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

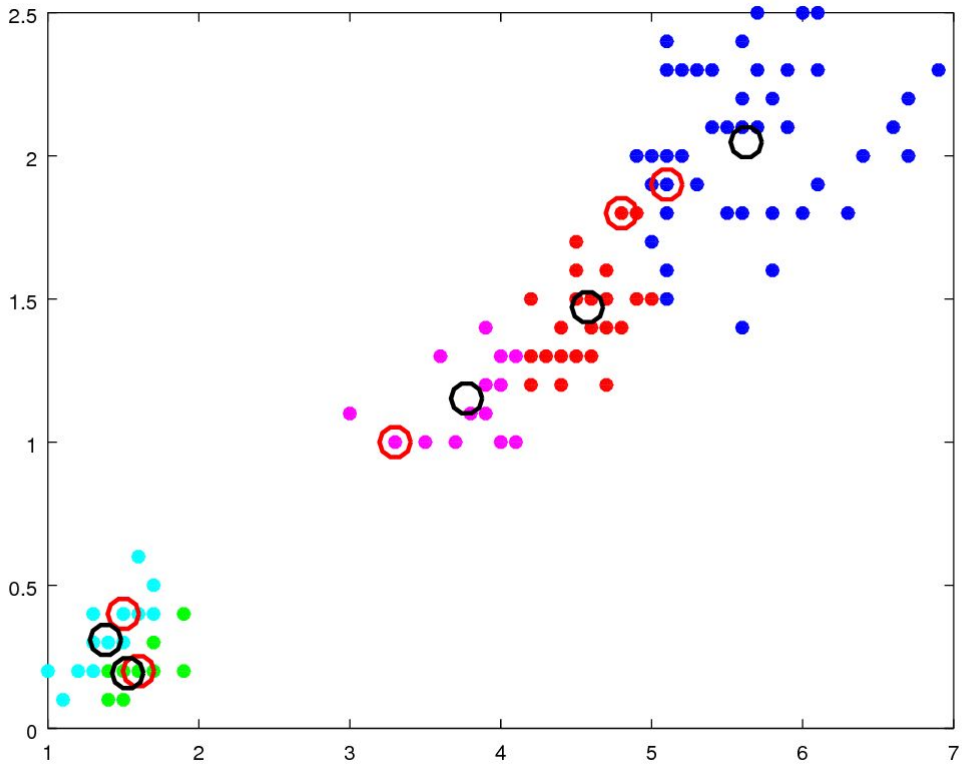


Imagem testando o algoritmo. Plotando apenas as dimensões 3 e 4. Os circulos vermelhos indicam os elementos que foram sorteados como centroides iniciais. Os circulos pretos indicam os novos centroides a partir do conjunto particionado a partir dos primeiros k centróides. A imagem está distorcida por conta da falta de escala entre as diferentes propriedades das plantas.

