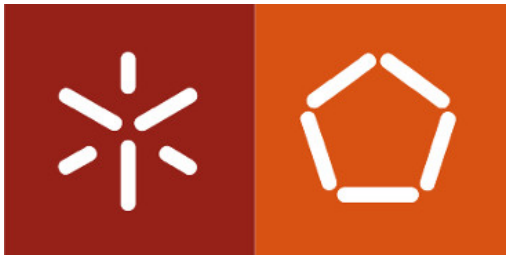

Universidade do Minho

Mestrado em Engenharia Informática

Previsão do número de infetados/mortes de doenças pulmonares

Diogo Tavares, pg42826

Hugo Nogueira, A81898

Luís Abreu, A82888

Aprendizagem Automática 2

4º Ano, 2º Semestre

Departamento de Informática

Junho 2021

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Identificação do Projeto e Objetivos

No âmbito da Unidade Curricular de Aprendizagem Automática 2 foi-nos proposta a realização de modelos de *Deep Learning* capaz de trabalhar séries temporais.

O principal objetivo passa pela criação de modelos de previsão do número de óbitos causados por uma determinada doença, e assim auxiliar uma melhor resposta no combate à mesma. Atualmente a pandemia COVID tem sobrecarregado os sistemas de saúde consumindo a maioria dos seus recursos. Neste trabalho, pretende-se o desenvolvimento de uma ferramenta, com recurso a métodos de deep learning e à linguagem Python, que seja capaz de prever o número de infectados e mortos provocados por doenças pulmonares. A previsão será feita com base nos dados fornecidos por diversas entidades de saúde sobre o número de pessoas com doenças pulmonares ao longo de um período de tempo, bem como por outras entidades dos mais diferenciados campos de atividades, como por exemplo meteorologia, aviação entre outros.

A previsão do número de mortes será realizada utilizando Séries Temporais com diferentes registos de periodos temporais, nomeadamente:

- registos diários;
- registos semanais;
- registos mensal.

1.2 Etapas de Trabalho

Conforme planeado com os orientadores do projeto, este trabalho foi dividido em três partes distintas. Uma vez que não existia um dataset criado para poder ser feita uma previsão do número de óbitos por uma doença pulmonar, uma dessas partes, a mais importante e a mais longa e onde ocorreram vários processos iterativos para melhoria desse dataset, consistiu na sua criação. Após a criação do dataset segui-se a fase de exploração e tratamento de dados, onde foram retirados features

de pouca importancia para uma previsão do número de óbitos bem como tratados os dados para melhoria do desempenho do modelo, em específico o tratamento de *missing values* Após esta fase procedeu-se à escolha de atributos relevantes para a previsão objetivo. Por fim, seguiu-se a implementação de modelos de previsão utilizando Redes Neuronais Recorrentes e Redes neuronais Convolucionais, e aqui foram experimentadas optimizações diferentes por forma a fazer um benchmark de qual o melhor melhor modelo a utilizar para este dataset.

1.3 Estrutura do relatório

Este relatório encontra-se dividido em 7 capítulos.

No capítulo 1.**Introdução** é feita uma breve abordagem ao que se pretende com a realização deste projeto, bem como quais as diferentes etapas do trabalho.

Seguidamente, no capítulo de 2.**Dataset** faz-se uma explicação detalhada dos diferentes *datasets* utilizados para a criação do *dataset* utilizado nas diferentes previsões do número de óbitos. Esta preparação ocupou uma boa parte da realização do trabalho.

No capítulo 3.**Exploração de Dados, Tratamento de dados e *Feature Selection*** , uma vez que o *dataset* final criado para a alimentação dos modelos de previsão de Séries Temporais contêm inúmeras features, e sabendo que quanto mais simples o *dataset* menos complexos são os modelos de previsão são apresentadas as várias observações realizadas para determinar quais as *features* mais pertinentes de se manter ou quais as menos interessantes e assim eliminá-las do *dataset* e obter um modelo mais simples mas sem nunca descurar a sua performance. Encontram-se também explicadas ao pormenor os vários tratamentos de dados aplicados ao dataset, em particular no seu tratamento de *missing values* e na sua transformação de *dataset* de fequência diária para frequencia semanal e mensal. Aqui é possível também encontrar a exploração de dados que nos permitiu conhecer melhor os dados do datset bem como nos auxiliou na escolha da melhor forma para realizar a *feature selection.*

No capítulo 4.**Modelos de previsão** são enumerados e explicados ao pormenor quais os modelos e suas layers constituintes utilizados neste trabalho para a previsão do número de óbitos por doença pulmonar em Portugal

No capitulo 5.**Otimização** é exposto os vários hiper-parâmetros que foram testados no intuíto de encontrar e melhorar os diferentes modelos de previsão.

