Universidade Federal de São Carlos
Aprendizado de Máquina
Prof. Tiago A. Almeida <talmeida@ufscar.br>

Exercício Prático 1: Método do K-Vizinhos mais Próximos

Introdução

Neste exercício você implementará o método dos k-vizinhos mais próximos (também conhecido como K-NN, ou *k-nearest neighbors*) e verá como ele utiliza os dados para fazer predições de amostras não vistas. Antes de começar este exercício, é recomendável que você revise os conceitos apresentados em aula.

Arquivos incluídos neste exercício

`ex01.m` – Script geral do exercício
`ex01Dados.mat` – Base de dados 1 com amostras de Iris Setosa e Versicolour
`ex01Dados2.mat` – Base de dados 2 com amostras de Iris Virginica e Versicolour
`visualizarDados.m` – Função para visualizar os dados
[⋆] `normalizar.m` – Função para normalizar as amostras
[⋆] `distancia.m` – Função para calcular a distância Euclidiana entre uma amostra e um conjunto de objetos
[⋆] `knn.m` – Rotina principal do K-NN para predizer a classe de uma amostra qualquer

⋆ indica os arquivos que você precisará completar.

O arquivo `ex01.m` conduzirá todo o processo desse exercício.

O Problema

Você foi contratado por uma grande empresa de cosméticos para desenvolver um método para classificar diferentes espécies de uma flor. Essencialmente, a empresa está interessada em separar automaticamente espécies de uma flor chamada *Iris*. Esse tipo de flor é composta por três espécies: Setosa, Virginica e Versicolour, apresentadas na Figura 1. As duas primeiras (Setosa e Virginica) possuem propriedades aromáticas de interesse da empresa, já a última (Versicolour) não pode ser utilizada.

Devido à forte semelhança visual entre elas, ocorreu a ideia de que, talvez, seja possível detectar cada espécie pelas medidas de comprimento e largura das pétalas. Com base nessa informação, a empresa criou duas base de dados pré-classificadas (Setosa + Versicolour e Virginica + Versicolour) com as respectivas medidas das pétalas das flores. A sua função é implementar o método do K-vizinhos mais próximos para determinar a espécie de uma Iris a partir dos dados das pétalas.

(a) Iris Setosa (b) Iris Virginica (c) Iris Versicolour

Figura 1: Espécies de Iris

Visualização dos Dados

Na primeira etapa do procedimento `ex01.m`, a base de dados com exemplos de Iris Setosa e Iris Versicolour é carregada e as amostras são plotadas em 2D para melhor visualização e interpretação dos dados (veja a Figura 2). Além disso, os valores da amostra $X^{(1)}$ são exibidos. Observe que, a primeira amostra da base de dados é uma Iris com comprimento de pétala igual à 4.9 cm e largura igual à 3.1 cm.

Normalização

Após plotar os dados, o procedimento `ex01.m` chama a função de normalização por padronização. O intuito é reduzir a escalabilidade dos atributos fazendo com que cada atributo X_j fique com média $\mu_j = 0$ e desvio padrão $\sigma_j = 1$ e, com isso, reduzir o impacto no cálculo das distâncias entre as amostras.

No processo de normalização, você deverá criar uma nova base normalizada $Xnorm$ a partir da base original X. Para isso, cada atributo j da amostra i deverá ser calculada por $Xnorm_j^{(i)} = \frac{X_j^{(i)} - \mu_j}{\sigma_j}$, sendo que μ_j corresponde à média de X_j e σ_j, ao desvio padrão de X_j.

Você precisará completar a função ***normalizar.m****.*

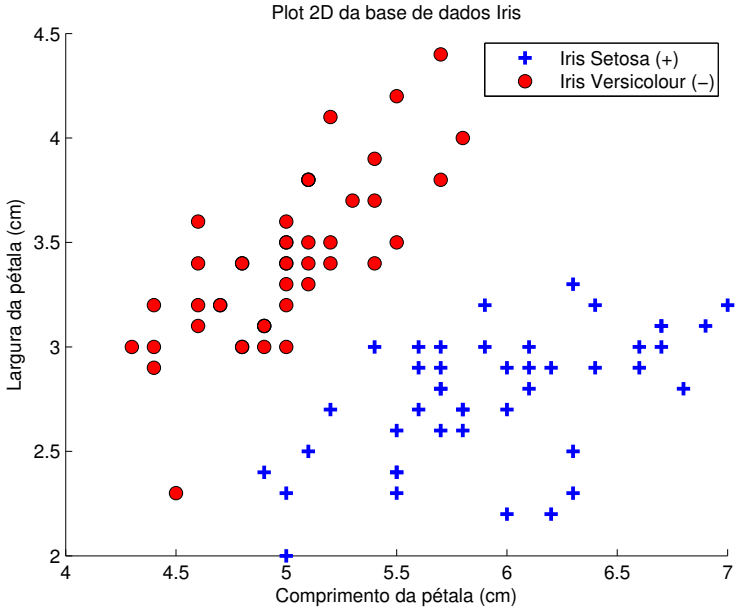

Figura 2: Visualização dos dados

Se a sua função de normalização estiver correta, espera-se que a primeira amostra seja normalizada para [-0.8615 0.0326].

