Universidade Federal de Pernambuco
DES/CTG

Aprendizagem de Máquina (ES456/2018-2)

Lista de exercícios 1

2018-09-22

professor: Daniel de Filgueiras Gomes (daniel.fgomes@ufpe.br)

1) Utilizando a base de dados Iris, disponível no link abaixo, e a linguagem de programação de sua preferência, calcule para cada classe:

a) O vetor médio
b) O vetor de desvio padrão para cada característica da base de dados
c) O vetor máximo para cada característica da base de dados
d) O vetor mínimo para cada característica da base de dados
e) A matriz de dispersão
f) A matriz de covariância
g) A matriz de correlação
h) As duas componentes principais de maior magnitude dos dados
i) A projeção dos dados nas duas maiores componentes principais

2) Implemente o algoritmo do k-means na linguagem de programação de sua preferência e inicialize este algoritmo com k=3 centros c1=[5.1,3.5,1.4,0.2], c2=[4.9,3.0,1.4,0.2], c3=[4.7,3.2,1.3,0.2].

a) Após a convergência, qual o valor dos centros?
b) Qual o centro mais próximo do centro da classe “Iris-setosa”
c) Qual o centro mais próximo do centro da classe “Iris-versicolor”
d) Qual o centro mais próximo do centro da classe “Iris-virginica”

3) Implemente o algoritmo do KNN na linguagem de programação de sua preferência.

4) Implemente um classificador de distância mínima na linguagem de sua preferência.

5) Implemente uma função discriminante utilizando o log da probabilidade a posteriori para cada uma das classes da base de dados Iris.

6) Implemente um classificador bayesiano para a base de dados Iris.

7) Implemente um classificador bayesiano para a base de dados Iris projetada sobre as duas maiores componentes principais.

8) Determine o erro percentual e o desvio padrão do erro(quando for possível) nos conjuntos de treinamento e validação dos classificadores implementados nas questões 3,4, 6 e 7 utilizando os seguintes métodos de particionamento de dados:

a) Houldout com 90% das amostras para treinamento e 10% para validação;
b) Cross-Validation 10-folds;
c) Monte Carlo Cross-Validation com 90% das amostras para treinamento e 10% para validação;

9) Repita as avaliações da questão 6, calculando a matriz de confusão do erro percentual para cada um dos métodos de particionamento (Houldout, Cross-Validation 10-folds, Monte Carlo Cross-Validation).

Link para os dados utilizados nos exercícios desta lista:
https://www.dropbox.com/sh/k2vnt1tpm3doz0w/AACSSzLBdNtxEnFNTdZl4RRia?dl=0

OBS: A resposta desta lista deve ser enviada por e-mail para daniel.fgomes@ufpe.br até a data de 04/10/2018. O material enviado deve ser formado pelo(s) código-fonte(s) do(s) programa(s) utilizado(s) nos cálculos/implementações devidamente comentado(s) e um arquivo .txt com a saída dos cálculos dos programas.