Processamento de Streams - Módulo II
Fase II - Final submission

João Funenga nº 61635, Catarina Afonso nº 62206

3 de junho de 2022

Abstract

Neste *paper*, será descrito o dataset com que se vai trabalhar neste *assignment*, bem como enunciar algumas questões que se quer analisar numa fase posterior sobre os dados que se tem.

Introdução

Neste projeto trabalhar-se-á sobre um *dataset* que engloba os dados dos tempos de espera nos hospitais na região de Lisboa. Com este conjunto de dados pretende-se resolver algumas perguntas que o grupo irá enunciar na secção seguinte recorrendo a técnicas de Machine Learning (ML) em *streaming*. Relativamente aos dados com que se vai trabalhar,o dataset será limitado para as observações feitas apenas no Santa Maria por ser o que contém mais observações (395858) e por ter menos tipos de urgência (apenas urgência geral e pediátrica).

Dataset

A tabela 1 representa o *schema* dos dados. O *dataset* é definido por 1603384 instâncias com 8 *features*, sendo estas representadas abaixo:

Table 1: Descrição do dataset

AcquisitionTime	Hospital	UrgencyType	Service
TimeStamp	int	string	string
EmergencyState	Waiting_Time	PeopleWaiting	H_name
string	int	int	string

Desafios

O desafio focar-se-á essencialmente no desenvolvimento do modelo preditivo do tempo de espera sabendo a cor da pulseira e a altura do dia. O tempo de espera dos hospitais representa tudo o que advém de algo anormal. Isto é, no Inverno em que é recorrente haver crises de gripes, os hospitais já têm mais gente para trabalhar logo os tempos de espera já terão isso em conta. Tudo o que foge destes padrões esperados é que entram para o tempo de espera.

Situação Atual em Portugal

Atualmente, de acordo com o ministério de saúde [1], a estimação de tempo de espera com uma determinada pulseira é recorrendo à mediana de todos os tempos de espera para cada um dos diferentes níveis de manchester, mas tendo em conta a dimensão da instituição. Para cada um dos diferentes níveis, os intervalos de tempo considerados para o cálculo da mediana dos dados variam. Embora estes cálculos sejam feitos com um intervalo de 5 minutos, ou seja, em contexto de *streams* isto corresponderia a uma janela deslizante de tamanho correspondente ao definido para cada nível de urgência e cujo *stride* seria de 5 minutos. Esta forma de estimação, parece ser algo rudimentar por apenas ser calculada uma mediana e não com um algoritmo de ML mais avançado. No entanto, perceber-se-á se é de facto a melhor escolha ou não no que toca a este problema.

State of the art

Neste momento, os modelos *State of the art* para fazerem este tipo de previsões baseiam-se em vários atributos e no fundo são um modelo de regressão cujo valor a estimar é o tempo de espera. No *paper* [2] é dito que esta metodologia obteve resultados bastante melhores que apenas uma média deslizante sobre uma janela (análogo ao que Portugal faz com a mediana). Com estas melhorias, é possível dar melhores estimativas para o bem estar dos pacientes. Este modelo não procura reduzir os tempos de espera porque esses são sempre devido a acontecimentos aleatórios e esporádicos mas sim de os prever.

Estudo Experimental

Nesta fase, será efetuada a revisão da literatura científica, desenvolvendo o contexto e a base conceptual apresentando simultaneamente a metodologia aplicada.

Exploratory Data Analysis

A Exploratory Data Analysis (EDA) consiste num processo importante de exploração de dados que possibilita descobrir informações relevantes sobre os mesmos. Permite identificar padrões, fazer testes de hipóteses e verificar os pressupostos através de estatística descritiva e de representações gráfica.

Como referido anteriormente, os dados que serão analisados são referentes apenas ao Hospital de Santa Maria, para ter um conhecimento geral sobre o *dataset*.

