Lista 5

K-médias e PCA

Instruções

Deverá ser enviado ao professor, um arquivo texto contendo os gráficos, resultados e comentários requeridos em cada item.

1. **K-médias**

- Carregue os dados contidos no arquivo ex5data1.data.

O arquivo contém uma matriz de dados. Esta matriz é composta de 150 linhas e 5 colunas. As 4 primeiras colunas representam 4 atributos e a coluna 5 representa a classe a qual pertence o exemplo. Nestes dados, existem 3 classes, sendo 50 exemplos de cada classe.

Os dados pertencem a um problema de reconhecimento flores (íris dataset). Os 4 atributos são tamanho e espessura da sépala e da pétala de cada flor. As três classes referem-se as flores 1-setosa, 2-versicolor e 3-virginica.

- Implemente o k-médias para a base de dados, utilizando somente os 4 primeiros atributos.
- Varie o número de clusters entre 2 e 5
- Calcule o somatório dos erros quadráticos em relação aos centroides para cada número de agrupamentos.

Apresentar: Gráfico do erro pelo número de agrupamentos
Apresentar: O número de agrupamentos para este problema, de acordo com a heurística apresentada em aula
Comentários: Comente sobre o número de classes obtido

- Execute o K-médias para o número de agrupamentos obtidos
- Compare o resultado com o valor real das classes

Comentários: Comente sobre o resultado obtido

2. **PCA**

- Utilizando a mesma base de dados da questão anterior, aplique o algoritmo PCA e reduza a dimensão de modo a preservar 99% da variância.

Apresentar: O número de atributos

Comentários: Comente sobre como foi obtido este número de atributos.

- Reduza a dimensão da base de dados original para 2.
Apresentar: Figura em 2 dimensões com os dados. Utilize cores diferentes para cada classe.
Comentários: Sabendo que uma classe é linearmente separável e as outras duas não são, verifique se este comportamento é mantido para o conjunto de dados com 2 dimensões.