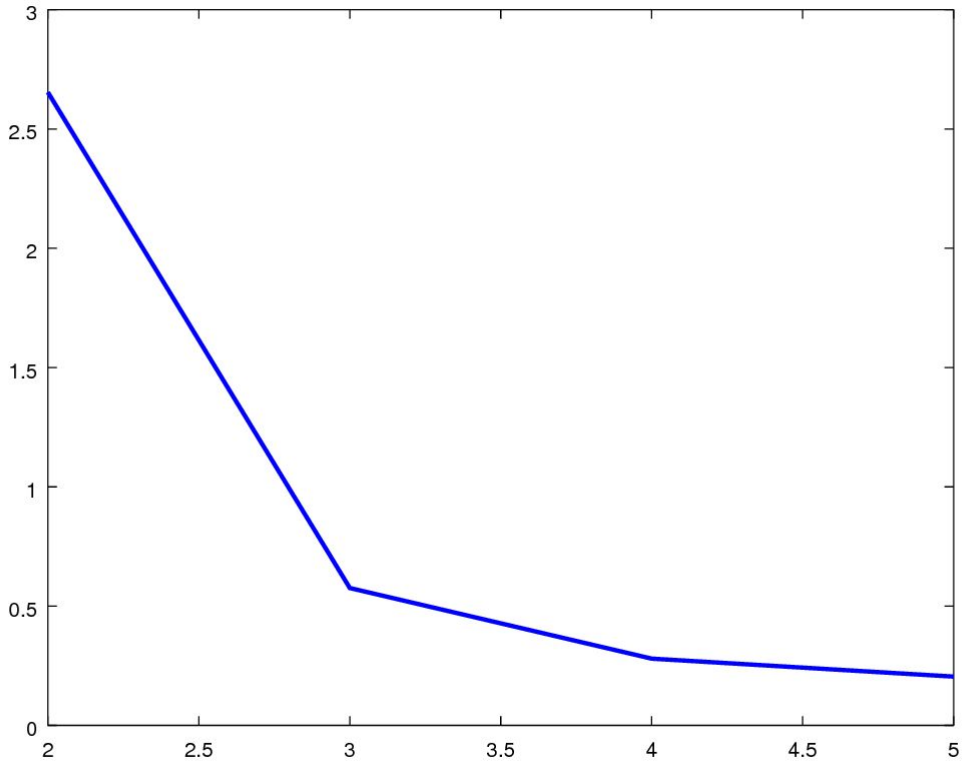
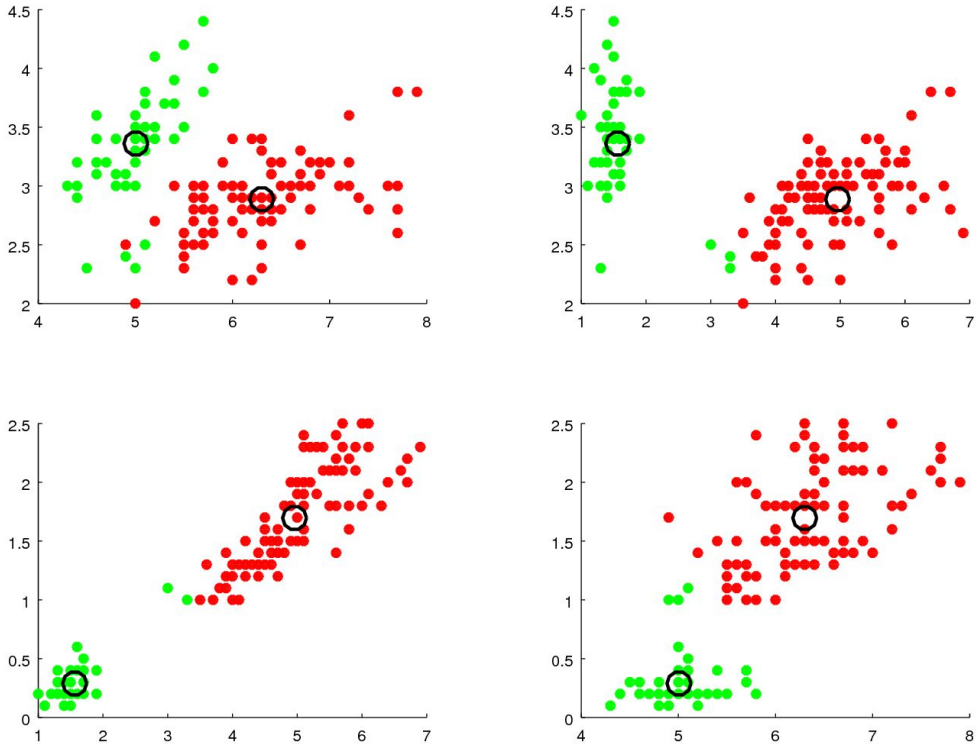


