Disciplina: Aprendizagem de Máquina
Período: 2024.1
Professor: César Lincoln Cavalcante Mattos

Lista 6 - K-médias e PCA

Instruções

- Com exceção dos casos explicitamente indicados, os algoritmos e modelos devem ser implementados do início em qualquer linguagem de programação (Python, R, Octave...).
- Pacotes auxiliares (sklearn, matplotlib, etc) podem ser usados somente para facilitar a manipulação dos dados e criar gráficos.
- A entrega da solução pode ser feita via pdf ou Jupyter notebook pelo SIGAA.

Questão 1

Considere o conjunto de dados disponível em **quake.csv**, organizado em 2 colunas de atributos. Os dados referem-se a latitudes e longitudes de locais em que foram registrados terremotos. Maiores detalhes sobre os dados podem ser conferidos em `https://www.openml.org/d/772`.

a) Avalie o algoritmo K-médias com distância Euclidiana na tarefa de agrupamento para tais dados. O número de grupos deve ser escolhido entre os valores $4, 5, 6, \ldots, 19, 20$ a partir do índice DB (Davies-Bouldin). Plote o melhor resultado de agrupamento obtido.

   Observação: Em cada avaliação repita múltiplas vezes (por exemplo, 20) a execução do algoritmo K-médias, escolhendo a solução com melhor erro de reconstrução.

b) Repita o item anterior considerando a distância de Mahalanobis.

Questão 2

Considere o conjunto de dados disponível em **penguins.csv**, organizado em 5 colunas, sendo 4 colunas de atributos e a última a classe do padrão. Os dados referem-se a medições anatômicas de pinguins da Antártida, classificados nas espécies Adelie, Chinstrap e Gentoo. Maiores detalhes sobre os dados podem ser conferidos em `https://allisonhorst.github.io/palmerpenguins/`.

a) Apresente a projeção em 2 dimensões dos padrões acima obtida pelo método PCA (análise dos componentes principais).

b) Ainda considerando o item anterior, calcule e mostre a variância explicada obtida quando a dimensão projetada é modificada (1,2,3 ou 4).

Observação: Não esqueça de normalizar os dados em ambas as questões.