COS868 - Probabilidade e Estatística para Aprendizado de Máquina

Segundo Semestre de 2023 - Professora: Rosa Maria Meri Leão

Projeto do Curso

1 Dataset

O objetivo deste trabalho é analisar um conjunto de dados aplicando a teoria aprendida em classe. É muito importante que seja realizada uma análise crítica dos resultados encontrados.

O projeto será baseado em dados reais fornecidos por um provedor de Internet de médio porte. Os dados representam a taxa de dados enviados em bps (taxa de upload) e a taxa de dados recebidos em bps (taxa de download) de/por um dispositivo na casa de um usuário do provedor. Dois tipos de dispositivos devem ser analisados: Smart-TV e Chromecast.

Dois arquivos contendo os dados serão disponibilizados: (1) dataset_chromecast.csv e (2) dataset_smart-tv.csv. Cada arquivo possui os seguintes campos: **device_id, date_hour, bytes_up, bytes_down**. O campo **device_id** é o identificador único do dispositivo que está na casa de um usuário do provedor. O campo **date_hour** é a data e a hora em que a coleta de dados foi realizada com a granularidade de 1 minuto. Os campos **bytes_up** e **bytes_down** contém, respectivamente, as taxas em bps dos dados enviados (taxa de upload) e dos dados recebidos (taxa de download) pelo dispositivo em um minuto.

Atenção reescalonar dados para log 10: como os valores das taxas de upload e de download variam diversas ordens de grandeza, para calcular as estatísticas é necessário reescalonar as taxas para log na base 10. Por exemplo, se o campo **bytes_up** é igual a 1000, o valor que você deve usar para fazer as análises é 3.

2 Estatísticas gerais

O objetivo desse estudo é avaliar os dados sem considerar o horário em que foram gerados, ou seja, você deve considerar todos os dados de cada um dos arquivos para obter as estatísticas descritas a seguir. Para cada tipo de dispositivo, Smart-TV e Chromecast, calcular: Histograma, Função Distribuição Empírica, Box Plot, Média, Variância e Desvio Padrão, para a taxa de upload e taxa de download. Com relação aos box plots, você deve gerar **um único gráfico** com os quatro box plots (taxa de upload (Smart-TV e Chromecast) e taxa de download (Smart-TV e Chromecast) de forma a poder compará-los.

Lembre-se que o tamanho do *bin* deve ser estimado de forma que o histograma represente de forma adequada os dados estudados. Use o método de Sturges apresentado em aula para estimar o tamanho do *bin*.

Comente o que você observou a partir dos gráficos e sobre as diferenças e/ou similaridades entre os dois tipos de dispositivos. **Mais importante que o cálculo das estatísticas, é a interpretação dos resultados obtidos.**

3 Estatísticas por horário

O objetivo dessa análise é avaliar os dados considerando o horário em que foram gerados independente do dia. Você deve considerar os dados coletados em cada hora para cada tipo de dispositivo para obter as estatísticas descritas a seguir. Para cada tipo de dispositivo, Smart-TV e Chromecast, para cada hora calcular: Box Plot, Média, Variância e Desvio Padrão, para a taxa de upload e taxa de download.

Com relação aos box plots, você deve gerar **4 gráficos**: taxa de upload (Smart-TV e Chromecast) e taxa de download (Smart-TV e Chromecast) de forma a poder compará-los. Cada gráfico deve conter 24 box plots (um para cada hora).

Para a média, variância e desvio padrão, você deve fazer **4 gráficos**, representando no eixo X a hora e no eixo Y os valores das três estatísticas para cada taxa coletada, para cada tipo de dispositivo.

Faça uma análise dos resultados obtidos. Comente sobre as diferenças ou similaridades entre os dois tipos de dispositivos. O que você pode concluir a respeito das estatísticas obtidas por horário?

4 Caracterizando os horários com maior valor de tráfego

Neste item o objetivo é analisar os dois horários com maior valor da média de cada taxa coletada para cada tipo de dispositivo. O **Passo 1** é escolher, a partir dos gráficos da seção 3, o horário com maior valor de média, para a taxa de upload e taxa de download, para cada tipo de dispositivo: Smart-TV e Chromecast.

Você terá 4 datasets, cada um contendo os dados com as seguintes características:

- Dataset 1: Horário com a maior média da taxa de upload em uma hora, Smart-TV
- Dataset 2: Horário com a maior média da taxa de download em uma hora, Smart-TV
- Dataset 3: Horário com a maior média da taxa de upload em uma hora, Chromecast
- Dataset 4: Horário com a maior média da taxa de download em uma hora, Chromecast

No **Passo 2**, faça um histograma para cada um dos 4 datasets. Lembre-se que você deve escolher o tamanho do *bin* usando o método de Sturges.

No **Passo 3** calcule o maximum likelihood estimator (MLE) para estimar os parâmetros das seguintes distribuições: Gaussiana e Gamma, para cada um dos 4 datasets. Explique como você calculou o MLE.

