Aprendizagem de Máquina Lista 1

Moisés de Siqueira Campos Neto

10 de Abril, 2016

1 Questão 1

Nessa questão, foram utilizados os seguintes banco de dados:

1. Breast Cancer Wisconcin (Original)
   Tipo dos atributos: inteiros
   Número de instâncias: 699
   Número de atributos: 10
   Nome do arquivo: *breast-cancer-wisconsin.data*

2. Wine
   Tipo dos atributos: reais
   Número de instâncias: 178
   Número de atributos: 13
   Nome do arquivo: *wine.data*

Alguns ajustes nos bancos foram necessários. No primeiro, existe uma pequena parcela de casos com dados ausentes. Para lidar com esse problema, foi escolhido ignorar esses casos e retirá-los do dataset. A classe se localiza como última característica de cada caso. No segundo, foi necessária apenas uma reordenação das colunas, de modo que a classe do caso fosse para a última coluna.

Os bancos foram divididos em uma proporção de 70% dele para treinamento e 30% para validação. Os datasets são randomizados antes da divisão, para evitar problemas relacionados a casos de uma mesma classe aglomerados. Além de randomizados, todos os dados, exceto a classe, são transformados para números reais e normalizados, de modo que todos os valores fiquem entre 0 e 1.

A distância utilizada foi a distância Euclidiana.

Os resultados são mostrados nas tabelas 1 e 2, e figuras 1 e 2.

Table 1: Resultados do K-NN no banco de dados 1 da questão 1

K	Quantidade de erros	Acurácia
1	15	92.68 %
2	14	93.17 %
3	14	93.17 %
5	13	93.66 %
7	9	95.60 %
9	19	90.73 %
11	18	91.22 %
13	12	94.15 %
15	16	92.19 %

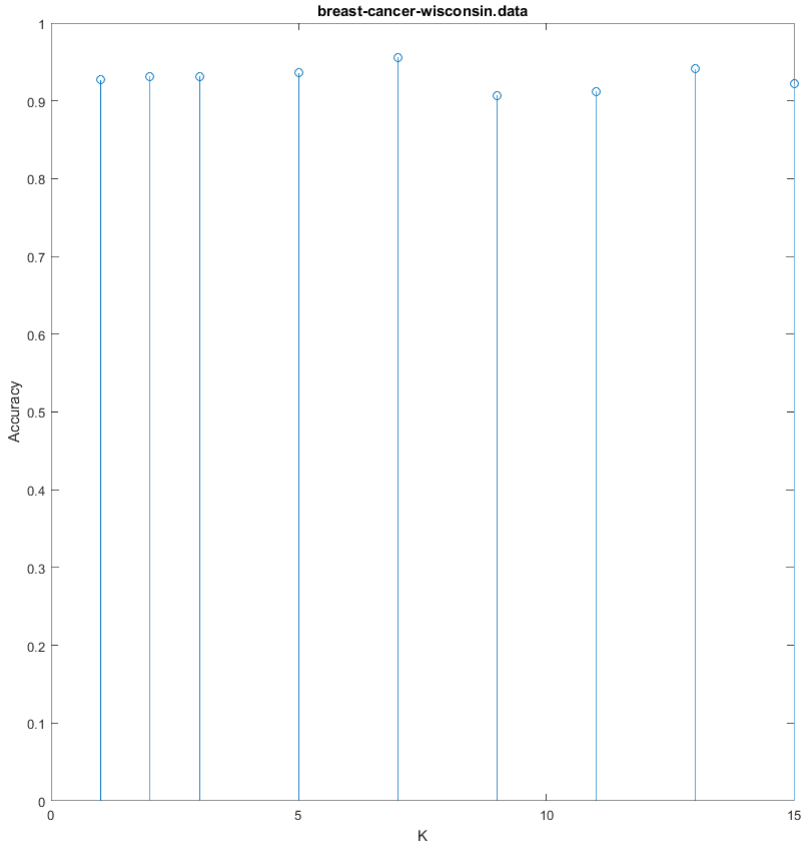

Figure 1: Gráfico da acurácia relacionada ao K utilizado no banco 1 da questão 1

Table 2: Resultados do K-NN no banco de dados 2 da questão 1

K	Quantidade de erros	Acurácia
1	31	41.50 %
2	35	33.96 %
3	29	45.28 %
5	40	24.52 %
7	34	35.84 %
9	34	35.84 %
11	35	33.96 %
13	44	16.98 %
15	40	24.52 %

