Aprendizagem de Máquina: Atividade 02 − K-NN

Carlos Emmanuel Pereira Alves
Curso de Bacharelado em Ciência da Computação
Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns, Brasil
carlos.emmanuel.236@gmail.com

1) Descreva um problema de classificação para o qual seria adequado utilizar o k-NN e descreva um problema de classificação para o qual não seria adequado utilizar este classificador. Justifique suas escolhas baseado nas vantagens e desvantagens do k-NN. Mostre pelo menos duas vantagens e duas desvantagens para cada exemplo.

Adequado: classificar se um e-mail é spam ou não

Vantagens:

- Simples
- Aprende nova informação facilmente

Desvantagens:

- O conjunto de testes é muito grande pois analisa vários e-mails para poder fazer a classificação, o que vai levar a lentidão.
- Quando o número de dimensões cresce a distância entre o vizinho mais próximo e mais distante se aproximam (Maldição da dimensionalidade), o que seria melhor nesse problema já que gostaríamos de analisar várias palavras.

Não adequado: Detecção de pessoas

Vantagens:

- Aplicável a problemas complexos
- Aproxima o erro ótimo de bayes quando o número de exemplos cresce

Desvantagens:

- Lento para classificar
- Sensível a medida de distância

2) Utilizando a base de dados archive.ics.uci.edu/ml/datasets/iris:

a) Selecione os três exemplos aleatórios de cada classe e construa a matriz de distância entre colocando um exemplo de cada classe como elemento de conjunto de teste e os outros 6 como conjunto de treinamento.

Matriz						
	Exemplo 1	Exemplo 2	Exemplo 3	Exemplo 4	Exemplo 5	Exemplo 6
Exemplo A	1,8	1,0	3,3	3,9	4,3	3,8
Exemplo B	5,1	4,4	0,7	1,1	0,9	0,4
Exemplo C	6,1	5,4	1,3	1,1	0,7	0,8

Exemplo 1	4,8	3	0,1	0,1	Iris-setosa
Exemplo 2	4,3	3	1,1	0,1	Iris-setosa
Exemplo 3	6,4	3,2	4,5	1,5	Iris-versicolor
Exemplo 4	6,9	3,1	4,9	1,5	Iris-versicolor
Exemplo 5	6,2	3,4	5,4	2,3	Iris-virginica
Exemplo 6	5,9	3	5,1	1,8	Iris-virginica

Exemplo A	4,8	3,4	1,9	0,2	Iris-setosa
Exemplo B	5,9	3,2	4,8	1,8	Iris-versicolor
Exemplo C	6,4	2,8	5,6	2,1	Iris-virginica

b) Utilizando a matriz de distância explique a classificação dos exemplos de teste utilizando 1-NN.

Matriz						
	Exemplo 1	Exemplo 2	Exemplo 3	Exemplo 4	Exemplo 5	Exemplo 6
Exemplo A	**1,8**	1,0	3,3	3,9	4,3	3,8
Exemplo B	5,1	4,4	0,7	1,1	0,9	**0,4**
Exemplo C	6,1	5,4	1,3	1,1	**0,7**	0,8

Aqui consideramos o primeiro vizinho mais próximo, ou seja, olhamos o primeiro elemento de menor distância e classificamos o exemplo igual a

esse vizinho. O A seria Iris-setosa, B Iris-virginica e C Iris-virginica. O que vemos que já induziu a erro no Exemplo B.

c) Utilizando a matriz de distância explique a classificação dos exemplos de teste utilizando 3-NN sem peso.

Matriz						
	Exemplo 1	Exemplo 2	Exemplo 3	Exemplo 4	Exemplo 5	Exemplo 6
Exemplo A	**1,8**	**1,0**	**3,3**	3,9	4,3	3,8
Exemplo B	5,1	4,4	**0,7**	1,1	**0,9**	**0,4**
Exemplo C	6,1	5,4	1,3	**1,1**	**0,7**	**0,8**

No 3-NN temos que olhar para os 3 vizinhos mais próximos, e, vamos classificar de acordo com a classe a qual a maioria dos vizinhos pertence.
No Exemplo A temos 2 Iris-setosa e 1 Iris-versicolor, ou seja ela é classificada como Iris-setosa.
No Exemplo B temos 1 Iris-versicolor e 2 Iris-virginica, ela é classificada como Iris-virginica.
No Exemplo C temos 1 Iris-versicolor e 2 Iris-virginica, ela é classificada como Iris-virginica.

d) Utilizando a matriz de distância explique a classificação dos exemplos de teste utilizando 3-NN com peso.