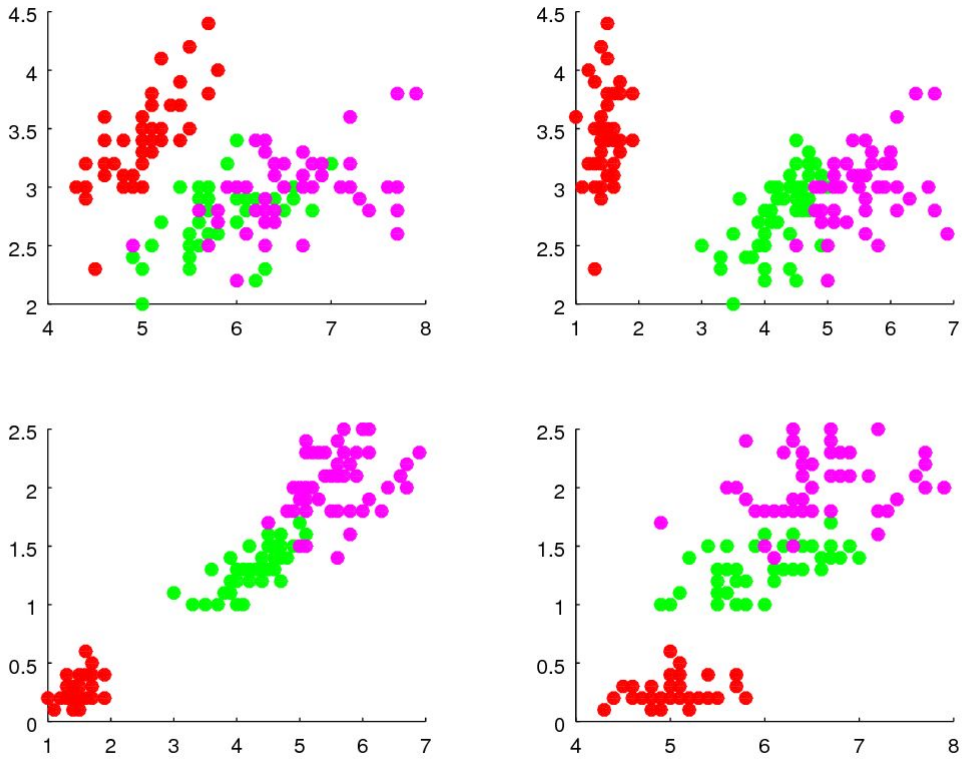
Gráfico do erro quadrático médio de cada ponto em relação ao seu centroide.
Faz sentido que o erro vá diminuindo a medida que novos centros são adicionados, uma vez que se você tiver o mesmo numero de centroides que o numero de elementos. Porém, se o erro diminuir consideravelmente menos em relação aos outros número de de centros, isso significa que podemos estar dividindo elementos que deveriam estar no mesmo grupo em grupos separados.

Portanto o valor K=3 parece ser o ideal para dividir o conjunto. E de fato, se observarmos o conjunto original de dados, ele é dividido em 3 classes.

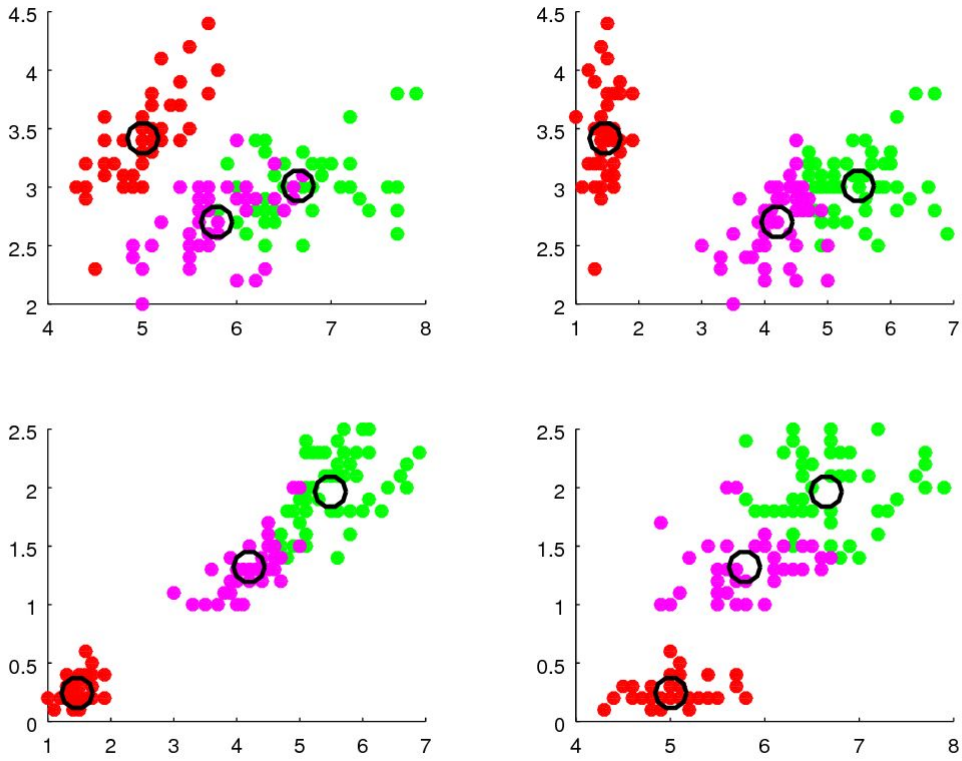
- Execute o K-médias para o número de agrupamentos obtidos
- Compare o resultado com o valor real das classes
Comentários: Comente sobre o resultado obtido