Primeiramente, verificou-se a que classe pertence cada variável nos dados e se os mesmos continham valores omissos (NaN). Através do método `.info()`, sabe-se que é necessário efetuar um *cast* em determinadas variáveis para um *data type* específico e, também, se certificou que não foi encontrado nenhum valor em falta:

```
RangeIndex: 395858 entries, 0 to 395857
Data columns (total 8 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Acquisition_Time  395858 non-null  object
 1   Hospital          395858 non-null  int64
 2   Urgency_Type      395858 non-null  object
 3   Service           395858 non-null  object
 4   Emergency_Stage   395858 non-null  int64
 5   Waiting_Time      395858 non-null  int64
 6   People_Waiting    395858 non-null  int64
 7   H_Name            395858 non-null  object
```

Figure 1: Sumário conciso sobre o *dataset*

Para este caso, efetuou-se uma mudança nas variáveis do tipo `object` e na variável 'Emergency_Stage' para o *type* que se encontra indicado na Tabela 1.

Seguidamente, procedeu-se ao cálculo das medidas descritivas características da amostra. Foi obtido o seguinte resumo estatístico para as duas variáveis numéricas:

	Waiting_Time	People_Waiting
count	395858.000000	395858.000000
mean	69.607175	4.148808
std	75.216639	6.291929
min	0.000000	0.000000
25%	24.000000	0.000000
50%	45.000000	2.000000
75%	85.000000	5.000000
max	599.000000	55.000000

Figure 2: Sumário estatístico sobre as variáveis numéricas

Observando os valores dados pela média, é de notar que os valores das duas variáveis são muito diferentes entre si. Pode-se afirmar que o tempo de espera médio é de aproximadamente 70 minutos e o número médio de pessoas em espera para o atendimento antes da triagem é de aproximadamente 4 pessoas. Estes valores parecem razoáveis uma vez que o cálculo da média funciona bem em conjuntos de dados muito grandes.
Quanto ao desvio padrão, pode-se verificar que o valor do desvio padrão da variável 'Waiting_Time' é relativamente alto, isto indica que os pontos dos dados se encontram dispersos em relação à média, ou seja, existe uma variação no período em que os pacientes estão à espera para receber o tratamento. Outro aspeto relevante do *dataset*, o máximo de tempo que um paciente teve esperar foi de 599 minutos, o que corresponde a aproximadamente a 10 horas.

Posto isto, visualizou-se a distribuição de algumas variáveis que fossem relevantes para a exploração do tema. Primeiramente, recorreu-se a um *bar chart* para visualizar a distribuição da variável categórica 'Emergency_Stage'. O comprimento de cada barra é proporcional à frequência da categoria correspondente:

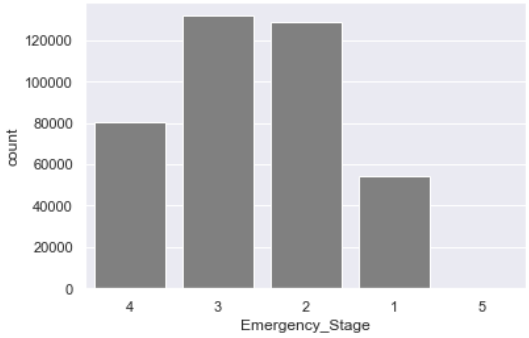

Figure 3: Representação gráfica da distribuição da variável 'Emergency_Stage'

Pela Figura 3, pode-se imediatamente observar que existe uma frequência muito baixa ao estado de emergência '5', isto significa, que os casos como risco iminente de morte tendem a acontecer raramente (não visível por causa da escala). Caso o desafio se tratasse de um problema de classificação, como a classe '5' se encontra em minoria, seria ótimo recorrer, a nível dos dados, a técnica de Random Over Sampling (ROS) para tratar de *imbalaced data*, para que as classes fiquem balanceadas embora tenhamos de ter cuidado pois isto pode levar a overfitting do modelo aos dados.

