Selección de características - Aplicaciones

En este proyecto se busca que el estudiante seleccione las características que mejor representan las un conjunto de datos multivariados. El estudiante deberá hacer uso de la descomposición en valores singulares para analizar la variabilidad de los datos y obtener proyecciones donde se maximice la varianza.

Procedimiento

Asumir que se quiere analizar un conjunto de datos $\mathbf{Y} \in \mathbb{R}^{N\times D}$, donde N es el numero de datos (o muestras) y D es la cantidad de caracter'isticas (o la dimensión de cada $\mathbf{y}_i \in \mathbb{R}^D$)

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1D} \\ Y_{21} & Y_{22} & \dots & Y_{2D} \\ \vdots & \vdots & & \vdots \\ Y_{N1} & Y_{N2} & \dots & Y_{ND} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^\top \\ \mathbf{y}_2^\top \\ \vdots \\ \mathbf{y}_N^\top \end{bmatrix}$$

A partir de dicha matriz se construye la siguiente matriz cuadrada $\mathbf{S} \in \mathbb{R}^{D\times D}$

$$\mathbf{S} = \mathbf{Y}^\top\mathbf{Y}$$

Para el análisis se obtiene la descomposición en valores singulares (SVD):

- Calcular los valores propios de la matriz $\mathbf{S}$.
- Estandarizar el vector de valores propios diviendolos con respecto a la suma total de ellos. De esta manera los valores propios estandarizados tendrán valores menores que 1, y su suma deber dar 1.
- Organizar los valores propios de mayor a menor. A partir de este orden también determinar el orden los vectores propios.
- Seleccionar los dos vectores propios correspondientes a los valores propios estandarizados más grandes.
- Cada vector propio $\mathbf{u}_d \in \mathbb{R}^D$ contiene una valores que ponderan cada dimensión o característica de los datos. Proyectar los datos $\mathbf{Y}$ sobre los dos vectores propios seleccionados ($\mathbf{U} \in \mathbb{R}^{D\times 2}$) y almacenar el resultado en $\mathbf{P} \in \mathbb{R}^{N\times 2}$

$$\mathbf{P} = \mathbf{YU}.$$

- Realizar una gráfica de la primera columna de $\mathbf{P}$ contra la otra. Cada muestra debe ser coloreada de acuerdo a la clase que pertenece. Observar si las clases son separables en esta nueva representación.
- Observar los vectores propios y seleccionar las características que tenga mayor ponderación a partir de los valores almacenados en los vectores propios seleccionados. De esta manera se puede seleccionar las características mas importantes.
- Compara el gráfico que se obtuvo previamente, con gráficos de formados a partir de pares de características seleccionadas en el paso anterior.

Bases de datos

Se pueden emplear datos relacionados con sus proyectos de investigación. Aunque también se pueden usar cualesquiera de las siguientes bases de datos:

- Breast Cancer, Iris plants, Wine, Diabetes. (Disponibles en la librería SKlearn).
- Una base de datos de su preferencia que sea orientada para Clasificación.

Referencias

[Boyd and Vandenberghe, 2018] Boyd, S. and Vandenberghe, L. (2018). *Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares*. Cambridge University Press.