Resultado com k = 2. A condição de parada é que os centroides não se movam mais do que 0.1 de uma iteração para a outra. O plot foi feito com as dimensões 1x2, 2x3, 3x4 e 4x1.



Esses são os dados originais, com suas respectivas classes.



Com K = 3 esse foi o resultado. Bem parecido com o original. Os centroides calculados são representados pelos circulos pretos.

O resultado para o conjunto vermelho de dados foi bem preciso pois os dados originais estavam mais separados do que os demais. Porém, os dados verde e roxo (cujo as cores estão invertidas nas imagens acima) mostram que alguns elementos no limiar dos dois conjuntos foram confundidos por estarem muito próximos em todas as dimensões.

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.

Verificando os dados de entrada, obtemos as seguintes informações:

Autovalores:
l1 = 4.224841

l2 = 0.242244
l3 = 0.078524
l4 = 0.023683

Autovetores correspondentes:
v4 = [-0.31725, 0.32409, 0.47972, -0.75112]
v3 = [0.580997, -0.596418, -0.072524, -0.549061]
v2 = [0.656540, 0.729712, -0.175767, -0.074706]
v1 = [0.361590, -0.082269, 0.856572, 0.358844]

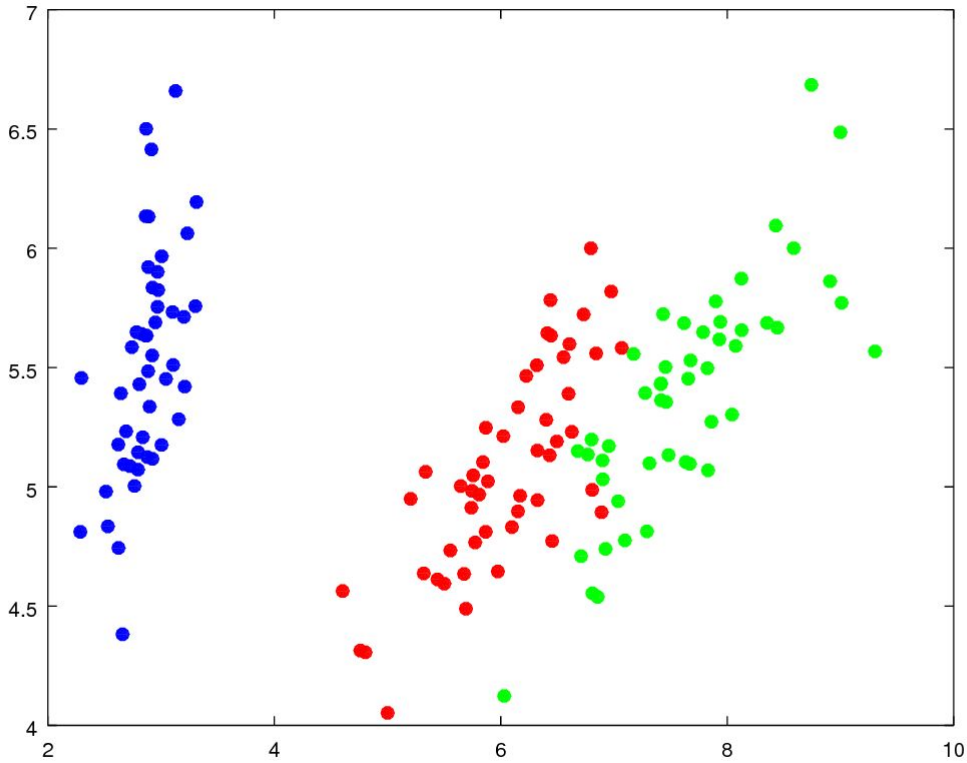
Porcentagem da variância de cada dimensão
%1 = 0.9246162
%2 = 0.0530156
%3 = 0.0171851
%4 = 0.0051831

