Relatório do Trabalho de Inteligência Artificial: Aprendizado de Máquina

Daniel Atkinson Oliveira[114054054]

Universidade Federal do Rio de Janeiro, Rio de Janeiro RJ, Brasil
danatkhoo@gmail.com

Objetivo

O trabalho descrito nesse relatório é uma das atividades da disciplina "Inteligência Artificial"da Universidade Federal do Rio de Janeiro. O objetivo desta atividade é prever as saídas de dois datasets: Iris Plants Database [1] e Discrimination in Salaries [2]. O modelo que foi implementado para a classificação é o perceptron e nenhuma biblioteca de aprendizado de máquina foi usada, uma das restrições impostas sobre o trabalho.

Metodologia

A linguagem escolhida para escrita do código [3] foi R. Nesta seção serão descritas todas as etapas que foram executadas para obter a implementação final do algoritmo de classificação de perceptron para ambos os datasets.

Análise dos Datasets

Antes de começar a treinar o classificador é importante compreender os datasets que serão usados.

O dataset Iris [1] contém três classes, cada uma representa uma espécie de planta. Essas são:

- setosa;
- versicolour;
- virginica.

Para cada uma dessas classes (atributo espécie chamado de "Species"no arquivo iris.dat) temos 50 observações, totalizando 150 observações.

Para cada uma das 3 classes temos 4 atributos numéricos, que são:

- comprimento da sépala ("Sepal.Length"no arquivo iris.dat)
- largura da sépala ("Sepal.Width"no arquivo iris.dat)
- comprimento da pétala ("Petal.Length"no arquivo iris.dat)
- largura da pétala ("Petal.Width"no arquivo iris.dat)

Tabela 1. Estatísticas do dataset.

Atributo	Min	Max	Esperança	DP	Correlação de classes
comprimento sépala	4.3	7.9	5.84	0.83	0.7826
largura sépala	2.0	4.4	3.05	0.43	-0.4194
comprimento pétala	1.0	6.9	3.76	1.76	0.9490
largura pétala	0.1	2.5	1.20	0.76	0.9565

Para cada uma das 150 observações os valores de todos esses atributos são conhecidos.

Estátisticas referentes ao dataset Iris [1] podem ser consultados na Tabela 1

Correlação baixa das classes nos indica que as observações desse atributo aparecem frequentemente em intervalos com bastante interseção entre classes. Podemos observar esse fenômeno na Figura 1

Conversamente, correlação alta das classes nos indica que as observações desse atributo geralmente possuem valores únicos à classe. Podemos observar esse fenômeno na Figura 2

O dataset Salary [2] contém duas classes, referente à variável "sexo"(sx no arquivo .dat):

- masculino (codificado como 0 no arquivo salary.dat)
- feminino (codificado como 1 no arquivo salary.dat)

Ao contrário do dataset Iris [1] não temos um número de observações iguais para cada classe.

Cada classe tem 5 atributos, que são:

- ranking (rk no arquivo salary.dat, com a seguinte codificação: 1 para professor assistente, 2 para professor associado e 3 para professor efetivado)
- número de anos no ranking atual (yr no arquivo salary.dat)
- maior grau atingido (dg no arquivo salary.dat, com a seguinte codificação: 0 para mestrado e 1 para doutorado)
- número de anos desde que atingiu grau mais alto (yd no arquivo salary.dat)
- salário acadêmico, em dólares (sl no arquivo salary.dat)

Como ambos os datasets tem quantidade de classes diferentes, diferentes estratégias serão usadas para a previsão de classe em cada um.

Validação Cruzada

Tanto para o dataset Iris [1] quanto para o dataset Salary [2], foi feita a validação cruzada 5-fold. Para isso ambos os datasets foram divididos, aleatóriamente, em 5 conjuntos com a mesma quantidade de observações. Uma rodada de validação cruzada 5-fold consiste em 4 dos 5 conjuntos serem utilizados para treino e 1 ser utilizado para teste. Foram feitas 5 rodadas de validação cruzada 5-fold, tal que o conjunto utilizado para teste muda a cada rodada, assim como o conjunto de treino. O processo está ilustrado na Figura 3.

