Álgebra Lineal Computacional

Segundo Cuatrimestre 2023

Trabajo Práctico N° 2

Sistemas de Recomendación

Definiciones y cálculos auxiliares

Matriz de covarianza

En probabilidad y estadística, la covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.
La covarianza entre dos variables se calcula como:

$$cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

donde x_i e y_i son cada una de las ocurrencias respectivas, y $\overline{x}$ e $\overline{y}$ son los valores promedio observados. En particular coincide con la varianza cuando el cálculo se hacer respecto de la misma variable:

$$cov_{x,x} = var_x = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{n}$$

La matriz de covarianza es una matriz cuadrada que contiene la covarianza entre los elementos de un vector. Es decir, si el vector tiene las componentes, $v = (x, y, z)$, se tiene:

$$\sum = \begin{bmatrix} var_x & cov_{x,y} & cov_{x,z} \\ cov_{y,x} & var_y & cov_{y,z} \\ cov_{z,x} & cov_{z,y} & var_z \end{bmatrix}$$

Por simplicidad vamos a suponer que todas las observaciones son la población total.

Análisis de Componentes Principales

El análisis por componentes principales (PCA) es una conocida técnica de reducción de la dimensionalidad de un problema, cuyo objetivo es transformar un conjunto de datos de alta dimensión en un espacio de menor dimensión, tratando de conservar la mayor parte de la información [1][2]. PCA funciona identificando las direcciones que capturan la mayor variación en los datos y proyectando los datos en esas direcciones, que se denominan componentes principales.
Dado un problema en el cual se identifican n variables independientes y una variable dependiente, y suponiendo que se realizan m observaciones, podemos organizar los datos de cada observación en una fila y las distintas columnas serán las distintas variables. Por tanto la matriz de variables independientes X tendrá m filas por n columnas, y la matriz de variables dependientes Y tendrá m filas por 1 columna.
El método de PCA consiste en:

(1) Centralización de las variables: se resta a cada valor la media de la variable a la que pertenece. Con esto se consigue que todas las variables tengan media cero. Como el proceso de PCA identifica aquellas direcciones en las que la varianza es mayor, deben estar todas las variables escaladas para que sean comparables, sino aquellas variables cuya escala sea mayor dominarán al resto. Dadas las observaciones $\mathbf{X_j}$ donde:

$\mathbf{X_j} = (X_1, X_2, ..., X_n)_j = (X_{j1}, X_{j2}, ..., X_{jn})$, con j cantidad de observaciones.

Se obtiene x_{ji} como:

$x_{ji} = (X_{ji} - \overline{X_i})/s_i$, con i para cada una de las variables independientes.

Donde:

$\overline{X_i} = (\sum_{j=1}^{m} X_{ji})/m$

$s_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(X_j i - \overline{X_i})^2}$

(2) La matriz $\mathbf{X}$ estará formada por los x_{ji} y a partir de ella se construye la matriz de covarianza definida anteriormente. El resultado es una matriz cuadrada y simétrica con dimensión n x n, con n la cantidad de variables independientes. Llamaremos a esta matriz de covarianza $\mathbf{C}$.

(3) El paso siguiente es encontrar los autovalores y autovectores de la matriz de covarianza del ítem anterior. Sean $\mathbf{V}$ y $\mathbf{D}$ la matriz con los autovectores por columnas y la matriz diagonal con los autovalores en la diagonal ordenados de mayor a menor, se tiene que:

$\mathbf{V^{-1} \cdot C \cdot V = D}$

Como interesa las principales componentes, que se corresponden con las máximas variaciones, sólo interesa encontrar el autovector que corresponde al máximo autovalor, y los siguientes en orden decreciente, sin la necesidad de hallar todos. Por tanto, se utiliza el método de la potencia y el método de la potencia con desplazamiento [3][4], para encontrar sucesivamente aquellas direcciones más relevantes de acuerdo a la necesidad.

