Redes Neurais Artificiais – Projeto Prático 1

Análise Exploratória e Visualização de Dados

Elloá B. Guedes

Escola Superior de Tecnologia
Universidade do Estado do Amazonas
Av. Darcy Vargas, 1200 – Manaus, AM
ebgcosta@uea.edu.br

5 de agosto de 2020

1. Contextualização

Em 11 de Março de 2020, a Organização Mundial de Saúde (OMS) declarou o surto de uma pandemia em nível global causada pelo coronavírus SARS-CoV-2, o qual foi identificado inicialmente na China em Dezembro de 2019 [1]. Este vírus é responsável pela COVID-19, que apresenta um espectro clínico variando de infecções assintomáticas a quadros graves, manifestando-se desde como uma Síndrome Gripal (sensação febril ou febre, dor de garganta, cefaléia, tosse e coriza) até uma pneumonia severa [2]. De acordo com dados da OMS, em 17 de Julho de 2020, em nível global foram registrados mais de $13,5$ mi de casos e 584.940 óbitos reportados.

No Brasil, o Ministério da Saúde declarou situação de transmissão comunitária em 20 de Março de 2020, em face de 904 casos confirmados em 24 estados e 11 óbitos, com o intuito de que todos os gestores nacionais adotassem medidas para promover o distanciamento social e evitar aglomerações [3]. Em 17 de Julho de 2020, o Brasil já ultrapassava a marca de 2 milhões de casos, 76 mil óbitos e $1,2$ milhões de casos recuperados, registrando taxa de letalidade de $3,8\%$, conforme painel de casos do Ministério da Saúde.

Considerando o contexto atual, o objetivo das equipes envolvidas no projeto consiste em analisar os dados da COVID-19 disponibilizados pela Prefeitura de Manaus, com o intuito de realizar uma análise exploratória e visualização de dados, passando também pelas etapas de limpeza e organização. As equipes serão compostas de 5 alunos e devem trabalhar de maneira organizada e sistemática para produzir os seguintes artefatos:

1. **Repositório no GitHub**. De caráter público (privado durante o desenvolvimento, público na ocasião da entrega), incluindo todos os integrantes da equipe e contendo as evidências de código e de progresso, registrando o trabalho dos múltiplos integrantes;

2. **Código-Fonte Python**. Deve ser incluído no repositório sob a forma de um único ou múltiplos Jupyter Notebooks, contendo todas as atividades de programação consideradas para atender ao que se pede no projeto;

3. **Relatório Técnico**. Um texto no formato LaTeX utilizando o template para artigos da Sociedade Brasileira de Computação. Este texto deve ser construído na forma de parágrafos, utilizando um encadeamento lógico, com figuras e tabelas que a equipe considerar apropriadas para esclarecer os questionamentos avaliativos. O tamanho mínimo do texto para avaliação é de 3 páginas considerando o template fornecido. Serão penalizados com 50% da nota aquelas equipes que usarem tamanhos de figuras exarcebados, espaços em branco ou qualquer outra estratégia que vise aumentar o número de páginas sem a inclusão de conteúdos significativos.

2. Detalhamento da Atividade

A base de dados a ser utilizada pelas equipes encontra-se disponível em <https://covid19.manaus.am.gov.br/wp-content/uploads/Manaus.csv>. Esta base de dados contém cerca de 25 MB de informações textuais no formato CSV (*Comma-Separated Values*) e codificação ISO 8859-1 em virtude dos acentos nas nomeclaturas dos bairros.

2..1 Visão Geral dos Casos Confirmados

O primeiro passo que as equipes devem efetuar após a importação da base de dados consiste em considerar apenas os casos confirmados, excluindo todos os demais registros distintos. Esta informação encontra-se na coluna `classificacao` da base de dados. Atenda ao que se pede, sempre utilizando as bibliotecas Python que preferir para responder às seguintes perguntas:

1. Quantos atributos descrevem cada exemplo? Quais são eles?

2. Quantos casos confirmados há em Manaus, cumulativamente?

3. A qual período de tempo a base de dados se refere, isto é, qual o registro mais antigo e qual o mais recente? Leve em conta a data de notificação.