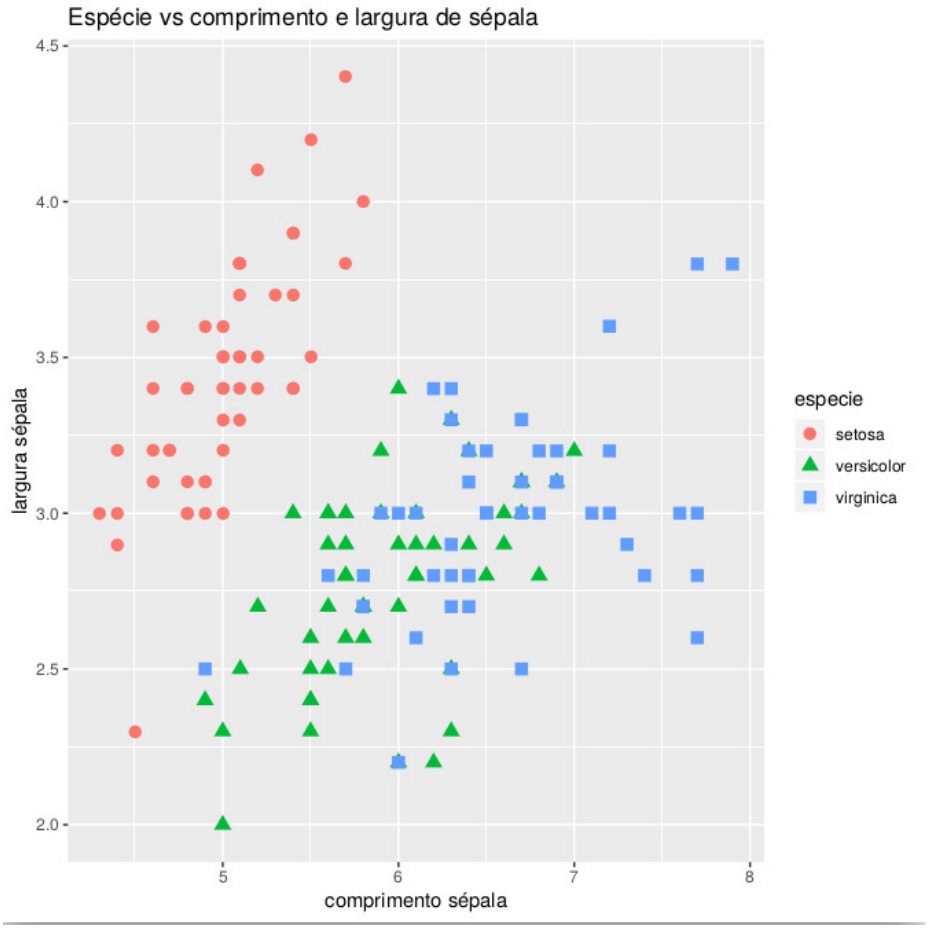

Figura 1. Podemos ver que, tanto no atributo comprimento sépala quanto em largura sépala, há bastante interseção entre os valores