(4) Una vez obtenidos los autovectores (componentes principales) se calcula el valor que toma cada componente para cada observación en función de las variables originales. Es decir, sea $\mathbf{W}$ la matriz que se obtiene a partir de las primeras 2 columnas de $\mathbf{V}$ (como ejemplo se toman 2, pero pueden ser de 1 a n la cantidad de autovectores), los puntos proyectados son las filas de la matriz $\mathbf{T}$ al realizar la siguiente operación:

$\mathbf{T = X \cdot W}$

La primer columna de $\mathbf{T}$ es la proyección de los puntos sobre la primer componente principal, y la segunda columna es la proyección sobre la segunda componente principal.

Finalmente lo que se obtiene es un conjunto de datos $\mathbf{T}$ que explica la variable dependiente $\mathbf{Y}$ a partir de una combinación lineal de los datos originales donde solamente se consideran los primeros dos términos para cada observación.

Cuando aplicamos PCA, es importante comprender qué parte de la variación de los datos explica cada componente principal. Una forma de medirlo es el concepto de "varianza explicada". La varianza explicada mide la proporción de varianza de los datos que explica cada componente principal. Y se calcula como la división entre cada uno de los autovalores y la suma de los mismos. Digamos que hay N autovectores, entonces la varianza explicada de cada autovector (componente principal) será: $Varianza\_explicada = \lambda_i/(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$

Evaluación del modelo de predicción

El objetivo original es poder asignarle una clasificación en 1, 2 o 3 a un vino según los gustos de los clientes, a partir de un conjunto de características o propiedades.
Para ello se cuenta con el conjunto de datos wine.csv que tiene el registro de 178 vinos, con la valoración de 13 características o propiedades (variables independientes) y la clasificación a la que pertenece (variable dependiente).
La idea de usar el método de PCA es encontrar la posibilidad de describir cada una de las categorías de clasificación con una cantidad menor de variables (1, 2 , 3 o 4, y no las 13 originales).
Una vez encontrada esta combinación, usarla para estimar un nuevo conjunto de datos y comparar el resultado obtenido con la clasificación dada.
Por tanto se debe dividir en dos el conjunto de datos con los que se cuenta, uno será el conjunto de entrenamiento ($\mathbf{Y_{train}}, \mathbf{X_{train}}$) (de alrededor del 80-90 % de los datos) y el otro el conjunto de testeo ($\mathbf{Y_{test}}, \mathbf{X_{test}}$) con los datos restantes. Es importante tomar en forma aleatoria cada uno de los conjuntos, notar que el conjunto de datos suministrado se encuentra ordenado por $'Segmento\_Cliente'$.
Con el método de PCA de la sección anterior podemos obtener los puntos proyectados en los componentes principales a partir del conjunto de entrenamiento:
$\mathbf{Y_{train}}$, $\mathbf{T_{train}} = \mathbf{X_{train}} \cdot \mathbf{W_{pca}}$
Para luego evaluar el conjunto de testeo:
$\mathbf{Y_{test}}$, $\mathbf{T_{test}} = \mathbf{X_{test}} \cdot \mathbf{W_{pca}}$
La matriz $\mathbf{W_{pca}}$ se calcula en función de cuántos componentes principales se desean emplear (en el ejercicio 2(a) se pide para 1 a 4 componentes principales), y se utiliza el mismo tanto en el conjunto de entrenamiento como en el de testeo. Es decir que se obtendrán 4 experimentos con 4 $\mathbf{W_{pca}}$, que claramente tienen distintas dimensiones.
Resta comparar $\mathbf{Y_{test}}$ con $\mathbf{Y_{estimado}}$ para cada uno de los experimentos.
$\mathbf{Y_{test}}$: son las clasificaciones esperadas dados los datos $\mathbf{X_{test}}$ que son conocidos.
$\mathbf{Y_{estimado}}$ son las clasificaciones encontradas por el método de primeros vecinos (KNN) a partir $\mathbf{T_{test}}$ y del conjunto de entrenamiento ($\mathbf{Y_{train}}, \mathbf{T_{train}}$)
En la siguiente sección se explica el método de primeros vecinos para comparar los resultados.