Matriz						
	Exemplo 1	Exemplo 2	Exemplo 3	Exemplo 4	Exemplo 5	Exemplo 6
Exemplo A	**1,8**	**1,0**	**3,3**	3,9	4,3	3,8
Exemplo B	5,1	4,4	**0,7**	1,1	**0,9**	**0,4**
Exemplo C	6,1	5,4	1,3	**1,1**	**0,7**	**0,8**

Aqui o 3-NN é calculado com peso, ou seja, quem está mais próximo tem um peso maior na hora de escolher a classificação, quanto menor a distância para o vizinho maior será a influência dele no peso.
Aqui teremos que fazer o cálculo para poder classificar:

Exemplo A:

- 2 vizinhos da classe Iris-setosa: (1/1,8) + (1/1) = 1,5
- 1 vizinho da classe Iris-versicolor: (1/3,3) = 0,3
- Classificação: Iris-setosa

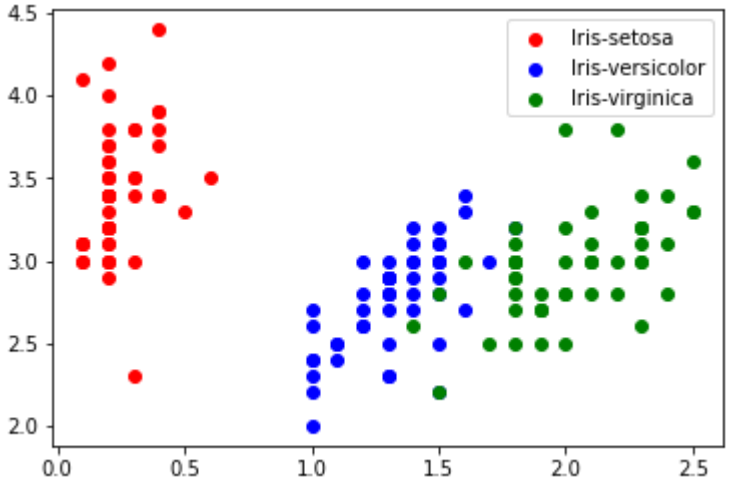
Exemplo B:

- 1 vizinho da classe Iris-versicolor: (1/0,7) = 1,4
- 2 vizinhos da classe Iris-virginica: (1/0,9) + (1/0,4) = 3,6
- Classificação: Iris-virginica

Exemplo C:

- 1 vizinho da classe Iris-versicolor: (1/1,1) = 0,9
- 2 vizinhos da classe Iris-virginica: (1/0,7) + (1/0,8) = 2,6
- Classificação: Iris-virginica

e) Selecione duas características da base Iris construa um diagrama de dispersão colocando símbolos ou cores distintas para cada classe. Características selecionadas: sepal width e petal width.



3) Utilize o classificador pelo vizinho mais próximo utilizando distância euclidiana. Avalie este classificador utilizando metade dos exemplos de cada classe da base Iris como conjunto de teste e o restante como conjunto de treinamento. Utilize uma biblioteca para o classificador 1-NN. Dica: você pode utilizar o sklearn

scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.
A taxa de acerto foi de 94,66%.

4) Utilize os classificadores 7-NN com e 7-NN sem peso e avalie os classificadores utilizando metade dos exemplos de cada classe da base Speaker Accent Recognition como conjunto de teste e a outra metade como conjunto de treinamento. Base: archive.ics.uci.edu/ml/datasets/Wine, arquivo accent-mfcc-data-1.csv. Dica: você pode utilizar o sklearn archive.ics.uci.edu/ml/datasets/Speaker+Accent+Recognition.
7-NN sem peso: taxa de acerto de 72,72%.
7-NN com peso: taxa de acerto de 76,96%.

5) Faça o mesmo da questão anterior para a base Wine archive.ics.uci.edu/ml/datasets/Wine.
7-NN sem peso: taxa de acerto de 61,79%.
7-NN com peso: taxa de acerto de 59,55%.

6) Faça o mesmo da questão anterior removendo a última coluna da base Wine.
7-NN sem peso: taxa de acerto de 68,53%.
7-NN com peso: taxa de acerto de 73,03%.