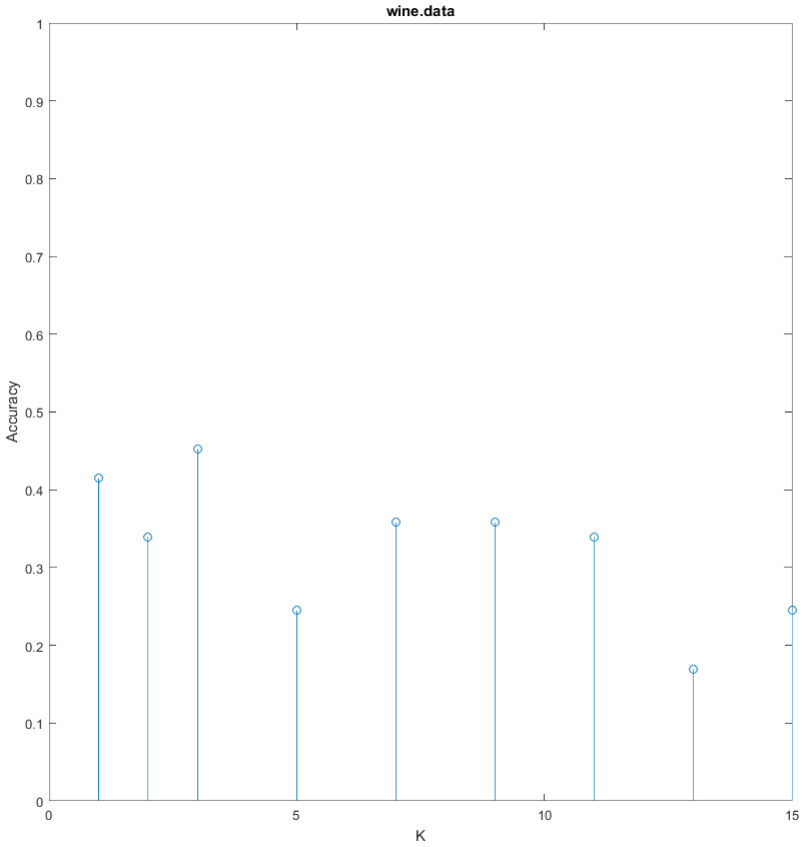

Figure 2: Gráfico da acurácia relacionada ao K utilizado no banco 2 da questão 1

2 Questão 2

Nessa questão, foram utilizados os seguintes banco de dados:

1. Acute Inflammations
   Tipo dos atributos: categóricos, inteiros
   Número de instâncias: 120
   Número de atributos: 6
   Nome do arquivo: *diagnosis2.data*

2. BLOGGER
   Tipo dos atributos: categórigos, inteiros
   Número de instâncias: 100
   Número de atributos: 6
   Nome do arquivo: *pro-bloggers.data*

O banco de dados *Acute Inflammations* precisou de um ajuste. Nos dados, o primeiro atributo era um número real relativo à temperatura do paciente. Tal atributo foi removido para manter o banco exclusivamente categórico. Foram necessárias reordenações de colunas nos dois bancos.

Os bancos foram divididos em uma proporção de 30% dele para treinamento e 70% para validação. Os datasets são randomizados antes da divisão, para evitar problemas relacionados a casos de uma mesma classe aglomerados.

Os resultados são mostrados nas tabelas 3 e 4, e figuras 3 e 4.

Table 3: Resultados do K-NN no banco de dados 1 da questão 2

K	Quantidade de erros	Acurácia
1	0	100.0 %
2	0	100.0 %
3	0	100.0 %
5	0	100.0 %
7	0	100.0 %
9	0	100.0 %
11	17	79.52 %
13	31	62.65 %
15	17	79.52 %

Table 4: Resultados do K-NN no banco de dados 2 da questão 2

K	Quantidade de erros	Acurácia
1	17	75.36 %
2	29	57.97 %
3	19	72.46 %
5	19	72.46 %
7	18	73.91 %
9	19	72.46 %
11	23	66.66 %
13	23	66.66 %
15	27	60.86 %

