación.

...

- Es un **método alternativo más adecuado a la regresión logística** cuando la **variable cualitativa tiene más de dos niveles** ($K \geq 2$).
- LDA Supone también un modelo más estable cuando el **tamaño muestral n es pequeño** y la distribución de los predictores es aproximadamente normal en cada una de sus clases.
- El propósito del LDA es encontrar la **combinación lineal de las variables originales** que permita la mejor **separación entre grupos** de un set de datos.
- El LDA está basado en el **clasificador Bayesiano**.

Ejemplo: LDA

- Predecir si el rendimiento del combustible (gas mileage) de un automóvil es alto o bajo en función del resto de predictores del set de datos auto del paquete ISLR.

Preguntas

- Alguna pregunta?