Phoebus - Estágio

Outubro 2019

Prova de Estágio

0.1 Dados Utilizados

Os problemas aqui apresentados são baseados nos seguintes conjuntos de dados:

1. *Titanic Machine Learning from Disaster*
   - `https://www.kaggle.com/c/titanic/data`
2. *Instacart Market Basket Analysis*
   - `https://www.kaggle.com/c/instacart-market-basket-analysis/data`

0.2 Especificação

Os problemas podem ser resolvidos nas linguagens:

1. Python
2. R

Para as visualizações, além das linguagens, fica aberta a utilização das ferramentas:

1. Tableau
2. Power BI

1 Problemas

1.1 Introdução - Programação

1. Implemente um programa que faça a ordenação do *array* [5, 9, 54, 12, 7, 5, 6, 1] utilizando algoritmo de *BubbleSort*.

2. Implemente um programa que faça a inserção e ordenação do *array* [23, 11, 7, 21, 59, 83, 38] utilizando algoritmo de *BinaryTree*.

3. Escreva uma classe Intervalo que representa um intervalo de coordenadas inteiras e identificador também inteiro. Os objetos desta classe devem ser criados passando-se três parâmetros nominais id, xmin e xmax com valores inteiros. A classe deve possuir também um método elo que recebe um outro intervalo desta classe e calcula o valor do elo entre os dois, definido como segue:

   - Se os intervalos tem interseção não vazia, o elo é o tamanho desta interseção;
   - Se os intervalos a e b não se encontram, o elo é o negativo da distância entre eles, definida como a menor distância entre um ponto de a e um ponto de b.

```
    Exemplos:
>>> a = Intervalo(id=1, xmin=2, xmax=3)
>>> b = Intervalo(id=2, xmin=3, xmax=4)
>>> a.elo(b)
    0
>>> b.elo(Intervalo(id=4, xmin=8, xmax=15))
    -4
>>> a.elo(Intervalo(id=8, xmin=2, xmax=4))
    1
```

1.2 Introdução Teórica - Ciência de Dados

1. Qual a diferença entre aprendizagem supervisionada e não supervisionada? Explique um pouco de cada.

2. O que você entende por validação cruzada?

3. O que é uma regressão linear?

4. O que são *outliers*? Apresente, ao menos, uma forma de lidar.

1.3 Titanic

1. Faça uma análise inicial das variáveis (*exploratory data analysis*) apresentadas em: `https://www.kaggle.com/c/titanic/download/train.csv`

2. Tente prever a variável ***Survival***, quais passageiros estavam mais propensos à sobrevivência, baseado nos atributos estudados na questão anterior.

1.4 Instacart

1. Utilizando o conjunto entre os dados de:

   - **Vendas**: `https://www.kaggle.com/c/instacart-market-basket-analysis/download/order_products__train.csv`
   - **Produtos**: `https://www.kaggle.com/c/instacart-market-basket-analysis/download/products.csv`

   Apresente visualizações demonstrando:

   (a) Quais produtos são os mais vendidos (product_id, product_name)?

   (b) Quais produtos, dentre os que são primeiro introduzidos no carrinho (add_to_cart_order), são os mais vendidos?

   (c) Qual a média de produtos(product_id) por compra(order_id)?