No capitulo 6.**Resultados** fazemos a reflexão e análise dos resultados obtido com a aplicação dos modelos de previsão.

Por fim, em 7.**Conclusões** é feita uma breve conclusão do trabalho realizado,

onde são retiradas ilações sobre o mesmo e se o grupo concluiu os objetivos iniciais a que se propôs.

2 Dataset

O *dataset* utilizado neste projeto para a previsão do número de ótbitos em Portugal foi criado de raíz pelo grupo.

Este processo foi algo moroso devido aos requisitos que os dados do *dataset* deveriam obedecer. Como neste projeto utilizamos *datasets* com frequência de registos diários, semanais e mensais tornou a procura de dados que obedecessem a estes requisitos algo complicado, assim o grupo optou por procurar dados cuja sua frequência de registos fosse diária, para mais tarde serem transformados de forma fácil, através da soma dos mesmos, em registos semanais e mensais.

Como o objetivo deste trabalho é a previsão do número de mortes por doença respiratória em Portugal, e devido aos recentes acontecimentos da atualidade no que à pandemia COVID-19, optamos por nos focar em encontrar datasets referentes ao COVID-19, bem como outros datasets que fossem pertinentes e que contribuíssem para um incremento na qualidade dos nossos dados. No entanto apesar de partir-mos do principio que seria fácil encontrar encontrar estes dados, isso não se veio a revelar, pois apesar de haverem muitos datasets sobre a doença COVID-19, o facto desta doença ser recente faz com que outros *datsets* que no nosso entender poderiam ser relevantes sejam difíceis de encontrar.

Como base para a construção do *dataset* recorremos ao repositório no *GitHub* criado pela *Data Science for Social Good Portugal*, onde nele é possível encontrar dados referentes a Portugal, nomeadamente o número de casos confirmados de COVID-19, número de casos cofirmados por regiões do país, número de casos por faixa etária e género, o número de óbitos e número de ótitos por regiões, nível de sintomas apresentados e número de recuperados e respetivas regiões. Ainda neste repositório da *Data Science for Social Good Portugal* encontramos dados referentes ao número de testes realizados pelos laboratórios em Portugal e ao número dos diferentes tipos de teste existentes, PCR e Antigénio, ralizados.

No entanto para incrementar valor ao *dataset* que servirá para realizar previsões acrescentamos ainda dados relativos à meteorologia em Portugal. utilizando a plataforma *Visualcrossing* foi-nos possível obter dados atmosféricos em Portugal,

onde obtivemos as temperaturas minimas, máximas, médias para Portugal bem como o nível de precipitação, humidade relativa, visibilidade, estado do céu (limpo, parcialmente nublado, nublado).

Foi ainda acrescentado dados relativos à evolução diária dos acionamentos de meios de emergência médica em Portugal. Esses dados foram obtidos junto do SNS - Serviço Nacional de Saúde. Nele constavam o número de acionamento de helicóptero, viaturas, ambulâncias, motociclos de emergência médica.

Por forma a enriquecer ainda mais o *dataset* foram adicionados dados relativos a todos os países do mundo no que toca ao número de casos registados, número de novos casos, numero total de mortos nesses países, número de testes realizados, número de vacinados parcialmente e completamente, número de paciente internados em unidade de cuidados intensivos, numeros de pacientes internados nos hospitais por COVID-19, rendimentos per capita do respetivo país, número de mortes cardiovasculares, número de pessoas com diabetes, dados relativos a condições sanitária entre outros. Todos estes dados foram obtidos pela plataforma *our World Data* através do seu *GitHub*.