Seguidamente, produziu-se os seguintes histogramas nas variáveis numéricas:

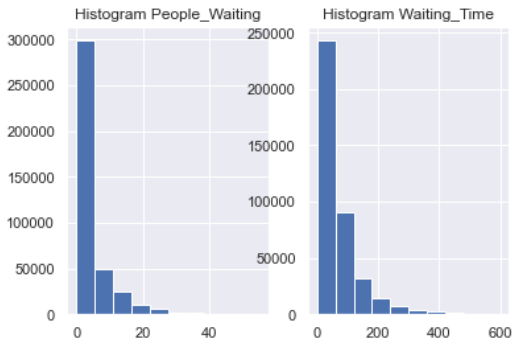

Figure 4: Representação gráfica das distribuições 'People_Waiting' e 'Waiting_Time' respetivamente

Por fim, para perceber a relação entre as duas variáveis numéricas, neste caso, 'Waiting_Time' e 'People_Waiting', produziu-se o seguinte *scatterplot*:

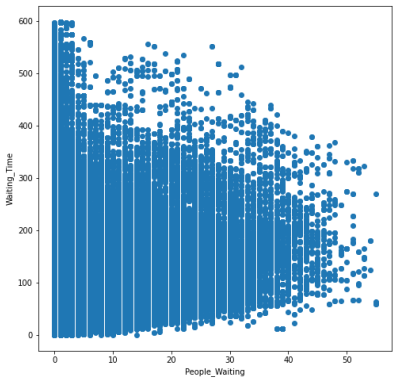

Figure 5: Representação gráfica da relação entre 'Waiting_Time' e 'People_Waiting'

Observando a figura 4 que relaciona o número de pessoas à espera com os tempos de espera, esta aparenta apresentar uma inconsistência, por exemplo, a existência de um tempo de espera associado à observação com um número de pessoas em espera igual a 0. Este tempo de espera positivo mesmo quando não há pessoas há espera ocorre porque este valor é calculado através da médias das pessoas à espera nas 2 últimas horas. Relativamente ao resto, o tempo à espera parece diminuir com o aumento do número de pessoas à espera o que parece contra-producente.

Análise do Concept Drift

Depois de efetuada a EDA, para se ter um entendimento geral dos dados que se tem, sem realizar qualquer tipo de modelação estatística, analisou-se outro aspeto importante antes de se passar para a fase preditiva, a análise do Concept Drift. Este é um problema muito recorrente e significante em *streams* porque a incerteza e o desconhecimento dos dados que se tem e das respetivas distribuições e variações das mesmas ao longo do tempo irão influenciar diretamente algumas decisões que se tem de fazer relativamente ao desenvolvimento do modelo. Esta análise envolve saber quando ocorre, de que forma ocorre e onde ocorre.

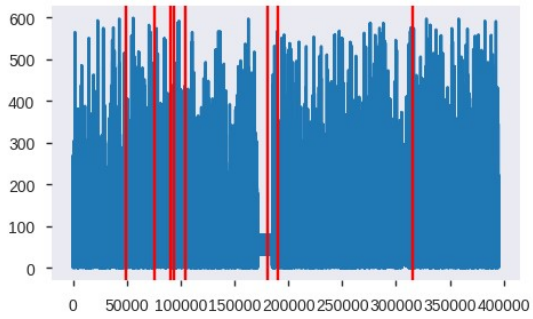

Figure 6: Concept Drift - Waiting Time

Waiting Time

Analisou-se agora o Concept Drift relativamente ao tempo de espera em minutos pelas pessoas. Como se pode ver pelo gráfico acima, foram detetados alguns momentos em que existiu *drif*t. Recorreu-se ao ADWIN para detetar as alterações. Esta ferramenta não necessita que se defina o tamanho das janelas apriori, apenas precisa de um valor representante de uma janela suficientemente grande para fazer a análise. Este examinará todos os possíveis valores de corte dessa janela e calcula valores apropriados para as janelas mais pequenas tendo em conta os dados históricos caso detete que haja diferenças significativas (entre os dados históricos e os atuais), o que é representante de haver *concept drift*. De uma forma mais formal, este usa a seguinte estatística de teste $\theta_{ADWIN} = |\mu_{Hist} - \mu_{New}|$ O corte ótimo é encontrado quando esta diferença excede um limiar pré-definido (no parâmetro delta). Como se pode observar pelo gráfico, existe algum *delay* entre o momento em que ocorre o *drift* e o momento em que é detetado. Isto é facilmente visível no meio do gráfico em que as duas deteções estão ligeiramente desfasadas do gráfico que sobrepõem. Este *concept drift* refere-se a um menor tempo de espera comparativamente aos restantes dias. Relativamente ao tipo de *concept drift* presente neste caso, parece ser do tipo "*Sudden drift*" pela facto da distribuição mudar drasticamente a meio comparativamente ao resto dos dados.

