Relatório Exercício 5

Aluno: João Guilherme Madeira Araújo

Número USP: 9725165

Relatório do Exercício 5:

O exercício pediu a implementação de um método de redução de dimensionalidade e análise de dados chamado Principal Component Analysis, PCA, no dataset íris. Devemos mostrar como PCA nos ajuda a melhor visualizar e entender o dataset.

Datasets

Foi usado o dataset íris, que contém 50 amostra de plantas do gênero íris. Os dados de cada amostra são o comprimento das sépalas, largura das sépalas, comprimento das pétalas e largura das pétalas, além da espécie de cada planta, Iris setosa, Iris virginica e Iris versicolor.

O Algoritmo

O código-fonte do Principal Component Analysis está no arquivo pca.py. O algoritmo calcula os autovetores e autovalores da matriz de covariância dos nossos dados, calculamos quanto da variância é explicada por cada PC, plotamos todos pontos do nosso dataset em função de todos pares de características e por fim projetamos os dados nos dois maiores Principal Components e plotamos eles para comparar.

Conclusões

Podemos observar que o *scatter plot* (Figura 3) que usa os PCAs como dimensões tem uma separabilidade das espécies bem melhor e mais visível que os demais gráficos (Figura 2), permitindo possivelmente uma melhora análise dos dados. Esse resultado já era esperado uma vez que juntos os dois primeiros componentes principais explicam 97% da variabilidade dos dados (Figura 1).