UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores
Acatlán

Practica 2:

PCA

Aprendizaje de maquina y minería de datos avanzados

Integrantes:

Gustavo Adolfo Alvarez Hernández

Profesor:

Eduardo Eloy Loza Pacheco

FES Acatlán, 9 de febrero 2023

Objetivo

Que el alumno aprenda a utilizar métodos de reducción de la dimensionalidad como lo es PCA, con el fin de quitar variables altamente correlacionadas, explicando en cierta medida los datos con menor número de variables; Esto además de mejorar el número de cálculo permite visualizar los datos y entender de mejor manera problemas multidimensionales.

Resumen

En la presente practica se utilizarán datos sobre vino, en el cual se tienen características de cada vino y la calidad asociada al vino, a dichos datos se les aplico PCA para solo generar dos variables y mostrar que con solo dos variables es posible generar una regresión logística que clasifique de una manera adecuada los tipos de vino, sin la necesidad de usar todas las variables directamente. Finalizando con una representación visual de la clasificación del clasificador.

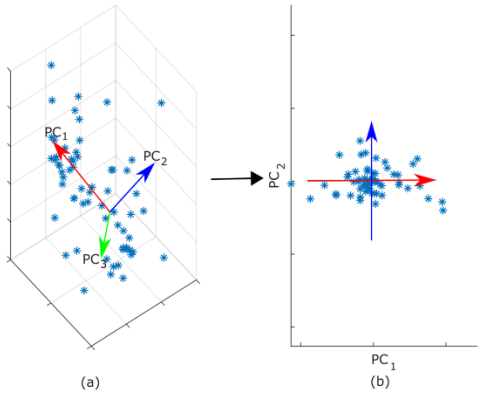
Antecedentes

Análisis de componentes principales (PCA), es una técnica de reducción de dimensionalidad, la cual parte de la idea de tomar datos con variables correlacionadas y generar n nuevas variables z_i no correlacionadas entre ellas y explican la misma varianza que las variables iniciales.

Su funcionamiento se basa en conceptos de algebra lineal como lo son eigenvalores y eigenvectores, su procedimiento se puede reducir a:

1. Se estandarizan los valores de X, esto es que tengan media cero y varianza unitaria.
2. Se genera una eigendescomposición sobre la matriz de covarianza
$$C_x = \frac{1}{n} X'X$$
Dicho proceso consiste encontrar los valores y vectores propios.
3. Los valores propios se ordenan en orden decreciente y representan la varianza de la componente
4. Y por medio de una operación de proyección que es el producto punto entre la matriz X y los vectores propios generan a cada una de las n componentes, donde $n = rank(X)$.
    a. Si se observa ahora cada componente será una combinación lineal de las variables originales y cada componente irá explicando en menor proporción la varianza de todos los datos.

De esta manera, uno puede explicar la mayor varianza de los datos usando una menor cantidad de variables que las originales.



Desarrollo

1. Importar los módulos necesarios

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.decomposition import PCA
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import  StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

2. Cargamos los datos sobre los vinos

```
data= pd.read_csv("Wine.csv")
```

3. Separamos nuestros datos en entrenamiento y prueba para validar posteriormente al modelo.

```
x_train,x_test,y_train,y_test= train_test_split(X,y, test_size=.25, random_state=0)
```

4. Para asignar una ponderación a cada variable justa, procedemos a escalar nuestros datos y servirá como paso intermedio para PCA

```
sx= StandardScaler()
x_train= sx.fit_transform(x_train)
x_test= sx.fit_transform(x_test)
```

5. Procedemos a instanciar PCA y transformar nuestros datos

```
pca= PCA(n_components=2)
x_train= pca.fit_transform(x_train)
x_test= pca.transform(x_test)
```

6. Ahora observamos que tanta de la varianza total de los datos explicamos con cada una de las dos componentes que generamos

```
pca.explained_variance_ratio_
```

```
array([0.37281068, 0.18739996])
```

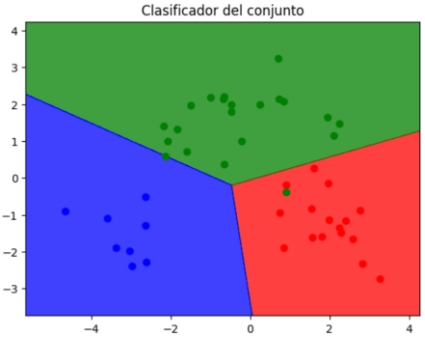
7. Entrenamos una regresión logística con las dos componentes que generamos y posteriormente hacemos la predicción sobre el test.

```
regLog= LogisticRegression(random_state=0)
regLog.fit(x_train, y_train)

y_predict= regLog.predict(x_test)
```

8. Graficamos su forma de decisión que realizo en el entrenamiento

9. Y por último su desempeño en datos que no uso para entrenar.



Conclusiones

Por medio de PCA se redujo de 11 variables a 2 componentes principales lo cual aligera la cantidad de cómputo, estas dos variables explican poco mas del 55% de la varianza total de los datos totales con sus 11 variables.

Por su parte aun sin usar toda la información se observa que la clasificación tiene un desempeño bastante bueno y generalizable, por lo cual además de reducir computo, tenemos un desempeño bueno a pesar de la reducción de información.

Por último, otra ventaja del método es que nos posibilita la generación de visualizaciones que nos permita entender de forma general como se realiza la clasificación y como se comportan nuestros datos.

Referencias

Serafeim Loukas, P. (30 de May de 2020). *Towards Data Science*. Obtenido de PCA clearly explained —When, Why, How to use it and feature importance: A guide in Python: https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

Zhou, Z.-H. (2016). *Machine Learning.* Singapore: Springer Nature Singapore.