Universidade de São Paulo – USP São Carlos
Instituto de Ciências Matemáticas e de Computação
SCC5809 – Redes Neurais – Profa. Dra. Roseli Romero
Damares Oliveira de Resende – #11022990

Relatório de Atividades – Exercício 5

Nesse exercício aplicou-se o método PCA na base *wine*. Esse método é capaz de projetar os atributos do dado de entrada em outro domínio, reduzindo o número de componentes significativas e conservando a significância dos dados.

A base *wine* possui 13 atributos e 3 classes. Neste trabalho o número de atributos foi reduzido para 2. Para tal, usa-se a padronização padrão onde de cada valor subtrai-se a média e divide-se o resultado pelo desvio padrão. Isso possibilita que a média da base fique em torno de 0 e a variância em torno de 1.

Em seguida aplicou-se o método do PCA, que foi implementado usando apenas a biblioteca *numpy* e está disponibilizado em conjunto com o relatório. Esse algoritmo segue 4 passos:

- Centralizar os dados
- Calcular a matriz de covariância
- Calcular os autovetores e autovalores
- Ordenar os valores e selecionar os *n* mais significantes

O PCA codificado foi testado contra o algoritmo disponibilizado pela biblioteca *sklearn* e obteve os mesmos resultados, o que indica que o código implementado está correto. Porém, a padronização dos dados de entrada no *sklearn* (*StandardScaler*) é feita de forma diferente. Para essa etapa os resultados divergiram. A Tabela 1 mostra os valores da razão da variância obtidas para ambos casos.

Tipo de Padronização	Razão da Componente 1	Razão da Componente 2
StandardScaler	0.3620	0.1921
Std Padrão	0.9981	0.0017

Tabela 1: Resultado da razão de variância para as componentes 1 e 2, onde o PCA utilizado foi codificado apenas com a biblioteca *numpy* e o método de padronização foi variado: o método *StandardScaler* é fornecido pelo *sklearn* e o método Std Padrão foi codificado. Nesse último subtrai-se de cada valor a média e divide-se pelo desvio padrão.

Os gráficos ilustrados nas Figura 1 e 2 mostram a diferença da projeção de cada abordagem. Quando se utiliza a padronização do *sklearn,* as classes ficam mais bem definidas e separadas em clusteres. Porém, as componentes não mantêm grande significância dos dados (somadas elas têm apenas 55% da significância). No caso do método de padronização codificado, quase 100% da significância dos dados é mantida em duas componentes, porém não é possível separar linearmente as classes 2 e 3. Contudo, como esse método mantém a significância dos dados, ele reflete melhor o que eles representam e, portanto, deve ser utilizado. Dessa forma, pode-se concluir que os dados não são linearmente separáveis e que algoritmos de classificação complexos devem ser considerados.

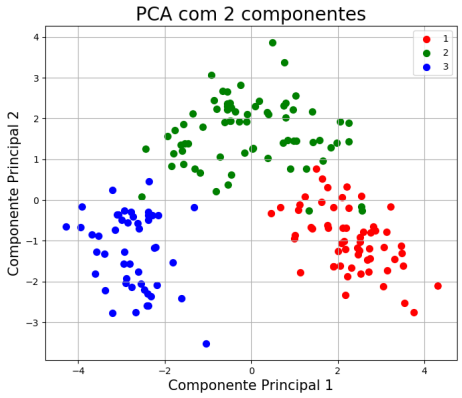

Figura 1: PCA com duas componentes para dados padronizados utilizando o *StandardScaler* do *sklearn*.

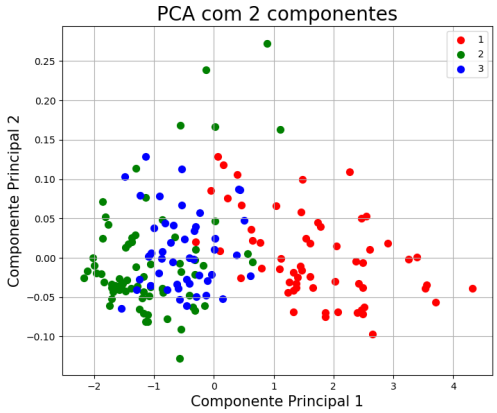

Figura 2: PCA com duas componentes para dados padronizados utilizando o método padrão (subtrai-se a média e divide-se pelo desvio padrão).