People Waiting

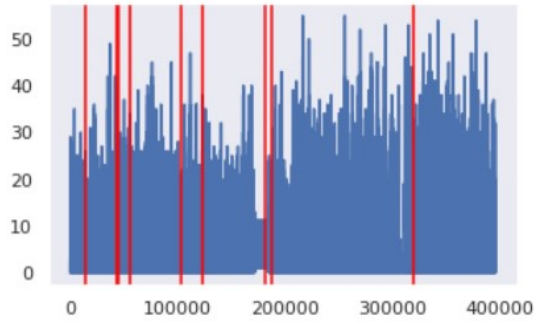

Figure 7: Concept Drift - People Waiting

Depois de se perceber que existe uma variação na distribuição dos dados numa altura no tempo relativamente ao tempo de espera, ver-se-á agora se tal acontece para as pessoas à espera. Como se pode verificar pelo gráfico acima, o tipo de *concept drift* detetado é muito semelhante ao do que ocorre para o tempo de espera. Isto deve-se a numa determinada altura ter-se muito menos pessoas à espera que nos restantes dias. Por ambos os *concept drifts* parecerem ocorrer da mesma forma, possivelmente estes estarão relacionados, o que faz sentido no que toca à interpretação visto que, caso se tenha menos pessoas à espera, pelo "Waiting_Time" ser calculado usando a média de pessoas das últimas 2 horas, é de esperar que esta redução drástica se reflita nesta variável. Deste modo, o tipo de *drift* também aparenta ser "*Sudden Drift*".

Modelos propostos

Aspetos Gerais dos modelos

Para a construção dos modelos preditivos, recorreu-se a algoritmos que utilizam Aprendizagem Supervisionada, ou seja, fornece-se ao algoritmo um conjunto de dados como *input* com as suas respetivas saídas desejadas como *output*. O objetivo é estimar o mapeamento entre os exemplos de entrada e os exemplos de saída. Como se pretende prever um valor contínuo, neste caso, o tempo de espera, então enfrentou-se o problema com algoritmos de Regressão. A regressão é um técnica para modelar e analisar dados que são compostos por uma variável dependente (a variável de resposta) e uma ou mais variáveis independentes (as variáveis de entrada).

A partir disto, selecionou-se as variáveis mais relevantes para a construção de cada modelo, que serão descritas nos próximos subcapítulos, e criou-se uma nova variável a partir dos dados disponíveis, de forma a melhorar o desempenho do modelo. Neste caso, criou-se uma *lag feature* denominada por 'lag_Waiting_Time_30_mn'. Esta variável consiste em usar o tempo médio de espera nos últimos 30 minutos. Outro conceito comum entre os modelos, é o "Pipelining". Isto representa uma sequência de operações que definem o *dataflow*. Para todos os 4 modelos apresentados de seguida usou-se inicialmente o Standard Scaler para escalar os dados devido às diferentes escalas mas, como falado nas teóricas, não seria necessário aplicar esta operação aos modelos baseados em árvores. Relativamente à métrica utilizada para avaliar a qualidade do modelo, usámos o MAE por ser o indicado em contexto de problemas de regressão e estar nas unidades dos dados (por exemplo, como estamos a prever o tempo de espera em minutos, o MAE virá em minutos).

Primeiro Modelo

O primeiro modelo que se construiu foi um modelo de Regressão Linear [3]. Este modelo é o mais simples de aprendizagem supervisionada. A variável y é calculada a partir de uma combinação linear de um conjunto de variáveis X. A notação vetorial generalizada do modelo de regressão linear é dada da seguinte forma:

$$\hat{y} = \mathrm{w}^T \mathrm{x} = \mathrm{x}^T \mathrm{w}$$

Onde $\mathrm{x} = \begin{bmatrix} 1 \\ x_1 \\ \dots \\ x_d \end{bmatrix}$ e $\mathrm{w} = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_d \end{bmatrix}$.

