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 uma base de dados fornecida.

2. Descrição do Trabalho

2.1. Base de dados (dataset)

O dataset incluem dados sobre apartamentos para aluguel nos EUA e seus preços (em dólares). Utilizando os dados disponíveis, prepare os dados e separe-os em dois conjuntos: treino e validação (para ajsutes de hiperparâmetros). Os dados incluem: ***ID, Price; Fee; Area (square feets); Bedrooms; Bathrooms; City name*** e outros atributos, num total de 22 atributos. A variável alvo é o preço do imóvel (**Price**) que é uma variável contínua (problema de regressão). A descrição dos dados e o arquivo de dados está disponível no Google Classroom. Mais informações sobre o dataset estão disponíveis no arquivo **apartments.txt.**

A preparação de dados deve verificar a existência de dados faltantes ou inconsistentes e definir como trata-los (se houver). Deve-se também avaliar a relevância de cada campo para a realização de uma melhor estimativa do preço, inclusive com análise exploratória de dados usando gráficos.

Cada grupo pode propor ao professor uma **Base de Dados Alternativa** para utilizar no projeto, desde que tenha complexidade igual ou superior a base de dados aqui proposta e trata-se de um problema de estimativa de valor numérico (regressão), tratável como um problema de aprendizado supervisionado. A base de dados proposta pode ser substituir a base descrita acima, apenas com a explícita concordância do professor.

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. Faça uma análise exploratória dos dados. Prepare os dados para serem apresentados aos modelos de estimativa (regressores).Os dados podem ter atributos faltantes ou com imprecisões em seu valor (ruído). O dataset está dividido em dois conjuntos: **apartments_for_rent_train.csv e apartments_for_rent_test.csv** respectivamente para treino e teste.

O ajuste de hiperparâmetros deve utilizar a técnica de Cross Validation com K-Folds (K=10) discutida em sala para realizar o ajuste de hiperparâmetros com o dataset de treino.

2. Crie três modelos usando diferentes técnicas de aprendzado de máquina para resolver o problema.

a) Modelo baseado em K-NN, Árvore de Decisão ou SVM

Crie um modelo baseado em K-NN (K Nearest Neigbhours), árvore de decisão ou SVM para fazer a estimativa da variável alvo.

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

Crie um modelo baseado em Redes Neurais do tipo MLP (MultiLayer Perceptron) para fazer a estimativa da variável alvo.

c) Modelo baseado em Comitês (Random Forests, AdamBoost, etc)

Crie um modelo baseado em Florestas Aleatórias (Random Forests) ou outra técnica vista de comitê vista (AdaBoost, XGBoost, etc) para fazer a estimativa da variável alvo.

3. Análise Comparativa do desempenho dos modelos.

Avalie comparativamente os dois modelos, utilize medidas apropriadas de desempenho de modelos (RMSE, MAE, MAPE, R^2, etc). Discuta os resultados e qual seria o modelo mais apropriado.

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

4. Aplicação da predição do Modelo Desenvolvido

Criar um trecho de código para aplicar o modelo gerado a um arquivo de dados não conhecido a priori, mas com mesmo formato do arquivo de dados fornecido (apartments_for_rent_test.csv). Observe que o arquivo de dados deverá ser preparado pelo código para ser apresentado ao regressor, de modo similar ao feito no item 1. O nome do arquivo de dados de testes será 'apartments_for_rent_final_test.csv'

3. Material a ser Entregue e Prazo

Deve ser entregue um jupyter notebook (formato .ipynb) com descrição, comentários e código-fonte

OBS: Entregar através do Google Classroom! Não compacte o arquivos em um zip (ou qq outro formato),!

A. Notebook com descricão, comentários e código-fonte (ver detalhes abaixo)

Prazo de Entrega: 1/julho/2024;

4. Apresentação do Trabalho:

Cada grupo dever apresentar o trabalho, análise, metodologia, conclusões e justificando suas conclusões e respondendo eventuais questões sobre os procedimentos adotados e conclusões.

Data da Apresentação:3/julho/2024

Estrutura do Notebook

OBS:(arquivo em jupyter Notebook) Intercarlar células(tags) de texto e código fonte observando estrutura indicada abaixo. Comentar o código nas célula de código

Título: Lab. Exame - CMC-13

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.

5. Aplicação da predição do Modelo Desenvolvido

Trecho de código para uso do modelo desenvolvido aplicado a um arquivo de dado do mesmo formato do arquivo de dados fornecido (apartments_for_rent_train.csv).O nome do arquivo de dados de testes será 'apartments_for_rent_final_test.csv'

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

Bom Trabalho!

Prof. Paulo André Castro