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.

Os dados foram normalizados.

1. **K-médias**

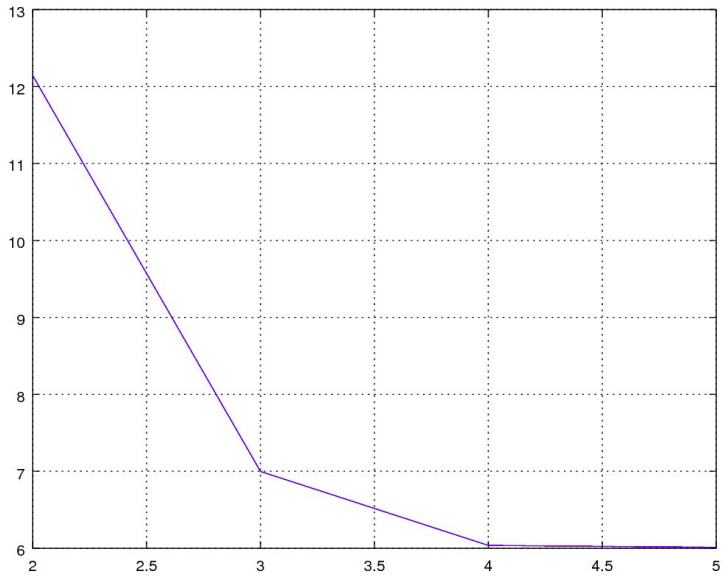
- Carregue os dados contidos no arquivo ex5data1.data.

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
Heurística: selecionar o K que mais melhorou (diminuiu o erro).
Nesse caso, o **K = 3.**

Comentários: Comente sobre o número de classes obtido
O número de agrupamento é exatamente o número de classes que há no dado, 3. Através do gráfico, é possível perceber que para K > 3, o erro diminuiu. Isso é fato: quanto mais o número de clusters se aproximar do número de exemplos, menos erro há. Porém, maior o overfitting. Pensando nisso, faz sentido selecionar o K que mais impactou na diminuição do erro, K = 3.

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

- Classe 1 - setosa

Para os 50 exemplos da classe 1, o k-means(k=3) agrupou **todos em 1 único grupo**. Coesão = 100%.

- Classe 2 - versicolor

Para os 50 exemplos da classe 2, o k-means(k=3) agrupou **41** em um grupo e **9** em outro. Coesão = 82%.

- Classe 3 - virginica

Para os 50 exemplos da classe 3, o k-means(k=3) agrupou **8** em um grupo e **42** em outro. Coesão = 84%.

Comentários: Comente sobre o resultado obtido.
A coesão nos agrupamentos foi elevada. Foi perceptível nos vários testes que há duas classes bem próximas. Seria necessário analisar a distribuição dos dados no espaço para predizer a causa de tal comportamento. Talvez estas duas classes não são linearmente separáveis.

2. PCA

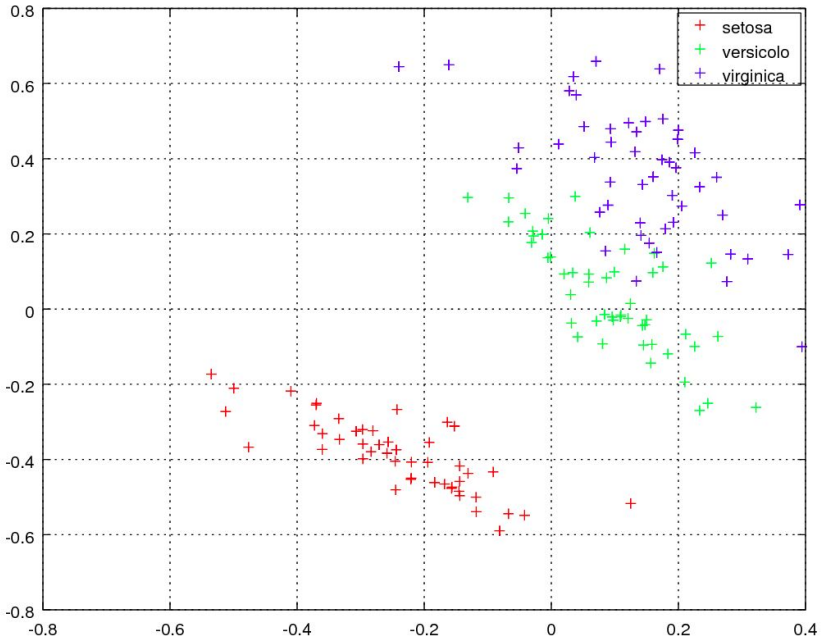
- Utilizando a mesma base de dados da questão anterior, aplique o algoritmo PCA e reduza a dimensão de modo a preservar 99% da variância.
Apresentar: O número de atributos
Número de atributos necessários para preservar 99% da variância: **3**

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

A partir dos autovalores, foram gerados os percentuais de variância para os top k=1...4 atributos com maiores variâncias. O atributo 4 (k=1), por si só, tem 84% da variância; os atributos 3 e 4(k=2), somados, 95% ; e, finalmente, os atributos **2, 3** e **4 (k=3)**, têm, somando suas variâncias, 99% da variância total.

- 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.

Analisando o gráfico, as duas classes que antes não eram linearmente separáveis (acredito que fossem as versicolor e virginica), agora podem sim ser separadas por uma linha, mesmo com erro para alguns exemplos.