Com todos estes acrescentos ao *dataset* de base obtido na plataforma *Data Science for Social Good Portugal*, oiginou um *dataset* com 422 registos e 4146 features. Como é possível observar, obtivemos um *dataset* algo extenso e para fazer uma redução de *features* por forma a que os nossos modelos sejam mais simples, iremos nos capítulos seguintes fazer uma exploração dos seus valores e *features*.

Os *datasets* utilizados para a criação do *dataset* que será posteriormente tratado e explorado com o intuíto de ser utilizado na previsão do n+umero de óbitos em Portugal, encontra-se em **Datasets auxiliares**, e os processos que foram aplicados para a sua criação encontram-se na pasta **tratamento data** .

Outros *datasets* foram também pesquisados, mas devido à dificuldade em encontrar dados foram postos de parte, e o grupo optou por se focar unica e exclusivamente no *dataset* com dados sobre a pandemia COVID-19 em Portugal. Encontrando-se estes em **Datasets ignorados**.

3 Exploração de Dados, Tratamento de dados e *Feature Selection*

A exploração de dados e o seu tratamento, é das fases mais importantes em todo o processo de criação de um modelo de *machine learning*, pois permite-nos aferir a qualidade dos dados que estamos a utilizar bem como os seus pontos fortes e fracos. Nesta fase é essencial entender os tipos dos dados, definir quais as variáveis independentes e qual a classe objetivo, e é também importante retirar dos dados medidas ou características das diferentes entidades que compõem o *dataset*.

Para entender melhor os dados foram realizada uma série de visualizações gráficas para identificar-mos possíveis padrões nos dados e a possível existência de valores nulos no *datase*. Toda a exploração efetuada ao *dataset* encontra-se na pasta **exploracao data**. Nesta pasta encontram-se o ficheiro *Jupyter Notebook* **Final exploration** que demonstra passo a passo as diferentes explorações realizadas sobre o *dataset* dataframe tratado.

Para esta exploração de dados ser possível de ser realizada, foi necessário tratar os dados, eliminado os valores nulos e possíveis atributos desnecessário. Este processo foi sendo realizado de forma iterativa e á medida que mais dados de diferentes *datasets* foram sendo adicionados. Daí a pasta tratamento data conter dois ficheiros *Jupyter notebook*, o ficheiro process data que contem o tratamento de dados relativos ao *dataset* base daily_covid_19_portugal.csv e tratamento relativo às condições meteorológicas que se encontram nos *datasets* temp_portugal_XXX.csv

Este tratamento inicial originou o *dataset* daily_covid.csv, a quem posteriormente foram adicionados mais dados. Dados estes que foram tratados no ficheiro tratamento.ipynb.

O tratamento dos dados foram focados especialmente na resolução dos valores nulos no *dataset*, que foram o principal problema encontrado nos dados. Para resolver esse problema, em alguns casos demos-lhes valor 0, pois os registos encontravam-se ainda muito no inicio do periodo de pandemia, por exemplo não haviam vacinas ou ainda não tinham ocorrido óbitos. Para a falta de valores pelo

meio do dataset utilizamos o valor médio ou valor dia anterior.

Findada toda esta etapa de tratamento de dados, foi criado o *dataset* dataframe_tratado.csv, que foi então utilizado para aprofundar a nossa exploração de dados e efetuar ***feature selection*** por forma a simplificar o modelo de aprendizagem e melhorar a sua qualidade.

Esta ***feature selection*** encontra-se no ficheiro Final_exploration.ipymb, e inicialmente algumas *features* foram eliminadas baseadas na intuição, mas sempre orientada ao objetivo do problema, após isso a seleção de atributos consistiu na utilização de matriz de correlações de Spearman, pois talm como visto na exploração de dados, estes são não paramétricos. Consistiu ainda na aplicação do algoritmo SelectKBest da biblioteca *Scikitlearn*, e ta,bém dessa mesma biblioteca a utilização do algoritmo *RandomForestRegressor* aplicado ás *Lag Variables*.

Após a aplicação deste métodos decidimos manter *features* cuja sua presença fosse comum em ambos os algoritmos e que tivessem uma correlação superior a 0.5, e mantivemos também as *features* que ambos os algoritmos sugeriram mas que não se encontravam em simultânio nos mesmos.

