2ª Avaliação de Inteligência Artificial

Professor: Jânio Coutinho Canuto Período: 2015-2

A última avaliação da disciplina aborda os conceitos de Aprendizagem Supervisionada e Não-Supervisionada. Para executar os experimentos, utilizaremos os dados do seguinte arquivo:

http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

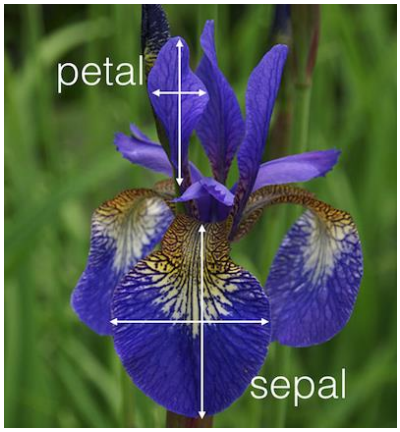
Esta é, provavelmente, a base de dados mais usada na literatura de reconhecimento de padrões. O artigo de Fisher ("The Use of Multiple Measurements in Taxonomic Problems" de 1936) é um clássico deste campo e é frequentemente citado até hoje. Os dados são referentes a 3 tipos de plantas (*Iris Setosa*, *Iris Virginica*, *Iris Versicolor*), com 50 amostras de cada tipo. É um conjunto bem simples de se trabalhar, e adequado para verificar a aprendizagem dos conceitos fundamentais.

Iris Setosa

Iris Virginica

Iris Versicolor

Cada linha do arquivo é uma instância (exemplo), com 5 elementos separados por vírgulas. Os primeiros quatro elementos (numéricos), são as características medidas nas plantas (comprimento da sépala – ***cs***, largura da sépala – ***ls***, comprimento da pétala – ***cp***, e largura da pétala – ***lp***, tudo em centímetros), o último elemento (texto) diz o tipo da planta (os primeiros 50 são *Setosa*, os 50 seguintes *Versicolor* e os 50 últimos *Virginica*).



Algumas regras:

- as implementações podem ser feitas na linguagem de programação de sua preferência;
- siga o roteiro e apresente os resultados e respostas às perguntas de forma organizada;
- os códigos deverão ser entregues;
- tudo deverá ser entregue de forma eletrônica;
- os trabalhos poderão ser feitos em grupos de até 10 pessoas;
- a data limite para entrega é 13/05;

Roteiro de Atividades:

1. **Visualização:** Esta é uma etapa comumente ignorada, mas de extrema importância. Muitas vezes é possível perceber padrões antes mesmo de decidir o que fazer. Muita gente aplica os métodos de reconhecimento de padrões sem sequer olhar para os dados que tem. Nosso conjunto de dados tem 4 dimensões, não dá para desenhar.
    a. ***Faça um gráfico de cada uma das 6 projeções bi-dimensionais possíveis:***
        i. ***cs*** x ***ls;***
        ii. ***cs*** x ***cp;***
        iii. ***cs*** x ***lp;***
        iv. ***ls*** x ***cp;***
        v. ***ls*** x ***lp;***
        vi. ***cp*** x ***lp.***
    b. ***Responda:***
        i. Você percebe algum padrão nos dados?
        ii. Existe algum grupo que é mais fácil de separar?
        iii. Qual dimensão parece ser mais informativa? (útil para classificação)

2. **Aprendizagem Não-Supervisionada (agrupamento de dados):** Nesta forma de aprendizagem os rótulos de classe (quinta coluna do nosso arquivo) não são utilizados no treinamento, apenas as características. Os algoritmos buscam "regularidades" nos dados.
    a. ***Implementação:***
        i. Aplique um agrupamento hierárquico (aglomerativo ou divisivo) aos dados.
        ii. Aplique o algoritmo k-means aos dados.
    b. ***Perguntas para os dois algoritmos:***
        i. Os grupos resultantes correspondem aos tipos das plantas?
        ii. Há algum grupo mais destacado dos demais?
        iii. A escolha da medida de distância entre grupos altera o resultado obtido? (se não souber responder teoricamente, teste)
        iv. Se você usar somente uma das características para realizar o agrupamento ainda é possível identificar grupos?
            1. Qual a melhor característica a escolher?
            2. Que utilidade isto teria?
        v. Para o k-means, o que acontece se:
            1. k = 2
            2. k = 4

3. **Aprendizagem Supervisionada:** Neste caso utilizamos as informações que já possuímos sobre qual tipo de planta cada exemplo é para treinar um classificador.
    a. ***Implementação:*** para o treinamento dos classificadores é preciso pensar em como representar os dados, tanto as características quanto os rótulos de classe.
        i. Treine uma árvore de decisão utilizando os dados.
        ii. Treine uma rede neural MLP utilizando os dados.
    b. ***Para a árvore de decisão:***
        i. Qual o primeiro "corte" feito pela árvore? Você esperava isso?
        ii. É possível classificar perfeitamente todos os pontos? Quantas perguntas são necessárias?
        iii. Escolha N subconjuntos aleatórios de exemplos e treine N árvores com no máximo K perguntas cada. Classifique os pontos usando as N árvores ao mesmo tempo e diga que ele pertence à classe que a maioria das árvores concordam.
            1. Como foi o resultado?
            2. Você consegue perceber como as escolhas de N e K afetam o resultado e o custo computacional?
    c. ***Para a rede neural:***
        i. Como o número de neurônios na camada escondida afeta o desempenho?
        ii. Como a escolha na forma de representar a saída (rótulos) afeta o desempenho?
        iii. Como a escolha na forma de representar a entrada afeta o desempenho?
    d. ***Validação:*** Utilizar métodos supervisionados para realizar classificações perfeitas de todos os exemplos é interessante para entender o funcionamento dos algoritmos, mas pouco útil na prática. O que gostaríamos é que os classificadores fossem capazes de generalizar a partir de alguns exemplos.
        i. Separe 70% dos dados para treinamento e 30% para teste.
        ii. Refaça o treinamento dos calssificadores utilizando apenas esse subconjunto de dados e verifique qual o desempenho no subconjunto de teste.
        iii. Quando devemos parar de treinar?
        iv. Já que o K-NN (K vizinhos mais próximos) não precisa de treinamento, vamos testá-lo agora. Classifique os dados de teste usando KNN com os dados de treinamento como referência.
        v. A escolha dos dados de treinamento e teste afeta os resultados obtidos com cada um dos métodos?