ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

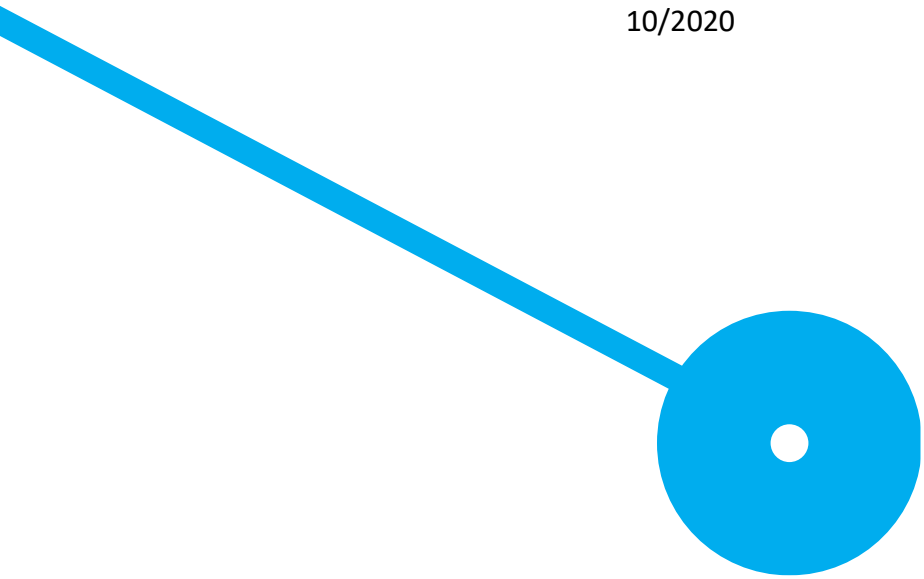
L

LICENCIATURA
ÁREA DE ESPECIALIZAÇÃO

Análise Comparativa de Algoritmos de Aprendizagem com Base em Séries Temporais

Luís Marques

10/2020



[Página propositadamente deixada em branco.]

ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

LICENCIATURA
ENGENHARIA INFORMÁTICA

Análise Comparativa de Algorítmos de Aprendizagem com Base em Séries Temporais

Luís Marques
Prof. Fábio Silva

[Página propositadamente deixada em branco.]

Biografia do Autor

Luís Marques é um estudante de Engenharia Informática na Escola Superior de Tecnologia e Gestão do Politécnico do Porto, sendo o ano atual o seu terceiro ano da Licenciatura. Nascido a 8 de junho de 1999 em Vila Nova de Famalicão, o autor nutre bastante interesse em tecnologia e gosta particularmente de desenvolvimento de software backend. Também demonstra bastante interesse na área de Inteligência Artificial e de *Machine Learning*.

[Página propositadamente deixada em branco.]

Resumo

De modo a fazer uma análise comparativa entre diferentes algoritmos de aprendizagem, é necessário fazer um estudo sobre eles, em primeiro lugar. Para isto foram utilizadas séries temporais.

Este projeto representa o estudo realizado em torno de alguns modelos, mais exatamente as variantes do modelo *ARIMA*. O algoritmo para cada modelo estudado foi feito utilizando alguns recursos disponibilizados pelo Professor Fábio Silva e com a ajuda de algumas bibliotecas de *Python*.

Palavras Chave: *ARIMA*, Inteligência Artificial, *Machine Learning*, *Python*

[Página propositadamente deixada em branco.]

Conteúdo

[Página propositadamente deixada em branco.]

Lista de Figuras

[Página propositadamente deixada em branco.]

Glossário

Autoregressive Integrated Moving Average with Exogenous Variable é um modelo de aprendizagem de média móvel integrado autoregressivo com variáveis exógenas. xiii

Autoregressive Integrated Moving Average é um modelo de aprendizagem de média móvel integrado autoregressivo. xiii, 3

dataset (Conjunto de dados) é uma coleção de dados normalmente tabulados. Por cada elemento destacam-se várias características. Cada coluna representa uma variável particular. Cada linha corresponde a um determinado membro do conjunto de dados em questão. Cada valor é conhecido como um dado. 1, 4

Google é uma empresa multinacional de serviços online e software dos Estados Unidos. x

Google Colab é um ambiente de desenvolvimento com a linguagem Python que não requer configuração e é executado utilizando a Google Cloud. 4, 5

Google Cloud é uma suíte de computação em nuvem oferecida pelo Google. x

grid search é o processo de coleção de dados para configurar os parâmetros ideais para um determinado modelo. 5

Inteligência Artificial é a inteligência similar à humana exibida por mecanismos ou software, para além de também ser um campo de estudo académico. i, iii, x, xiii

Machine Learning é um subcampo da Engenharia e da ciência da computação que evoluiu do estudo de reconhecimento de padrões e da teoria da aprendizagem computacional em Inteligência Artificial. i, iii, xiii, 4

Python é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. iii, x, 4, 5

roadmap é um recurso visual de alto nível que mapeia a evolução do produto/projeto ao longo do tempo. 6

script é um programa de computador, normalmente executado com um interpretador. 5

Seasonal Autoregressive Integrated Moving Average with Exogenous Variable é um modelo de aprendizagem de média móvel integrado autoregressivo sazonal com variáveis exógenas. xiii

Seasonal Autoregressive Integrated Moving Average é um modelo de aprendizagem de média móvel integrado autoregressivo sazonal. xiii

Variáveis Exógenas são variáveis cujos valores são determinados fora do modelo e são impostas ao modelo (no caso do ARIMA serão utilizadas para ajudar a fazer as previsões). 5

[Página propositadamente deixada em branco.]