No **Passo 4** você deve fazer um gráfico para cada um dos 4 datasets contendo 3 curvas: o histograma, a função densidade Gaussiana com os parâmetros obtidos com o MLE e a função densidade Gamma com os parâmetros obtidos com o MLE. Observando os gráficos você deve comentar se existe ou não uma variável aleatória da literatura que possivelmente possa ser usada para representar os dados de cada um dos 4 datasets.

O **Passo 5** consiste em fazer o gráfico *Probability Plot* comparando os dados de cada dataset com as distribuições parametrizadas. No total são 8 gráficos, comparando os dados reais dos 4 datasets com cada uma das duas distribuições parametrizadas.

O **Passo 6** consiste em fazer o gráfico *QQ Plot* comparando os dados do dataset 1 com aqueles do dataset 3, e os dados do dataset 2 com aqueles do dataset 4. Serão dois gráficos, um para cada tipo de taxa: download e upload.

Como o tamanho dos datasets que tem ser comparados através do *QQ Plot* são diferentes. Segue uma sugestão do procedimento para compará-los.

Primeiramente começamos ordenando ambos os conjuntos de dados. Esse é um procedimento padrão na criação de um *QQ plot*.

Para o primeiro conjunto de dados, com menor número de amostras, calculamos os quantis diretamente.

Para encontrar os pontos no segundo conjunto de dados (que possui o maior número de amostras) que correspondam às posições dos quantis do primeiro conjunto de dados, usamos a interpolação linear. A equação básica da interpolação linear entre dois pontos é:

$$y = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)} \tag{1}$$

Onde:

- x é o valor do quantil para o qual queremos encontrar um valor interpolado (neste caso, são todos os quantis do primeiro conjunto de dados).
- x_1 e x_2 são valores conhecidos de quantis do segundo conjunto de dados.
- y_1 e y_2 são valores do segundo conjunto de dados correspondentes aos quantis x_1 e x_2, respectivamente.
- y é o valor interpolado que estamos calculando.

Dessa forma estamos alinhando os quantis do conjunto de dados menor com quantis equivalentes no conjunto de dados maior através da interpolação. Isso nos permite comparar conjuntos de dados de tamanhos diferentes em um QQ plot, *redimensionando* o conjunto de dados maior para corresponder à distribuição de quantis do menor.

A partir dos resultados dessa seção você deve analisar as seguintes questões:

1. Quais foram os horários escolhidos para cada dataset?
2. O que você pôde observar a partir dos histogramas dos datasets?
3. Comente sobre as diferenças e/ou similaridades entre os datasets 1, 2, 3 e 4.
4. É possível caracterizar os datasets acima por uma variável aleatória da literatura?
5. Se a resposta for não, qual o motivo?
6. O que você pôde observar a partir dos gráficos *QQ Plot* e *Probability Plot*?

5 Análise da correlação entre as taxas de upload e download para os horários com o maior valor de tráfego

.

Neste item o objetivo é analisar se existe alguma correlação entre a taxa de upload e a taxa de download de um mesmo dispositivo. Você deve calcular o coeficiente de correlação amostral e fazer o gráfico scatter plot comparando as taxas dos seguintes datasets: **dataset 1 e dataset 2**, **dataset 3 e dataset 4**. Você deve gerar dois gráficos scatter plot, um para cada dispositivo, comparando as suas taxas.

Você deve analisar os resultados e indicar se existe alguma correlação entre as taxas de download e upload dos dispositivos. O que você pôde concluir a partir dos resultados?

6 Comparação dos dados gerados pelos dispositivos Smart-TV e Chromecast

O objetivo deste estudo é avaliar se os dois dispositivos que são usados prioritariamente para assistir vídeo, possuem distribuição de probabilidade das taxas de upload e download semelhante nos horários de maior tráfego. Você deve usar a estatística G do teste Chi-Square for goodness of fit (apresentado em classe) para fazer essa análise. Esse teste também é chamado de G-test (https://en.wikipedia.org/wiki/G-test).

Você deve realizar o G-test comparando as taxas dos seguintes datasets: **dataset 1 e dataset 3**, **dataset 2 e dataset 4**.

Note que para realizar o G-test entre cada par de datasets, o número de bins e os valores dos bins dos histogramas dos datasets a serem comparados devem ser os mesmos. Portanto se os bins dos histogramas que você obteve na seção 4 para um determinado par de datasets for diferente, escolha o número de bins e os valores dos bins de um deles e use como referência para aquele par de datasets.

7 Relatório

Você deve fazer um relatório contendo todos os resultados que você obteve e explicando como você os obteve. É importante comentar cada um dos resultados e explicar se o resultado que você obteve poderá auxiliar o provedor de serviço de Internet a entender os dados que passam pela sua rede. A avaliação do projeto será feita com base na qualidade do relatório.

Você deve fazer upload do seu relatório (arquivo pdf) na plataforma do Google Classroom na atividade Projeto do Curso.

No relatório deve estar indicado um link para o código que você usou para obter os resultados do trabalho.