Método do K-vizinhos mais Próximos

A próxima etapa do procedimento `ex01.m` é plotar o caso de teste $x = [5.5000\ 3.2000]$ (Figura 3). Em seguida, a amostra de teste é normalizada e espera-se que os valores dos atributos fiquem iguais à [0.0489 0.2422].

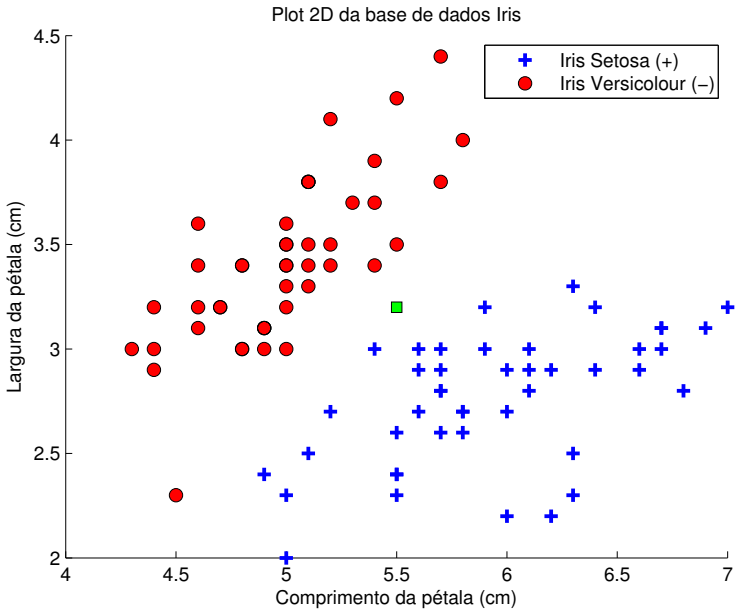

Figura 3: Visualização do caso de teste

Um importante parâmetro do método K-NN é a quantidade de vizinhos K que será usada para a escolha da classe da amostra de teste. Esse parâmetro normalmente é um valor ímpar

1, 3 ou 5.

Para encontrar os K vizinhos de x o método K-NN calcula a distância entre x e cada amostra $i = 1 \dots m$ de $X^{(i)}$. Para tal, primeiro é preciso completar a função `distancia.m`. A rotina em questão deverá receber um vetor qualquer $x_{(1\times n)}$ e uma matriz qualquer $X_{(m\times n)}$ e retornar um vetor de distâncias $D_{(m\times 1)}$, sendo que $D^{(i)} = dist(x, X^{(i)}), i \in \{1 \dots m\}$ e $dist$ corresponde a uma função de distância entre vetores.

Para este exercício, você precisará implementar a função de distância Eucliana, que é dada por:

$$dist(x, y) = \sqrt{\sum (x - y)^2},$$

ou a expressão equivalente

$$dist(x, y) = ||x - y||_2,$$

sendo x e y vetores de mesma dimensão.

Como é necessário calcular a distância entre x e cada amostra de $X^{(i)}$, você poderá usar um `loop-for` para computar o valor de cada $D^{(i)}, i \in \{1 \dots m\}$.

*Você precisará completar a função **distancia.m**.*

Após calcular as distâncias D, é necessário encontrar os K menores valores de D, que correspondem aos objetos mais próximos da amostra de teste e selecionar a classe majoritária entre as K amostras da base como a classe correspondente da amostra de teste (y). Adicionalmente, você precisará retornar os índices $ind\_viz_{(K\times 1)}$ correspondentes às posições (linhas) dos K-vizinhos encontrados em X.

*Você precisará completar a função **knn.m**.*

Se tudo estiver correto, o algoritmo fará a predição y da amostra de teste x e os K-vizinhos serão plotados para visualização (Figura 4).