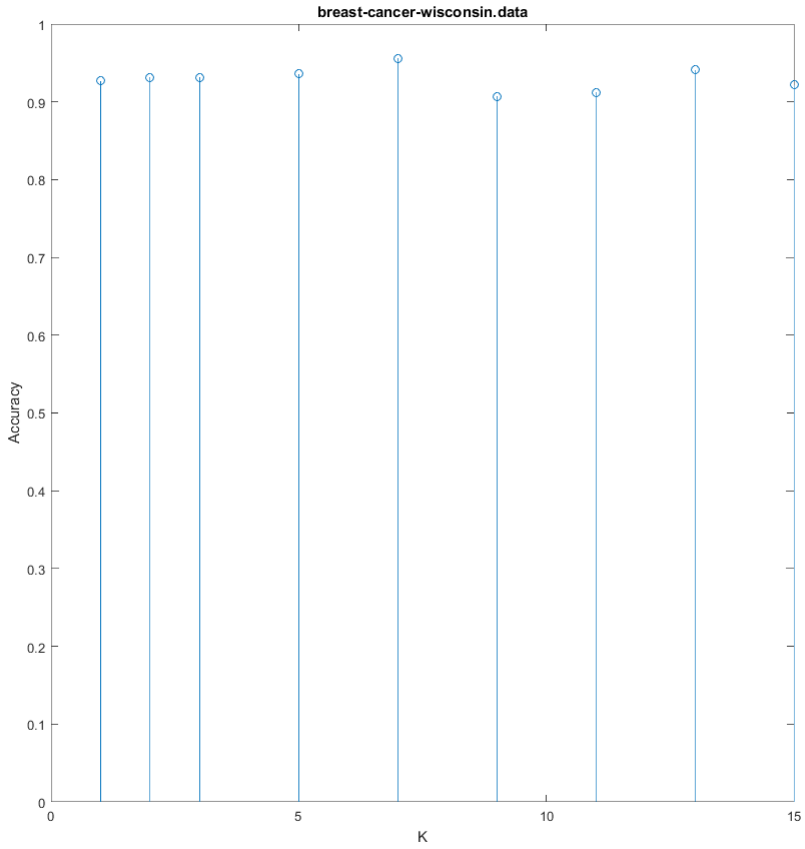

Figure 3: Gráfico da acurácia relacionada ao K utilizado no banco 1 da questão 2

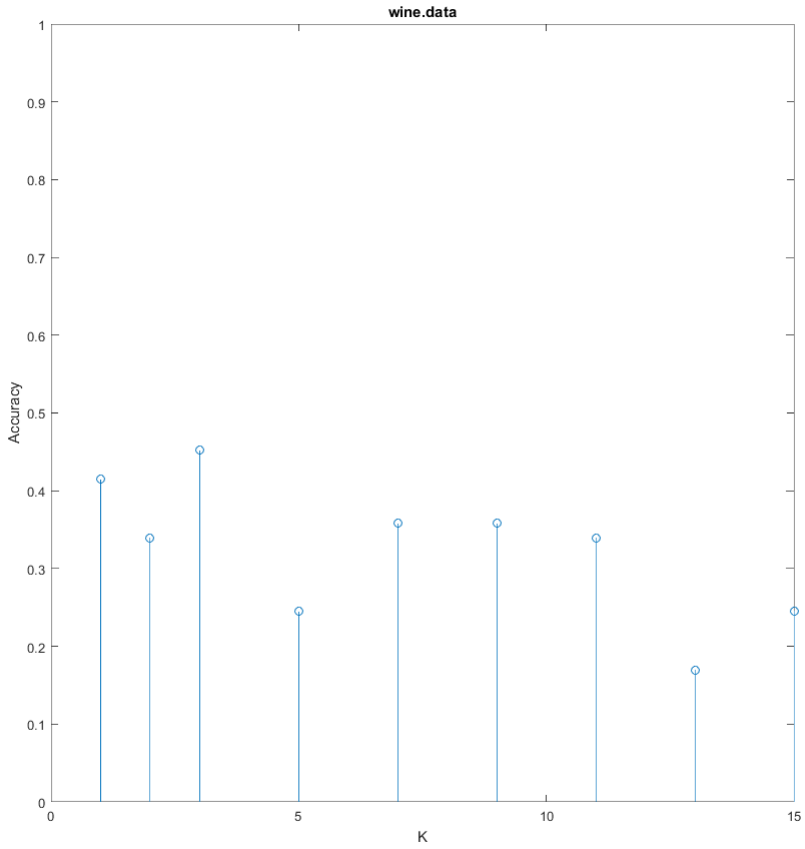

Figure 4: Gráfico da acurácia relacionada ao K utilizado no banco 2 da questão 2