AlgLin - Principal Component Analysis

João Pedro Paiva Cardoso, Marcos Paulo Sousa Santos

December 2022

1 O Dataset

Selecionamos um dataset do Instituto Nacional de Diabates e Doenças Digestórias e Renais (Reino Unido) que foi utilizado em um estudo que visava prever casos de diabetes utilizando dados do sangue dos pacientes.

Este dataset consiste em diversas variáveis médicas como IMC, nível de insulina, idade etc e de uma váriavel alvo, Outcome (resultado). Outcome = 1 representa resultados positivos (o paciente possui diabetes) e Outcome = 0 representa resultados negativos (o paciente não possui diabetes)

.

Link do dataset:
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

2 O Algoritmo

Primeiro, para normalizar os dados, obtivemos a média (mean) e a subtraímos do total de dados. Dessa forma deixando os valores das colunas mais uniformes.

```
mean = np.mean(data,axis=0)
meanReducedData = data - mean
```

Então, calculamos a matriz de covariância e a transpomos pois o algoritmo estava trabalhando com base nas linhas e não nas colunas:

```
covMatrix = np.cov(MeanReducedData.T)
```

Com isso já podemos calcular os autovetores (eigVec) e autovalores (eigVal):

```
eigVal,eigVec = np.linalg.eig(covMatrix)
```

Sabendo os autovalores, os ordenamos do maior para o menor

```
eigVal = np.array(np.sort(eigVal))[::-1]
```

Descobrimos assim que os maiores autovalores são 1.34565776e+04 e 9.32085966e+02, que correspondem as colunas Pregnancies (nº de gravidez) e Glucose (glicose) respectivamente. Por fim, basta representar estas colunas em um gráfico.

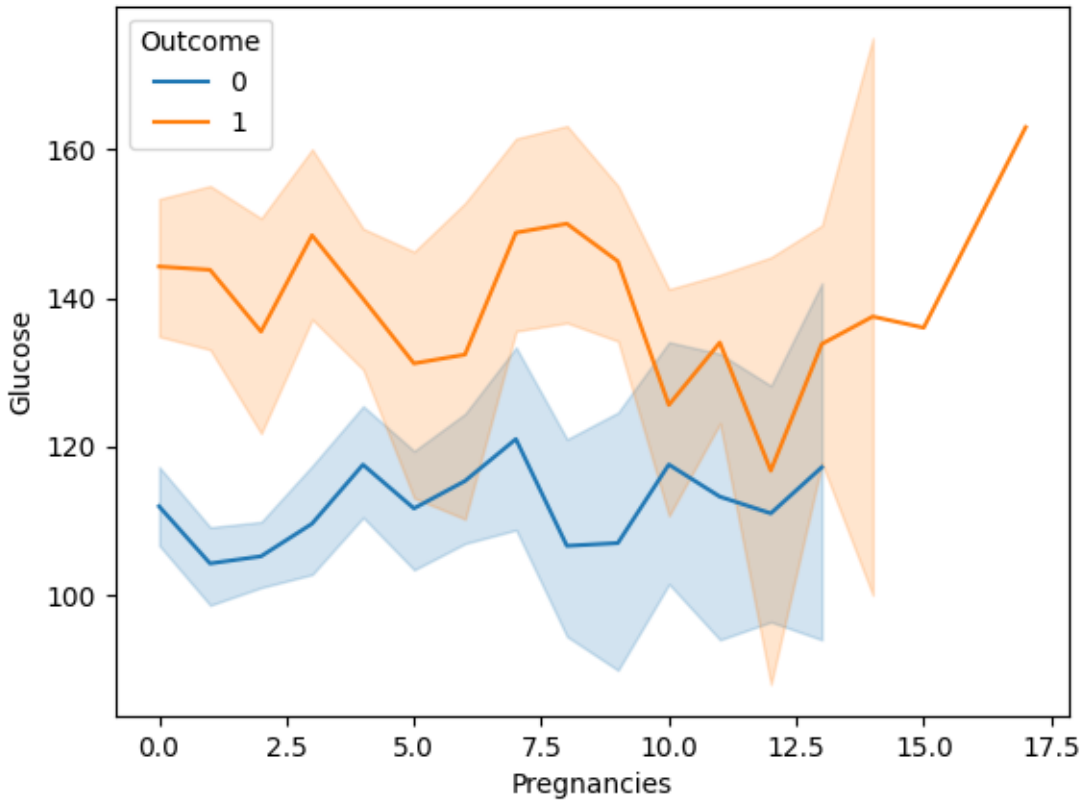

Figure 1: Gráfico dos Maiores Autovalores

```
sns.lineplot(x="Pregnancies",y="Glucose",data=data,hue="Outcome")
plt.show()
```

Observando o gráfico a seguir vemos que conforme a Glicose e as Gravidezes aumentam, o Outcome 1 (resultado positivo) é maior. Ou seja, são as melhores variáveis para se analisar.

Link do repositório:
https://github.com/Mr-marcs/PCA