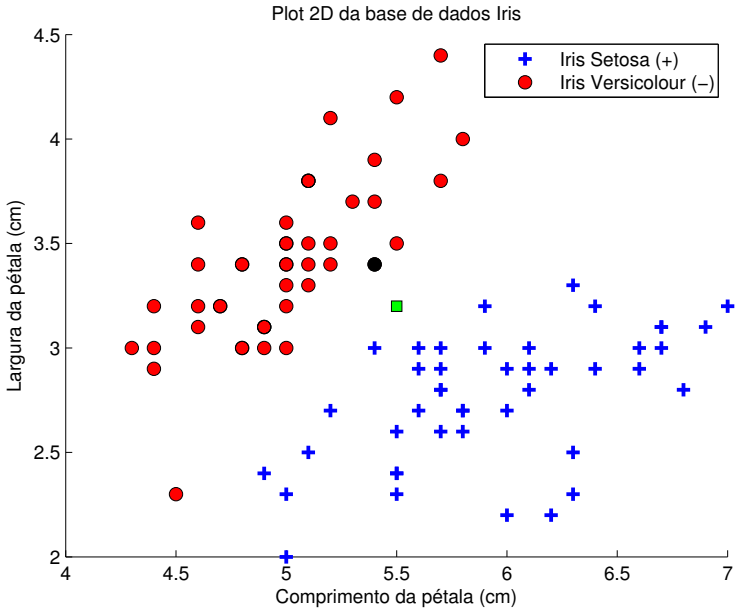

Figura 4: Visualização dos K-vizinhos

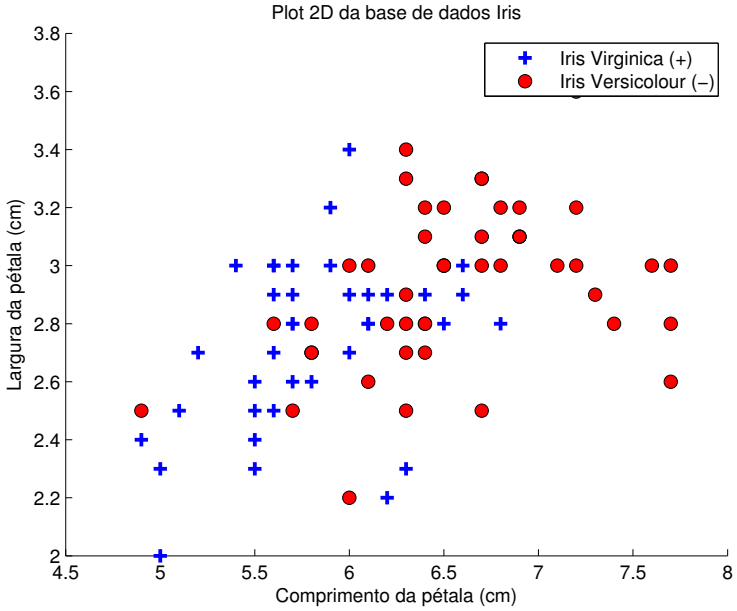

Figura 5: Visualização dos dados

Em seguida, o K-NN é testado para classificar amostras de uma base de dados com exemplos de Iris Virginica e Versicolour (Figura 5).

A amostra de teste escolhida é a [5.9000 2.8400] (Figura6).

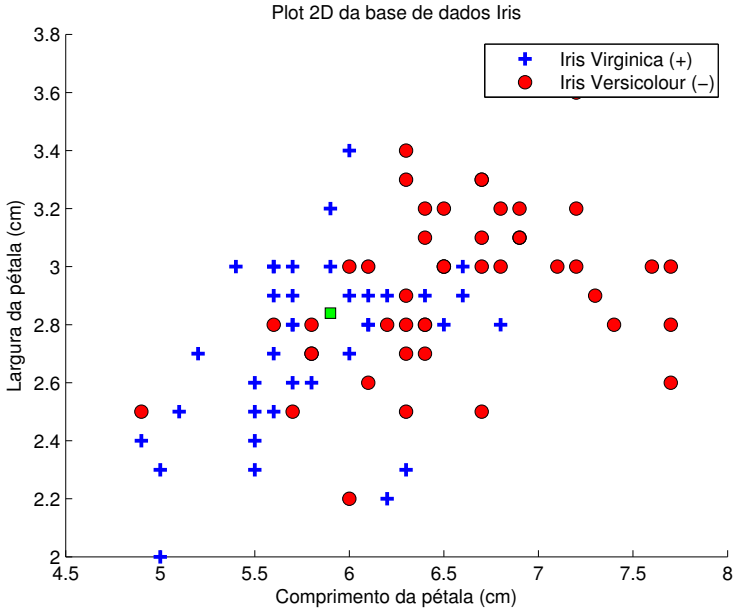

Figura 6: Visualização do caso de teste

Os atributos da base e o caso de teste são normalizados e o K-NN é empregado para fazer a predição. O resultado da classificação pode ser conferido na Figura 7.

Variação de K

Tanto os valores de K quanto as amostras de teste podem ser alterados para ambos os experimentos (base Iris Setosa + Versicolour e base Iris Virginica + Versicolour).

