Projeto Preparação de Dados e Árvore de Decisão
CMC-13 Introducão a Ciência de Dados
(Trabalho em Grupo de dois, três ou quatro alunos)

Prof. Paulo André Castro

1. Objetivo

Exercitar e fixar conhecimentos adquiridos sobre: preparação de dados e criação de modelos e avaliação de desempenho utilizando uma base de dados fornecida. Podem ser utilizados livremente frameworks para implementação dos métodos de aprendizado de máquina. Sugere-se o uso do scikit learn (sklearn), pandas, numpy e matplolib mas outros também podem ser usados.

2. Descrição do Trabalho

2.1. Base de dados (dataset)

O dataset inclui revisões de livros (0 a 10) feitas por vários usuários. Há catorze atributos e 131179 linhas de dados. Os atributos são os seguintes:

- 'user_id' : identificador do usuário (numérico)
- 'age',: idade do usuário
- 'isbn' : identificador do livro
- 'rating': classificação do livro dado pelo usuário (0 a 10)
- 'book_title' - Título do livro em inglês,
- 'book_author': Nome do autor do livro
- 'year_of_publication' : Ano de publicação do livro
- 'publisher',: Editora
- 'img_l': Link para imagem de capa do livro
- 'Language': Idioma no qual foi escrito o livro
- 'Category': Tipo de livro, observe que um livro pode pertencer a mais de um tipo (string)
- 'city': Cidade do usuário (identificado por user_id)
- 'state': Estado do usuário
- 'country': País do usuário

Observe que há dois arquivos de dados (dados_treinamento.csv, dados_teste.csv), com mesmo formato, que devem ser usados no treinamento e teste dos modelos, respectivamente. Não utilize os dados de teste para ajuste de hiperparâmetros, faça isto usando apenas os dados de treinamento.

2.2. Tarefas a Realizar

1. Preparação dos Dados

Avalie se todos os campos são úteis para o trabalho. Se houver campos não úteis, exclua-os dando justificativa. Prepare os dados para serem apresentados aos modelos de classificação.Os dados podem ter atributos faltantes ou com imprecisões em seu valor (ruído).

2. Análise Exploratória e Visualização dos Dados

Selecionados os campos relevantes para o trabalho, avalie cada atributo/campo utilizando alguma ferramenta gráfica (boxplot, scatter-plot, gráfico de linha, gráfico de barra ou histograma). Não é necessário utilizar todos os tipos de gráficos no trabalho, mas selecione pelo menos dois tipos de gráficos diferentes, de acordo com aquilo que vc julgar mais adequado para este dataset. Verifique se algum atributo necessita ser transformado (normalizado, discretizado, etc) e se for o caso, faça a transformação.

3. Modelo baseado em Árvore de Decisão ou KNN.

Crie um modelo usando a técnica escolhida para classificar o livro (0 a 10), dadas as outras informações da linha (todos atributos exceto rating).

4. Classificador a priori

Crie um classificador a priori (as vezes chamado de zero regra), isto é que não usa nenhuma informação além da própria identificação do livro. A classificação é a média truncada ou moda das avaliações disponíveis para aquele livro.

5. Análise Comparativa do desempenho dos modelos.

Avalie comparativamente os dois modelos, utilize medidas apropriadas de desempenho de modelos (acurácia, precision, recall), utilizando os dados de teste. Discuta os resultados e qual seria o modelo mais apropriado, para um classificador automático de livros.

Verifique o desempenho nos dados de treinamento e teste. Há variação de desempenho significativa? Em caso positivo, explique porquê.

3. Material a ser Entregue e Prazo

Devem ser entregues um relatório e um notebook com o código

Entregar através do Google Classroom!

OBS: Não compacte os arquivos em um zip (ou qq outro formato), faça os uploads dos dois arquivos!

A. Relatório em formato pdf (ver detalhes abaixo)

B. Código em formato Notebook (ve detalhes abaixo)

Prazo de Entrega: 21/abril/2023;

Item A: Relatório

Estrutura do Relatório do Projeto (arquivo em formato pdf)

Título: CMC-13 Preparação de Dados

Equipe: Nomes do membros da Equipe

1. Preparação dos dados

Descrever procedimentos realizados para concluir esta tarefa

2. Análise Exploratória e Visualização dos Dados

3. Modelo baseado em Árvore de Decisão ou KNN

Descrever procedimentos realizados para concluir esta tarefa

4. Classificador a priori

Descrever procedimentos realizados para concluir esta tarefa

5. Análise Comparativa do desempenho dos modelos.

Apresente os dados e discussões sobre os resultados, inclusive dados sobre o desempenho no dataset de treino e testes.

Item B: Código do Projeto

Código do Projeto (Formato jupyter notebook, Linguagem: Python , R ou Julia)

Siga a estrutura do relatório, para organizar o código no notebook

1. Preparação dos dados

Códigos correspondentes

2. Análise Exploratória e Visualização dos Dados

Códigos correspondentes

3. Modelo baseado em Árvore de Decisão ou KNN

Códigos correspondentes

4. Classificador a priori

Códigos correspondentes

5. Análise Comparativa do desempenho dos modelos.

Códigos correspondentes

.

Bom Trabalho!

Prof. Paulo André Castro