Inteligência Artificial

Thiago Henrique Leite da Silva, RA: 139920

AULA9: Exercício prático aprendizado supervisionado

(1,0) Implemente o algoritmo KNN, Naive Bayes e Hunt e aplique no dataset IRIS:

https://www.kaggle.com/uciml/iris

Separe aleatoriamente 70% dos dados para treino e 30% para teste e reporte com um print da saída qual a acurácia do algoritmo (número de acertos).

Parte 01: Algoritmo KNN

A linguagem escolhida para o problema em questão desta vez foi Java, pois é uma linguagem que me agrada bastante, portanto foi treinar desenvolvimento nela, além de aproveitar as vantagens da orientação a objetos.

Utilizei uma biblioteca JXL do Java para ler o dataset diretamente da planilha, como forma de evitar o cansativo processo de inserirmos as entradas manualmente. Esta biblioteca pode ser baixada neste link:

http://jexcelapi.sourceforge.net/

O algoritmo funciona da seguinte maneira:

Fornecemos o path para nossas duas planilhas, uma contendo o dataset, onde utilizei apenas 70% dele, que totalizam 105 registros de espécies de Íris; e outra para os elementos de teste com os demais 30%, que totalizam 45 espécies. Além disso fornecemos também o valor K que será nosso delimitador para avaliar os melhores vizinhos.

Feito isso, podemos ler as planilhas, o que fazemos é percorrer cada uma delas salvando os atributos lidos em ArrayLists, pois serão a partir deles que teremos acesso à informação.

Definida estas funcionalidades do algoritmo, nós temos duas formas de utilização, uma onde rodamos os testes, rodando o método knnAlgorithm(), e outra onde o usuário fornece os atributos e o algoritmo classifica a espécie com base nos dados lidos; este é chamado pelo método perform().

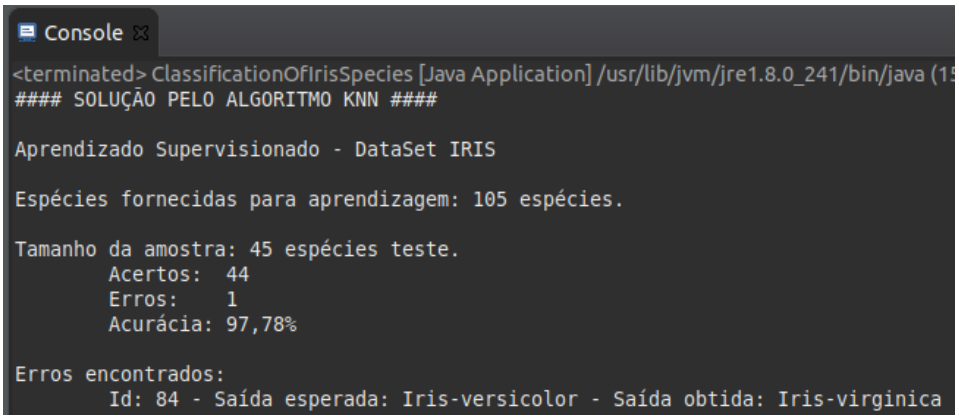
Em relação ao algoritmo, o que fazemos primeiramente é iterarmos sobre todos os elementos salvos nos Arrays com os valores teste, onde a cada elemento i, calculamos a distância do mesmo em relação a todos os elementos do dataset, então, ordenamos as distâncias, pegamos as K menores, e contabilizamos qual a espécie com maior ocorrência dentre as que sobraram, desta forma, o resultado disto será a classificação da espécie.

Então, nós imprimimos os resultados, que carregam:

Quantidade de acertos e erros nos testes, quantidade de dados testados e utilizados no treinamento, além de imprimir os erros, caso existam.

Em relação aos erros, imprimimos o ID da Íris, para uma melhor identificação, a saída esperada e a saída obtida.

Executando o código para o dataset fornecido, temos os seguintes resultados:



Em relação ao algoritmo de Bayes e Hunt, apesar de serem algoritmos relativamente simples, não tive tempo para implementar professora, estou tendo muita dificuldade para conseguir conciliar o tempo do estágio, matérias da Unifesp e descanso, por isto acredito estar rendendo menos do que poderia na disciplina, que por sinal está sendo a que mais estou gostando neste semestre. Aprendi bastante sobre Java implementando este algoritmo do Knn, porém a questão da documentação deixei de lado para conseguir implementar os outros, mas não obtive sucesso. Espero poder entregar um melhor resultado nas próximas aulas.