ACH2016 Inteligência Artificial

Exercício Programa 2 (EP2)
Classificação
Prazo limite de entrega no Tidia: 17/06/2019 às 23:55

Composição dos grupos: de 1-4 alunos
Enviar no Tidia o código fonte e o relatório em formato pdf

1. Base de dados

Escolher uma das seguintes bases de dados:

1.1 Spam
A base de dados Spam (http://archive.ics.uci.edu/ml/datasets/Spambase) contém informação de mensagens de e-mail de spam e não-spam a qual está disponível no repositório UCI Machine Learning (http://www.ics.uci.edu/~mlearn/MLRepository.html).

Esta base de dados contém 57 atributos numéricos e a classe de email que é nominal e pode ser spam (1) ou não spam (0). A base contém 4601 exemplos de e-mails dos quais 1.813 são de Spams (39.4%) e 2.788 são de Não-Spams (60.6%).

1.2 Transportation Mode Detection with Unconstrained Smartphones Sensors

A base de dados TMD contém informação de modos de transporte a qual está diponível em **https://www.kaggle.com/fschwartzer/tmd-dataset-5-seconds-sliding-window¶**.

Esta base de dados contém 13 atributos numéricos e a classe de transporte que é nominal, que pode ser Bus, Car, Still, Train e Walking. A base contém 5893 exemplos.

2. Conjunto de treinamento e conjunto de teste

A partir da base de dados completa escolhida, selecione uma amostra e coloque 70 % dos exemplos no conjunto de treinamento e 30% dos exemplos no conjunto de teste. Tente manter a mesma proporção de exemplos de cada classe do conjunto de dados completo.

3. KNN

Faça o seguinte:

- Execute uma implementação do algoritmo KNN com K=3 usando distância Euclidiana e distância de Manhattan. Calcule a acurácia.

- Execute uma implementação do algoritmo KNN com valores de K de 1 até 10 usando distância Euclidiana. Calcule a acurácia para analisar o efeito do parâmetro K na classificação. Faça um gráfico de K versus acurácia.
- Realize a padronização dos atributos (com média 0 e variância 1) e execute uma implementação do algoritmo KNN com o melhor valor de K encontrado no item anterior.

4. Naive Bayes

Execute uma implementação do Naive Bayes, por exemplo, a implementação disponível no simulador Waikato Environment for Knoledge Analysis -- WEKA (http://www.cs.waikato.ac.nz/ml/weka). Para tal faça o seguinte:

- Discretize os atributos de valor contínuo, isto é, para cada atributo, pegar o intervalo definido por [valorMin, valorMax] e dividí-lo em 6 intervalos de tamanho aproximadamente iguais.
- Treine um classificador Naive Bayes com e sem padronização, isto é a partir dos dados de treinamento calcular as probabilidades necessárias. Utilize o classificador Naive Bayes em cada exemplo do conjunto de teste e calcule a acurácia.

Obs: Substituir qualquer probabilidade igual a 0 por 0,00001 para evitar a multiplicação por zero.

5. Árvores de Decisão

Execute um algoritmo de árvores de decisão (por exemplo o J4.8 do WEKA) sem poda e depois com poda sobre o conjunto de treinamento com e sem padronização. Utilize o conjunto de teste para avaliar as árvores geradas e calcule a acurácia.

6. Verifique qual dos classificadores é mais robusto com relação à presença de ruídos

Escolha aleatoriamente 0%, 5%, 10%, 15%, 20% e 25% dos exemplos do conjunto de treinamento e mude o valor de suas classes, por exemplo na base de dados Spam mude o valor de 0 para 1 e vice-versa. Execute os algoritmos KNN com o melhor valor de K encontrado, Naive Bayes e Árvore de Decisão com o conjunto de teste e calcule a acurácia. Faça um gráfico da percentagem de ruído versus a acurácia.

7. Relatório

Elabore um relatório analisando os resultados encontrados de até 10 páginas.

Trabalho baseado em:
http://www.ppgia.pucpr.br/~alekoe/AM/2009/TrabalhoArvoresDecisao-AM2009.pdf