4 Modelos de previsão

O objetivo deste trabalho centrava-se na previsão do número de óbitos em Portugal. Para alcançar tal objetivo utilizamos modelos de *deep learning* de redes neuronais recorrentes **LSTM - Long Short Term Memory** e **CNN - Concolutional Neural Network**. Estes modelos de previsão são os mais utilizados na atualidade na previsão utilizando Séries Temporais. Na figura 1 observamos a constituição dos modelos que utilizamos no nosso trabalho.

Figura 1: Modelos de *Deep learning* LSTM (esquerda) e CNN (direita)

A construção e treino destes diferentes modelos, uma vez que utilizamos *datasets* de frequencia diária, semanal e mensal encontram-se nas pastas **Daily Model**, **Weekly Model** e **Daily Model** respetivamente.

5 Hiperparâmetros de otimização

Como forma de obtermos e encontrar-mos o melhor modelo possível, e de forma a poder-mos comparar a *performance* dos diferentes modelos, optmamos por aplicar a ambos os modelos de *machine learning* os mesmos parâmetros de otimização. Os hiperparâmetros que optamos por experimentar foram o número de épocas de treino, taxa de aprendizagem e o *batch size*.

	Número de épocas	Timesteps	Taxa Aprendizagem	*Batch size*
1 Otimização	100	2	0.1	1
2 Otimização	200	4	0.01	2
3 Otimização	300	6	0.001	3
4 Otimização	100	2	0.1	1
5 Otimização	200	4	0.01	2
6 Otimização	300	6	0.001	3
7 Otimização	100	2	0.1	1
8 Otimização	200	4	0.01	2
9 Otimização	300	6	0.001	3

Tabela 1: Valores utlizado nas optimizações

VER MELHOR ESTAS OTIMIZAÇÕES, EM ESPECIAL O BATCH SIZE POR CAUSA DOS MESES E SEMANAS

Estas otimizações são possíveis de serem observadas nos *notebooks* do repositório onde são feitas a construção e treino dos diferentes modelos, nas pastas **Daily Model**, **Weekly Model** e **Daily Model**.

6 Resultados

7 Conclusão

Após a experimentação com vários tipos de modelos de redes neuronais nos datasets Traffic Incidents Braga e S&P500 Index Price Prediction, concluímos que os modelos que melhores previsões são capazes de realizar, são os modelos CNN multivariável com a otimização 6 (timestep = 8, epochs = 50, batch size = 5) com valor de rmse de 0.107235, LSTM multivariável v2 com o modelo 1 e otimização 9 (timestep = 10, epochs = 120, batch size = 8) cuja rmse obtida é de 0.0213, respetivamente.

No entanto, seria interessante também implementar um modelo LSTM multivariável multistep com modelo auxiliar de previsão de variável independente para verificar se o modelo CNN multistep multivariável escolhido, continuaria ou não, a ser o modelo com melhor desempenho. Pertinente também, seria a constituição de um modelo CNN multistep multivariável com modelo auxiliar, pois a nossa suspeita é de que este seria o modelo ideal para o dataset em questão.

Relativamente ao *benchmark* sobre o dataset S&P500 Index Price Prediction, seria curioso modelar uma CNN multistep multivariável com modelo auxiliar, visto a CNN multistep multivariável ter obtido um valor muito próximo de RMSE do melhor modelo encontrado.

Para além fazer das sugestões de melhorias de *benchmark* sugeridas acima, uma otimização dos parâmetros *learning rate*, que poderia permitir ao nossos modelos não convergir tão rápido, assim como otimizar os *betas* usados no *optimizer Adam*. Seria também interessante ver como os modelos se comportavam usando outros tipos de *optimizers*, bem como o aumento da profundidade dos diferentes modelos, e também a alteração do número de neurónios de cada layer.

8 Referências

Comprehensive data exploration with Python https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Training Loss and Validation in Deep Learning https://stackoverflow.com/questions/48226086/training-loss-and-validation-loss-in-deep-learning

How To Prevent Overfitting https://programming-review.com/machine-learning/overfitting