Projeto Aprendizado de Máquina

CTC- 17 Inteligência Artificial
(Trabalho em Dupla)
Prof. Paulo André L. Castro

1. Objetivo

Exercitar e fixar conhecimentos adquiridos sobre Aprendizado de Máquina utilizando árvores de decisão e/ou redes bayesianas ingênuas (*Näive Bayes*) utilizando uma base de dados fornecida.

2. Descrição do Trabalho

2.1. Base de dados (dataset)

Opção 1: Usar as informações de classificações de filmes (1 a cinco estrelas) para os filmes e usuários fornecidos no pacote de dados (ml-1m.zip) obtidas por MovieLen (http://grouplens.org/datasets/movielens/) e disponibilizada no site da disciplina. A base de dados ml-1m oferece informações sobre usuários (UserID::Gender::Age::Occupation::Zip-code), filmes (MovieID::Title::Genres) e as classificações dadas (UserID::MovieID::Rating::Timestamp), aproximadamente 1 milhão de classificações dadas por 6000 usuários para 4000 diferentes filmes. Veja o arquivo README para mais detalhes. O objetivo é dadas as informações sobre o usuário, sugerir filmes que ele irá apreciar (alta classificação).

Opção 2: Selecionar a sua escolha um dataset com no mínimo 50000 instâncias e 5 variáveis para classificação binária ou multi-classe. Descrever o dataset e o objetivo esperado. Esta opção é considerada mais complexa que a opção 1 e esta complexidade adicional será considerada na avaliação. Uma boa fonte de datasets é https://archive.ics.uci.edu/ml/index.php.

2.2. Classificador baseado em árvore de decisão (ou Naive Bayes)

Utilizando a base de dados fornecida, criar um **classificador baseado em árvore de decisão** que classifique um dado filme com base nas informações disponíveis sobre o filme e sobre o usuário (gênero, idade e ocupação) . Você pode optar por um classificador baseado na técnica ***Naive Bayes,*** ao invés de árvore de decisão.

2.3. Classificador a priori

Crie um classificador *a priori*, isto é que não usa nenhuma informação. Isto é aponta como classificação a média truncada ou a moda das classificações dos filmes (ou instâncias).

2.4. Análise Comparativa

Compare os dois classificadores utilizando: taxa de acerto, matriz de confusão, erro quadrático médio e estatística kappa.

Para fazer a comparação, selecione pelo menos dez filmes (ou instâncias) que você assistiu dentro da base e dê sua classificação em estrelas. Faça isso antes de ver a avaliações dadas para cada um destes filmes (ratings.dat). Discuta qual classificador entre os dois (2.2 ou 2.3) é melhor. Proponha alguma alteração no classificador construído no item 2.2 que poderia torná-lo um classificador ainda melhor.

3. Material a ser Entregue e Prazo

Material: Relatório e Código Prazo de Entrega: 24/setembro/2018

Relatório do Projeto (**arquivo em formato pdf até 4 páginas**) com:

Título e Nomes dos integrantes do Grupo

2. Descrição e Resultados Obtidos

2.1. Descrição dos Classificadores e do dataset (se for opção 2)

2.2. Dados e Resultados da comparação

2.3. Discussão e sugestão de melhorias para o classificador

3. Conclusões: Comentários e sugestões sobre o trabalho (complexidade/facilidade, sugestões, etc.).

4. Descrição da Implementação: Linguagem e IDE utilizados, outros comentários eventualmente necessários para a execução do projeto.

Código do Projeto : Código-fonte do Sistema (em C, C++, C#, Julia, Python ou Java).

Bom Trabalho!

Prof. Paulo André Castro
pauloac@ita.br