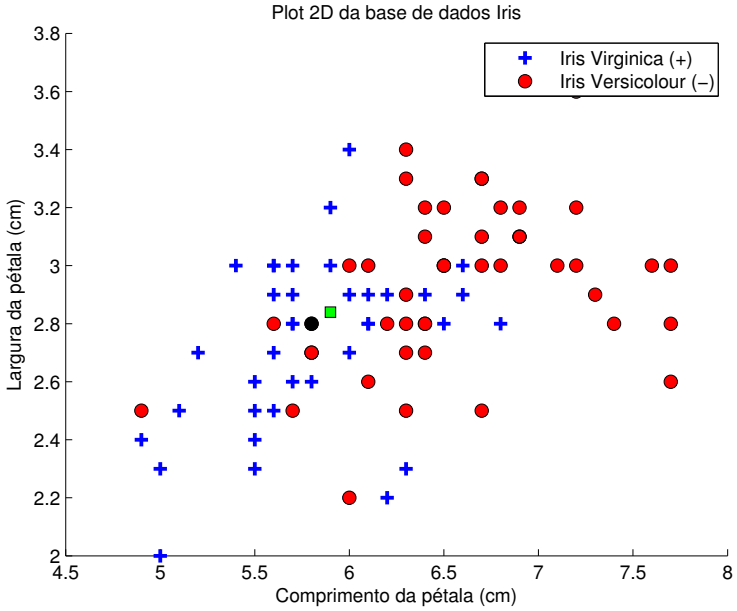

Figura 7: Visualização dos K-vizinhos

É interessante observar que, para as duas bases testadas e os dois casos propostos, a alteração no valor de K ocasiona a alteração da classe de predição dos dois exemplos. Para $K = 1$, ambos os casos de teste são classificados como Iris Versicolour. Porém, para $K = 3$, a primeira amostra de teste é classificada como Iris Setosa (Figura 8) e a segunda como Iris Virginica (Figura 9), e o mesmo ocorre para $K = 5$.

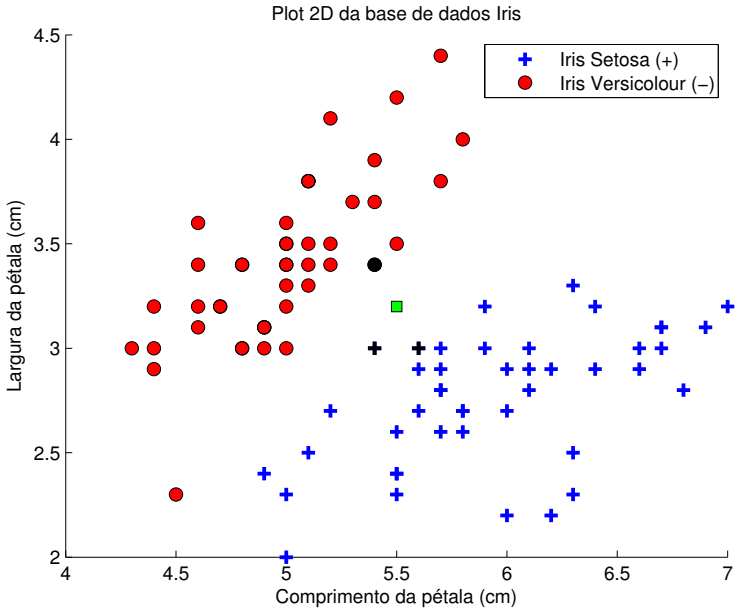

Figura 8: Visualização dos K-vizinhos

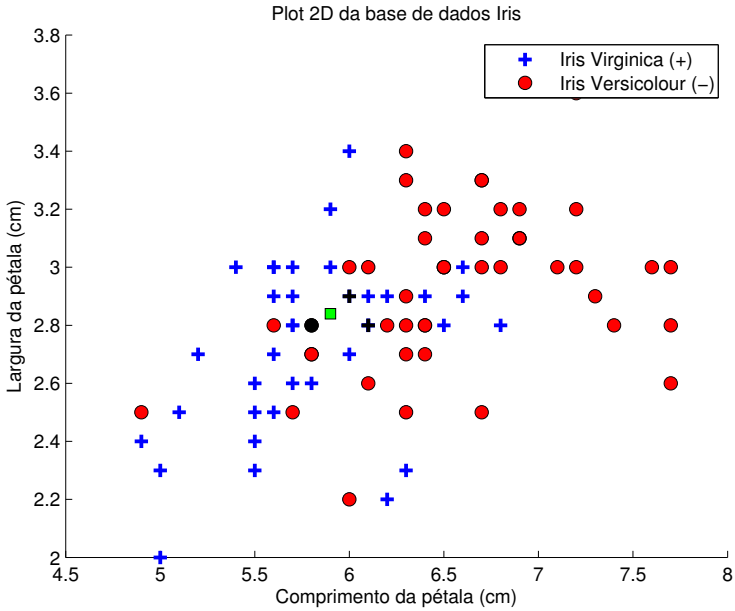

Figura 9: Visualização dos K-vizinhos