Se considerarmos apenas as dimensões 1 e 2 teremos 97,763% da variância preservada. Se considerarmos as dimensões 1, 2 e 3 teremos 99,482% da variância.

Portanto para preservar 99% da variância teriamos de usar pelo menos 3 atributos.

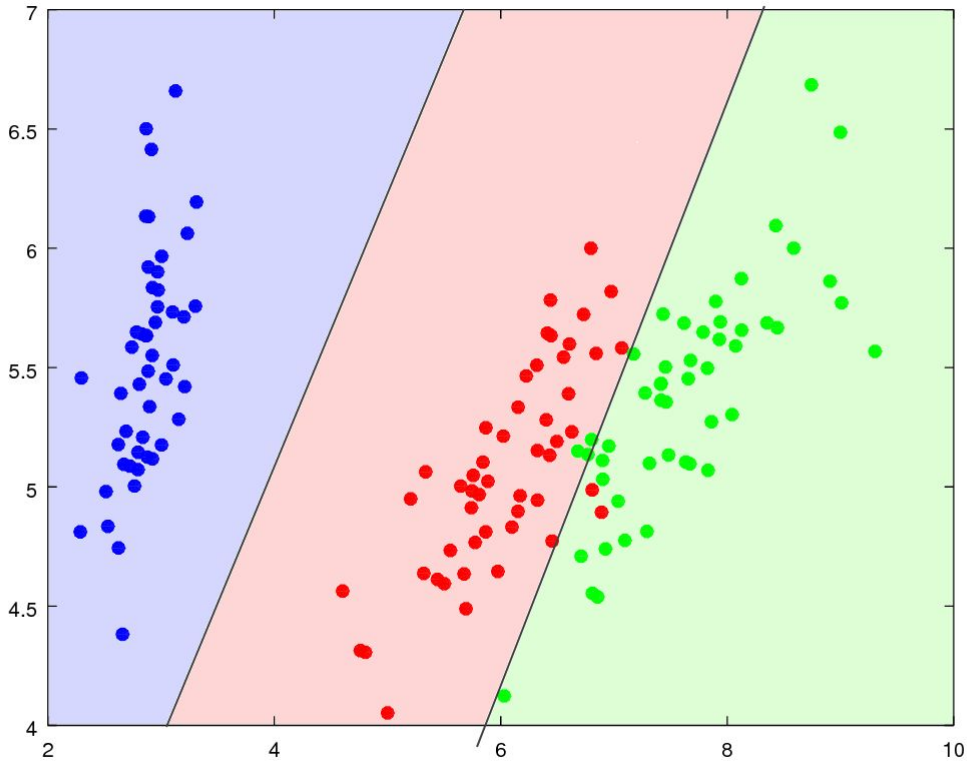
- Reduza a dimensão da base de dados original para 2.
Apresentar: Figura em 2 dimensões com os dados. Utilize cores diferentes para cada classe.
Comentários: Sabendo que uma classe é linearmente separável e as outras duas não são, verifique se este comportamento é mantido para o conjunto de dados com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

Para isso são usados as 2 principais componentes, que estão na direção dos autovetores 1 e 2 respectivamente.



este é o resultado da projeção usando as duas componentes principais.

Apesar dos dados estarem bem juntos, parece ser quase possível separá-los linearmente por completo agora.


7
6.5
6
5.5
5
4.5
4
2
4
6
8
10