Lab. Exame

CMC-13 Introdução 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 Ciência de Dados e preparação de dados utilizando uma base de dados fornecida. ***Neste trabalho, podem ser utilizados livremente frameworks para implementação dos métodos de aprendizado de máquina. Sugere-se o uso do scikit learn, mas outros também podem ser usados.***

2. Descrição do Trabalho

2.1. Base de dados (dataset)

O dataset inclue 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 (exame_cmc13_dados_treinamento.csv, exame_cmc13_dados_teste.csv), com mesmo formato, que devem ser usados no treinamento e teste dos modelos, respectivamente.

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. Modelo baseado em Redes Neurais do tipo MLP (MultiLayer Perceptron)

Crie um modelo baseado em Redes Neurais do tipo MLP (MultiLayer Perceptron) para classificar o livro (0 a 10), dadas as outras informações da linha (todos atributos exceto rating).

3. Modelo baseado em Árvores de Decisão ou Em Florestas Aleatórias (Random Forests)

Crie um modelo baseado em Redes Neurais do tipo MLP (MultiLayer Perceptron) para classificar o livro (0 a 10), dadas as outras informações da linha (todos atributos exceto rating).

4. Análise Comparativa do desempenho dos modelos.

Avalie comparativamente os três modelos, utilize medidas apropriadas de desempenho de modelos (acurácia, precision, recall, F1-score, Kappa statistic, etc). 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: 15/julho/2022;

Item A: Relatório

Estrutura do Relatório do Projeto (arquivo em formato pdf)

Título: Lab. 3 - Aprendizado de Máquina Probabilístico
Equipe: Nomes do membros da Equipe

1. Preparação dos dados
Descrever procedimentos realizados para concluir esta tarefa
2. Modelo baseado em Redes Neurais do tipo MLP (MultiLayer Perceptron)
Descrever procedimentos realizados para concluir esta tarefa
3. Modelo baseado em Árvores de Decisão ou Em Florestas Aleatórias (Random Forests)
Descrever procedimentos realizados para concluir esta tarefa
4. 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. Modelo baseado em Redes Neurais do tipo MLP (MultiLayer Perceptron)
Códigos correspondentes
3. Modelo baseado em Árvores de Decisão ou Em Florestas Aleatórias (Random Forests)
Códigos correspondentes
4. Análise Comparativa do desempenho dos modelos.
Códigos correspondentes

.
Obs:
Use células markdown antes de cada célula de código, para descrever qual o propósito do código em seguida

Bom Trabalho!
Prof. Paulo André Castro