Algoritmo de primeros vecinos o KNN (k-Nearest Neighbors)

El algoritmo de primeros vecinos (k-Nearest Neighbors o KNN) es un método sencillo que permite encontrar los primeros k puntos vecinos respecto de un punto en particular, y en función de eso otorgarle un valor de acuerdo al valor que tienen la mayoría de sus k vecinos.
Dicho de otra forma, cuando se busca la clasificación para un nuevo registro u observación, se localizan primero los k registros más cercanos o similares dentro del conjunto de datos de entrenamiento. A partir de estos vecinos, se realiza una predicción en función de la clase a la que pertenecen la mayoría de sus vecinos. La clasificación resultante se asigna como predicción del nuevo registro.
La similitud entre registros puede medirse de muchas formas distintas. Se puede utilizar un método específico para cada problema o dato. Generalmente, un buen punto de partida es la distancia euclídea.
El método KNN puede resumirse en los siguientes pasos:

(1) Calcular la distancia euclídea.

El primer paso consiste en calcular la distancia entre dos filas de un conjunto de datos.

Podemos calcular la distancia en línea recta entre dos vectores, que representan puntos en el espacio n-dimensional, utilizando la medida de distancia euclídea (podrían utilizarse otras medidas). Ésta se calcula como la raíz cuadrada de la suma de las diferencias al cuadrado entre los dos vectores:

$Distancia = \sqrt{\sum_{i=1}^{N}(x1_i - x2_i)^2)}$

Donde $x1$ es la primera fila de datos, $x2$ es la segunda fila de datos e i es el índice de una columna específica, ya que sumamos todas las columnas.

Con la distancia euclídea, cuanto menor sea el valor, más similares serán dos registros. Un valor de 0 significa que no hay diferencia entre dos registros.

(2) Obtener los vecinos más cercanos.

Los vecinos de un nuevo dato en el conjunto de datos son las k instancias más cercanas, definidas por nuestra medida de distancia.

Para localizar a los vecinos de un nuevo dato dentro de un conjunto de datos, primero debemos calcular la distancia entre cada registro del conjunto de datos y el nuevo dato. Podemos hacerlo utilizando nuestra función de distancia preparada anteriormente.

Una vez calculadas las distancias, debemos ordenar todos los registros del conjunto de datos de entrenamiento por su distancia al nuevo dato. A continuación, se seleccionan los k que sean más chicos. Para ello, podemos guardar la distancia de cada registro del conjunto de datos como una tupla, ordenar la lista de tuplas de menor a mayor distancia y obtener los primeros k vecinos.

Por último, se devuelve una lista con el número de vecinos más similares.

(3) Clasificación del nuevo dato.

Los vecinos más similares del conjunto de datos de entrenamiento pueden utilizarse para clasificar el nuevo dato. Es decir, que se clasifica al nuevo dato como la clase más representada entre los k vecinos.

De esta forma es posible obtener a partir del conjunto de datos ($\mathbf{Y_{train}}$, $\mathbf{T_{train}}$), los valores estimados $\mathbf{Y_{estimado}}$ asociados a $\mathbf{T_{test}}$.

Matriz de confusión

Para evaluar cuán bien o mal estimamos la clasificación de un vino, necesitamos comparar $\mathbf{Y_{test}}$ con $\mathbf{Y_{estimado}}$, y ver si coinciden o no cada una de sus componentes.
Una sencilla opción es contabilizar la cantidad de aciertos y la cantidad de fallos y dividirlos por la suma total, esto nos dará la proporción de aciertos y fallos.
También se puede construir una matriz de confusión, que es una matriz de dos dimensiones, la cual tiene en una de sus direcciones los valores de testeo verdaderos y en la otra dirección los valores estimados. Esto permite contabilizar y visualizar de a pares, si existe por ejemplo, algún caso en que debería clasificarse como 1 y se clasifica como 2, o viceversa, y para cada una de las categorías de clasificación.

Bibliografía

[1] Mixtures of probabilistic principal component analysers, Neural Computation 11(2), pp 443–482. MIT Press. 1999 (https://direct.mit.edu/neco)
[2] Clases teóricas del 24, 26 y 31/10 2023.
[3] Ver ejercicios 18, 19 y 20 de la Guía de Ejercicios 4 de Autovalores y Autovectores.
[4] Python Programming And Numerical Methods: A Guide For Engineers And Scientists. Capítulo 15. Qingkai Kong, Timmy Siauw and Alexandre Bayen.
(https://pythonnumericalmethods.berkeley.edu/notebooks/Index.html)