Para o desenvolvimento do modelo, selecionou-se as variáveis numéricas 'People_Waiting', "lag_Waiting_Time_30_mn" e a variável categórica "Emergency_Stage". Antes de treinar o modelo, como as unidades em que as variáveis se encontram são diferentes, e, para além disso, verificou-se através do EDA que a distribuição das features são distintas, bem como se verificou a existência de *outliers*, standardizou-se os atributos, ou seja, escalou-se os dados de forma a que todas as dimensões ficassem com média zero e variância unitária (μ=0, σ^2= 1). Este pré-processamento é um método robusto e é ideal para que o *optimizer* possa convergir. No decorrer do treino, escolheu-se o *Stochastic Gradient Descent* como *optimizer* e a medida utilizada para avaliar a performance foi o MAE.

Segundo Modelo

Para o segundo modelo seguiu-se uma abordagem um pouco diferente ao usar um regressor diferente ao invés do simples linear. Como se viu nas aulas, as árvores poderiam ser uma boa alternativa quando se tem dados que não tenham sofrido grande pré-processamento e que podem estar um pouco *messy* (sem serem normalizados, ou por ter valores em falta). Deste modo, a árvore utilizada neste caso foi a *Hoeffding tree* [4]. Esta apenas vê um exemplo de cada vez e, caso tenha dados suficientes, consegue generalizar bem para dados nunca antes vistos que aparecerão na stream, daí o nome igual ao da desigualdade falada no início deste módulo. Tendo esta ideia em mente, aplicou-se o modelo. Para este, decidiu-se que as features deveriam ser a "Emergency_Stage", a "People_Waiting" e o "Lag" utilizadas, também no primeiro modelo. Inicialmente utilizou-se também a timestamp, mas rapidamente percebeu-se que esta não fazia sentido por si só ser uma *feature* por ser algo que nunca se repete. Importante de notar que, relativamente à versão usada do regressor da *Hoeffding Tree*, esta foi a versão *adaptive* que implementa técnicas extra para lidar com o *drift*, que como se viu na secção anterior, existe nos dados.

Terceiro Modelo

O terceiro modelo foi feito para tentar tirar partido do atributo que não foi usado no segundo modelo, a timestamp. Pelo que foi explicado acima, uma timestamp é algo que não faz sentido ser usado como atributo por nunca se repetir mas, caso se use o dia e a hora da timestamp, estes já podem ser utilizados como atributos por se repetirem e poderem, possivelmente ajudar o modelo caso existam por exemplo mais casos aos fins de semana ou à noite do que noutras alturas. Deste modo, os atributos usados foram os mesmos que no segundo modelo ('People_Waiting', 'Emergency_Stage', 'Lag_30Mins') bem como este adicional, relativo à hora e ao dia da semana. Para o regressor a usar, tnetou-se utilizar agora ao invés da árvore, uma regressão linear igual ao primeiro com o Stochastic Gradient Descent como *optimizer*.

Quarto Modelo

Por fim, construi-se um quarto modelo recorrendo a métodos de ensemble. Para a construção do mesmo, utilizou-se os mesmos atributos que o modelo anterior (incluindo as *features* adicionadas) e aplicou-se o mesmo pré-processamento nos dados (*StandardScaler*). Posto isto, treinou-se de forma independente e paralela 4 modelos, sendo que os 3 primeiros modelos correspondem a um modelo de regressão linear, experimentando com diferentes *optimizers*: Stochastic Gradient Descent, RMSProp e o Adam, respetivamente [5]. A utilização de diferentes *optimizers* é importante, dado que estes mudam a forma como o modelo converge para a solução. A escolha do último modelo, relaciona-se com

a análise feita anteriormente sobre o *concept drift*. Tendo em conta que se detetou a existência de *drift* nos dados, o algoritmo Hoeffding Apdative Drift pareceu ser a escolha mais apropriada. A utilização de árvores de decisão, pode-se tornar instável quando existem pequenas variações nos dados. Ora, tal como se verificou anteriormente, percebeu-se que existe uma variação na distribuição nos dados, no entanto, este facto não terá implicações uma vez que se utilizou este modelo num *ensemble*.