Abreviaturas

ARIMA Autoregressive Integrated Moving Average. iii, xi, 1, 3, 4, 5

ARIMAX Autoregressive Integrated Moving Average with Exogenous Variable. 5

IA Inteligência Artificial. i, iii

ML Machine Learning. i, iii, 4

SARIMA Seasonal Autoregressive Integrated Moving Average. 5

SARIMAX Seasonal Autoregressive Integrated Moving Average with Exogenous Variable. 5

[Página propositadamente deixada em branco.]

Capítulo 1

Contextualização e Motivação

1.1 Introdução

1.1.1 Contextualização

O presente documento descreve o trabalho realizado na análise comparativa de algoritmos de aprendizagem com base em séries temporais. O trabalho foi proposto pelo Professor Fábio Silva cuja motivação é desenvolver uma plataforma de seleção dos melhores algoritmos para um determinado conjunto de dados em determinadas condições. O trabalho realizado pelo autor decorre do desenvolvimento do projeto final da Licenciatura em Engenharia Informática da Escola Superior de Tecnologia e Gestão do Politécnico do Porto.

1.1.2 Objetivos

Este projeto foi focado no estudo de um modelo de aprendizagem (ARIMA) e as suas variantes, com o objetivo de testar vários *datasets* e desenvolver uma plataforma que informe quais os melhores modelos e as melhores configurações dos mesmo para cada *dataset*. Posto isto, cada modelo teve de ser estudado e compreendido de forma parametrizada para que a integração com qualquer conjunto de dados não cause nenhum problema.

1.1.3 Resultados

≪Resultados obtidos≫

1.1.4 Estrutura

O presente documento está dividido em 5 capítulo principais:

1. Contextualização e Motivação - este capítulo pretende apresentar de forma sucinta e objetiva as circunstâncias a que surgiu o projeto, contendo o contexto, objetivos e resultados do mesmo. Neste capítulo está contida ainda uma parte destinada à fundamentação teórica do projeto, dando a conhecer, de uma forma um pouco abrangente, os principais tópicos.

2. Concetualização do Problema - nesta secção estão detalhados os aspetos mais técnicos do problema, visto que contém a definição dos requisitos e a definição de uma arquitetura concetual.

3. Metodologia de Operacionalização do Trabalho - neste capítulo é explorada e a metodologia de trabalho utilizada. Contém também, no final, uma descrição sobre aspetos mais técnicos do projeto.

4. Discussão dos Resultados - é nesta secção que serão discutidos os resultados obtidos e a explicação de possíveis controversias.

5. Conclusão - este capítulo visa realizar uma conclusão global do projeto e, por fim, uma descrição de possíveis melhoramentos futuros no trabalho desenvolvido.

1.2 Fundamentação Teórica

1.2.1 Série Temporal

Tal como é abordado em temas como a estatística, economia e matemática aplicada, uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. Em modelos de regressão linear a ordem das observações é irrelevante, mas em séries temporais a ordem dos dados é fundamental.

A análise de séries temporais compreende métodos para analisar os dados, a fim de extrair estatísticas significativas e outras características. Para fazer uma previsão com uma série temporal é utilizado um modelo para prever valores futuros com base em valores observados anteriormente.

Existem 4 componentes que uma série temporal pode ter: i) Nível: valor base da série se fosse uma linha reta; ii) Tendência (opcional): comportamento, normalmente linear, crescente ou decrescente ao longo do tempo; iii) Sazonalidade (opcional): padrões repetitivos ou ciclos de comportamento ao longo do tempo; iv) Ruído (opcional): variações nas observações que não podem ser explicadas pelo modelo.

Para concluir, todas as séries temporais têm nível e a maior parte também tem ruído. No entanto, a tendência e a sazonalidade são ocasionais.

1.2.2 Grid Search

Grid search (ou pesquisa em grelha) é o processo de coleção de dados para configurar os parâmetros ideais para um determinado modelo. Dependendo do tipo de modelo utilizado, alguns parâmetros são necessários. Esta pesquisa não se aplica apenas a um tipo de modelo, ela pode ser aplicada ao modelo de aprendizagem para calcular os melhores parâmetros a serem usados para qualquer modelo.

Um ponto importante a sublinhar é que esta pesquisa pode ser extremamente cara em termos computacionais e pode levar muito tempo até obter resultados. O *grid search* construirá um modelo em cada combinação de parâmetros possível. Ele itera por meio de cada combinação de parâmetros e armazena um modelo para cada combinação.

