CI-0148 Aprendizaje Máquina
Prof. Pablo Sauma Chacón

Laboratorio 1

Objetivo: Familiarizar al estudiante con el análisis de componentes principales, la lectura de gráficos y la utilización de bibliotecas en Python.

Enunciado: Lea las instrucciones del documento y resuelva el enunciado en Python. Al inicio de su documento adjunte su nombre y carnet como un comentario. Este trabajo es de carácter individual, tampoco se aceptarán códigos que no hayan sido desarrollados por su persona.

Para este laboratorio se trabajará con el set de datos titanic.csv, que consiste en datos tomados de algunos de los pasajeros a bordo del famoso barco Titanic. Dicho set de datos cuenta con las siguientes columnas:

`PassengerId`: Identificador numérico único para cada pasajero.
`Survived`: 1 si el pasajero sobrevivió al hundimiento
`Pclass`: Clase en la que viajaba el pasajero (primera clase, segunda clase, tercera clase)
`Name`: Nombre del pasajero
`Sex`: Sexo del pasajero
`Age`: Edad del pasajero, para algunos menores se incluye una fracción que representa los meses
`SibSp`: Total de hermanos y esposa a bordo
`Parch`: Total de padres e hijos a bordo
`Ticket`: Identificador del tiquete del pasajero
`Fare`: Precio de venta del pasaje
`Cabin`: Identificador de cabina del pasajero
`Embarked`: Puerto en el que embarcó el pasajero

1. Implemente un método que cargue el set de datos a memoria (puede utilizar la biblioteca `pandas` para esto). Revise el contenido del set de datos. Elimine las columnas que considere que no son relevantes para su análisis y utilice documentación interna para justificar su razonamiento. Luego con las columnas restantes elimine cualquier entrada que posea datos faltantes. Utilice *one-hot encoding* para convertir las variables categóricas en variables numéricas.

2. Convierta los datos obtenidos en el punto anterior en una matriz de numpy.

3. Implemente una clase llamada myPCA que permita recibir una matriz `numpy` de datos y obtener la matriz de componentes principales C, así como las inercias y los puntos necesarios para dibujar el círculo de correlación. Este método debe implementarlo usted, por lo que **no** se permitirá el uso de la biblioteca `scikit-learn` ni ninguna otra que permita ejecutar PCA de manera directa. Para calcular la matriz C deberá ejecutar los pasos del algoritmo PCA:
    a. Centrar y reducir la matriz

   b. Calcular la matriz de correlaciones
   c. Calcular los valores y vectores propios. Para ello puede serle útil el método `numpy.linalg.eigh`, que recibe una matriz y retorna sus valores propios y sus vectores propios en forma de matriz.
   d. Ordene los valores y vectores propios de mayor a menor, formando la matriz V.
   e. Puede calcular C realizando la multiplicación X * V
   f. Las inercia de cada componente principal corresponde a su valor propio dividido entre el total de variables (m).
   g. Los puntos necesarios del círculo de correlación corresponden a los valores de la matriz V, en las columnas 0 y 1 para todas las filas. Estos valores deben ir multiplicados por la raíz cuadrada de su valor propio correspondiente.

4. Utilice la clase implementada para ejecutar el algoritmo PCA con los datos obtenidos de titanic.csv. Grafique los datos sobre sus componentes principales y coloree los puntos según su valor "`Survived`". Grafique el círculo de correlación del modelo.

5. Agregue en documentación interna lo que muestran estos gráficos. ¿Cuántos grupos de datos parece haber?¿Qué comportamientos se pueden observar? ¿Qué podría explicar estos comportamientos? ¿Qué nos indica el círculo de correlación?

6. Si yo fuera un pasajero del Titanic, ¿qué atributos o características maximizarían mi probabilidad de sobrevivencia?

7. Por motivos de verificación, repita el experimento pero esta vez utilizando la biblioteca `scikit-learn` (`sklearn`). Nuevamente, vuelva a cargar el documento y realice el proceso para ejecutar el algoritmo haciendo uso de la poderosa biblioteca `sklearn`. Grafique nuevamente los datos sobre sus componentes principales haciendo uso del coloreo y también grafique el círculo de correlación del modelo.

8. ¿Hay alguna diferencia entre las gráficas? De ser así, ¿por qué cree que ocurrió esto? ¿Impacta el resultado de alguna manera?