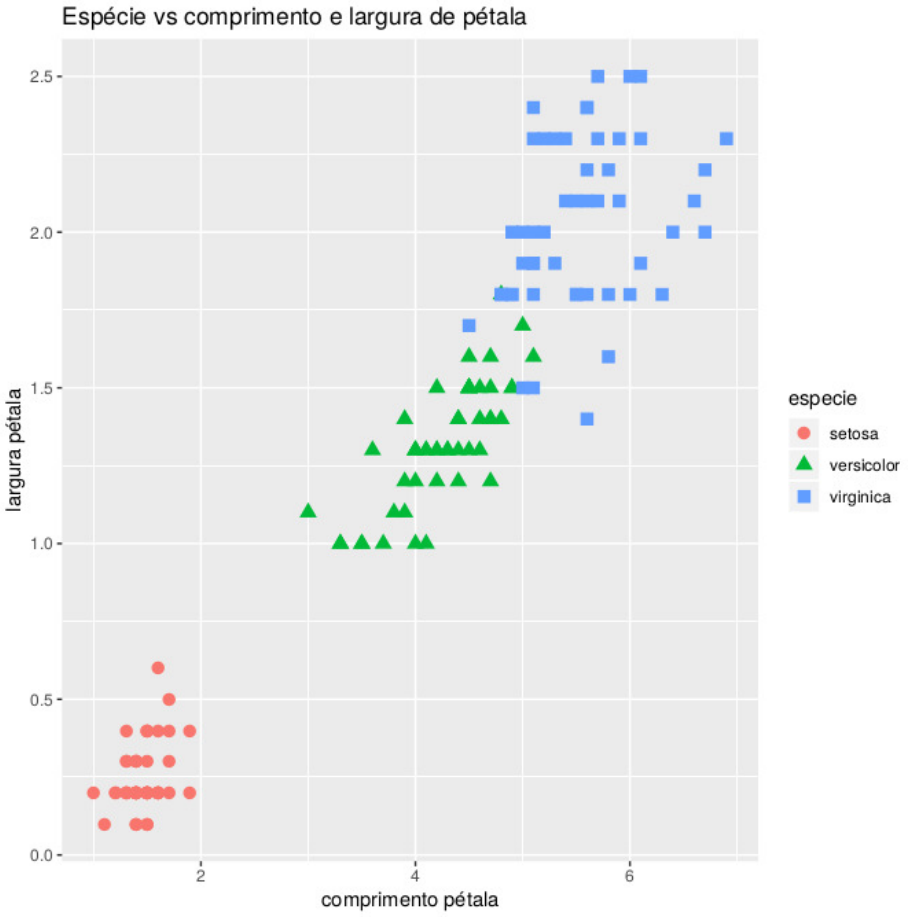

Figura 2. Podemos ver que, tanto no atributo comprimento pétala quanto em largura pétala, não há muita interseção entre os valores

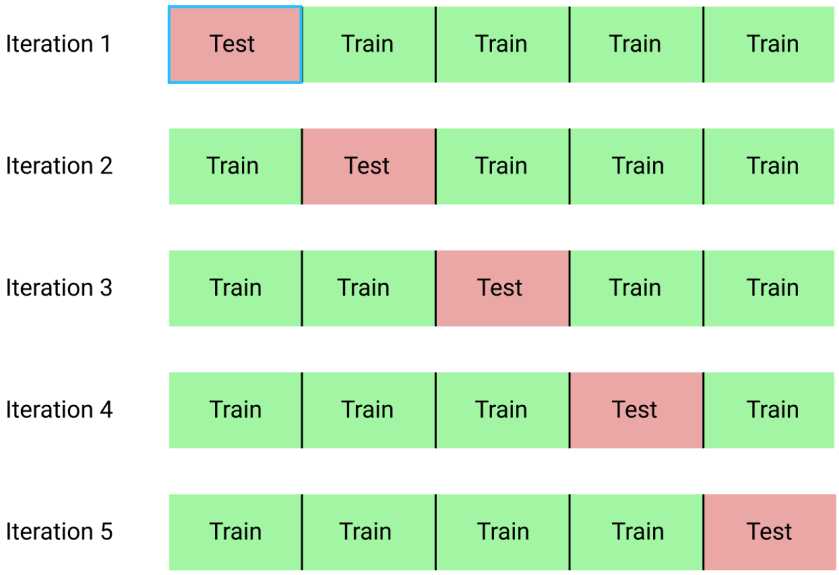

Figura 3. Processo de valicação cruzada 5-fold

Perceptron

Como dito anteriormente, por causa da diferença no número de classes em cada dataset duas técnicas diferentes serão usadas para implementar o algoritmo de perceptron, uma para cada dataset.

Como o dataset Salary [2] possui duas classes (0 ou 1), podemos implementar o perceptron convencional, que faz classificação binária. Para cada uma das 5 rodadas da validação cruzada 5-fold, uma matriz de treino $X_{120\times4}$, um vetor de classes c, tal que $c_i = 0, 1, i = 1, 2, ..., 120$ e n o número de épocas que o perceptron será treinado, são passados como argumentos para a função.

Resultados

Conclusão

Referências

1. Iris Plants Database
   `https://archive.ics.uci.edu/ml/machine-learning-databases/iris/`
2. Discrimination in Salaries
   `http://data.princeton.edu/wws509/datasets/#salary`

3. Código do Perceptron
   `http://data.princeton.edu/wws509/datasets/#salary`