1.2.3 Modelo ARIMA

O modelo de Média Móvel Integrada Autoregressiva - *Autoregressive Integrated Moving Average (ARIMA)* - foi o modelo utilizado para realizar este trabalho, sendo que, todo o projeto foi abordado em torno deste modelo e variações do mesmo. *ARIMA* significa:

- AR: "*Autoregression*"(Autoregressão) - Um modelo que usa a relação entre uma observação e um número de observações atrasadas.
- I: "*Integrated*"(Integrado) - O uso de diferenciação de observações (por exemplo, retirar uma observação de uma observação, no passo anterior) de forma a manter a série temporal estacionária.
- MA: "*Moving Average*"(Média Móvel) - Um modelo que usa a dependência entre uma observação e o erro residual de um modelo de média móvel aplicado a observações atrasadas.

Cada um destes componentes está especificado no modelo como um parâmetro. A notação utilizada é *ARIMA*(p, d, q):

- p: Ordem de atraso - número de observações atrasadas incluídas no modelo.
- d: Grau de diferenciação - número de vezes que observações brutas são diferenciadas.
- q: Ordem da média móvel - tamanho da janela de media móvel

Capítulo 2

Concetualização do Problema

2.1 Requisitos

Ao longo da realização do projeto, foram sendo definidos os requisitos para o desenvolvimento do mesmo, sendo eles:

- Estudar modelo *ARIMA* e variantes;
- Construir um serviço de testes de *datasets* com os modelos estudados anteriormente;
- Testar varios *datasets* e guardar melhores modelos e configurações dos mesmos;
- Integrar projeto com *Google Colab*.

Destas necessidades definidas em cima, podemos retirar algumas tarefas e subtarefas implícitas em cada uma.

2.1.1 Estudo do Modelo

- Estudo de algumas bibliotecas de *Python* utilzadas ao longo do trabalho;
- Realização de alguns exercícios de *Machine Learning* com *Python* (e.g., abrir e manusear um *dataset*, formatar as datas, imprimir ou exportar os dados, mostrar ou exportar um gráfico);
- Estudo do modelo *ARIMA* (e.g., propriedades, funcionamento, características)

- Estudo das variantes do modelo *ARIMA* que são: *ARIMAX* (Média Móvel Integrada Autoregressiva com Variáveis Exógenas), *SARIMA* (Média Móvel Integrada Autoregressiva Sazonal) e *SARIMAX* (Média Móvel Integrada Autoregressiva Sazonal com Variáveis Exógenas).

2.1.2 Construção do Serviço de Testes

- Estruturação dos *scripts*;
- Realização dos *scripts* na linguagem *Python*;
- Testar a veracidade e qualidade do programa desenvolvido.

2.1.3 Teste dos Datasets

- Executar uma *grid search* para encontrar melhores modelos e melhores configurações;
- Executar os melhores modelos com as melhores configurações e guardar resultados.

2.1.4 Integração com *Google Colab*

- Estudo da plataforma *Google Colab* (e.g., funcionalidades, compatiblidade);
- Colocar *scripts* na plataforma;
- Correr alguns testes.

2.2 Arquitetura Concetual

≪Arquitetura concetual - Nesta arquitetura deverá ser claro o que foi efetivamente desenvolvido pelo Estudante e aquilo que foi desenvolvido por terceiros. A arquitetura deverá realçar aspetos relacionadas com integração, protocolos, entre outros, e que o estudante deverá clarificar.≫

Capítulo 3

Metodologia de Operacionalização do Trabalho

3.1 Processo e Metodologia de Trabalho

Foi adotada uma abordagem iterativa incremental, com algumas reuniões de acompanhamento ao longo do projeto. No decorrer de 15 semanas, houve a necessidade de criar 4 iterações principais pelas quais foram divididas as tarefas em cima descritas.

O *roadmap* do projeto, que se encontra representado na forma de um diagrama de Gantt, apresenta, mais detalhadamente, a atribuição das tarefas nas diferentes iterações e a duração das mesmas.

4 sprints principais: - realização de exercicios de ML com Python - estudo do modelo ARIMA - estudo de variantes do modelo ARIMA - construção do serviço de testes, teste dos datasets e integração com google colab

Roadmap do projeto (Excel)

3.2 Desenvolvimento da Solução

Capítulo 4

Discussão dos Resultados

4.1 Apresentação e Discussão dos Resultados

4.2 Apresentação dos Impedimentos e/ou Constrangimentos

Capítulo 5

Conclusão

5.1 Reflexão Crítica dos Resultados

5.2 Conclusão e Trabalho Futuro

[Página propositadamente deixada em branco.]

Referências

[1] *www.newthinktank.com/2019/01/latex-tutorial*

[2] *pt.overleaf.com*

[3] *www.dickimaw-books.com/gallery/glossaries-styles*

[4] *www.wikipedia.org*

[5] *en.wikipedia.org/wiki/Time_series*

[6] *www.ime.unicamp.br/ hlachos/MaterialSeries.pdf*