Comparação dos modelos

Nesta fase, comparou-se a performance dos modelos obtidos. Por este ser um problema de regressão, usou-se o MAE como métrica para avaliar a qualidade dos modelos. Este denomina-se por "Mean Absolute Error" e representa o erro médio entre o valor previsto e o verdadeiro. Deste modo, é calculado somando os erros absolutos e dividido pelo tamanho da amostra e, assim, está nas mesmas unidades do que está a ser medido.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}(|f(x_i) - y_i|) \quad (1)$$

Neste caso, por se estar a prever o tempo de espera em minutos, o valor do MAE será também o erro em minutos.

Table 2: Comparação dos MAE's

Modelo	MAE (minutos)
1	16.23
2	15.44
3	16.99
4	69.61

Para o primeiro modelo temos um MAE de 16.23 minutos, cujo valor está entre o MAE do segundo e do terceiro modelo. Deste modo, uma simples regressão linear teve mais minutos de erro do que com a *Hoeffding tree*. Isto demonstra que uma árvore pode ser um melhor modelo para quando se tem muitos dados como é o caso.

Relativamente ao segundo modelo, este teve um MAE de 15.44 minutos enquanto que o terceiro teve um MAE de 16.99 minutos. Em teoria, as *features* do terceiro modelo são iguais às do segundo e ainda têm atributos adicionais relativamente à hora e ao dia da semana, mas acaba por ter um erro mais elevado. Isto demonstra que a Hoeffding Tree acaba por ser melhor que uma regressão linear neste tipo de problemas devido à não linearidade dos dados.

Analisando agora os resultados produzidos pelo MAE do quarto modelo, pode-se afirmar que não são satisfatórios (MAE = 69.61 min) . O modelo é relativamente instável comparativamente aos restantes modelos. Podem existir diferentes explicações para este resultado pouco favorável, uma delas pode ter a haver com o facto da a árvore ser um *strong learner*, e os métodos de *ensemble* utilizarem apenas *weak learners*. No entanto, experimentou-se também com apenas modelos de regressão linear, e os resultados foram iguais. Uma possível melhoria, seria utilizar outras combinações de modelos no *ensemble*.

Conclusão

Com este trabalho teve-se a oportunidade de experimentar vários modelos para prever o tempo de espera dos pacientes nos hospitais. Utilizou-se *Linear Regressors*, *Hoeffding trees* e *Ensemble methods* com mais do que um modelo a ser treinado em paralelo. Quanto à análise dos resultados, conseguiu-se observar que os resultados para os modelos foram todos muito semelhantes mas que para o *ensemble*, os resultados foram bastante piores. Teve-se também uma primeira abordagem ao conceito de *Concept Drift* e aos problemas que este pode trazer ao construir modelos para prever algo num contexto de *streams* levando, assim, ao uso de métodos *Adaptative* para tentar ultrapassar este problema (ao terem integrados um detetor de *drift* no próprio modelo).

References

Serviço Nacional de Saúde (2019, 15 de Janeiro). *Informação do cálculo de tempos de espera nos hospitais*
Acedido a 06/05/2022, em http://tempos.min-saude.pt/attachments/Informacao_calculo_tempos_2019.pdf

Pak A., Gannon B. Staib A *Predicting waiting time to treatment for emergency department patients*
Acedido a 06/05/2022, em https://www.sciencedirect.com/science/article/pii/S1386505620305219

RiverML *LinearRegression*
Acedido a 29/05/2022, em https://riverml.xyz/0.11.0/api/linear-model/LinearRegression/

RiverML *HoeffdingAdaptiveTreeRegressor*
Acedido a 29/05/2022, em https://riverml.xyz/0.11.0/api/tree/HoeffdingAdaptiveTreeRegressor/

RiverML *EWARegressor*
Acedido a 29/05/2022, em https://riverml.xyz/0.11.0/api/ensemble/EWARegressor/