Para fins da análise considerada no escopo deste projeto, vamos excluir todos os atributos relativos às comorbidades, sintomas, etnia, profissão, outras datas que não a de notificação, origem e outros que não estiverem envolvidos no contexto do trabalho solicitado. Estes atributos serão considerados irrelevantes para fins de simplificação. Exclua todas as linhas em que houver dados faltantes para os atributos remanescentes.

Visando efetuar uma análise exploratória dos dados, respondam, pelo menos, às seguintes questões:

1. Quantos exemplos e atributos há na base de dados após a limpeza e organização?

2. Qual a porcentagem de indivíduos recuperados em relação ao todo?
3. Os casos acometeram mais indivíduos do sexo masculino ou feminino?
4. Qual a média e desvio padrão de idade dos indivíduos que contraíram COVID-19? Qual o indivíduo mais jovem e o mais idoso a contraírem tal enfermidade?
5. Qual o bairro com maior incidência de casos?
6. Quais os três bairros com maior incidência de casos recuperados?
7. Quais os tipos de testes efetuados, segundo os dados? Indique os dados de maneira quantitativa e percentual.
8. Qual taxa de letalidade pode ser calculada a partir do conjunto de dados? Para calcular esta taxa, considere a fração do total de óbitos pelo total de casos;
9. Qual o tipo de correlação, mediante coeficiente de correlação de Pearson, entre a idade e o número de casos? Para responder a esta pergunte, agrupe o número de casos por idade e efetue o cálculo de tal coeficiente. Indique, a partir do resultado, a natureza desta correlação, se é positiva ou negativa, e qual sua intensidade.

2..2 Visualização de Dados

Para a visualização de dados, apresentem gráficos que denotem as informações a seguir. Explicitem o que denotam os eixos x e y claramente, escolham um esquema de cores compatível com a escrita técnico-científica e utilizem legendas ao incluí-los no texto.

1. Construa um histograma denotando a quantidade de casos nos 10 bairros em que houve mais casos registrados. Inclua todos os bairros remanescentes em uma categoria denominada "Outros." Denote as informações de maneira percentual;
2. Denote, por sexo, o boxplot da idade dos casos confirmados. Há outliers?
3. Denote em um gráfico de barras o número de novos casos por dia, considerando os 10 últimos dias existentes na base de dados;
4. Repita o gráfico anterior considerando o número de casos recuperado;
5. Construa um histograma que denote a quantidade percentual de casos por grupo etário, considerando que cada grupo contempla uma década (0 a 10 anos, 11 a 20 anos, etc.);
6. Elabore um gráfico que mostra o cumulativo de casos notificados ao longo do tempo;
7. Faça um gráfico do tipo *scatterplot* que denote a idade versus o número total de casos registrado para aquela idade. Aproveite o processamento efetuado para o cálculo da correlação. É possível observar alguma tendência?

2..3 Tipos de Tarefas

Recapitulem que dados fornecem experiência sobre um problema. No caso em questão, sugira:

1. Uma tarefa de classificação mediante Aprendizado Supervisionado que poderia ser feita com esta base de dados. Qual seria o atributo-alvo? Quais métricas de desempenho poderiam ser aplicadas? Que tipo de validação seria apropriado?
2. Uma tarefa de regressão mediante Aprendizado Supervisionado que poderia ser feita com esta base de dados. Qual seria o atributo-alvo? Quais atributos preditores a equipe considera relevantes para o cenário?
3. Bônus: Qual tarefa de Aprendizado Não-Supervisionado poderia ser concebida neste contexto?

2..4 Tecnologias e Sugestões

Para a realização desta tarefa, é obrigatório o uso da linguagem de programação Python 3.6+ e das bibliotecas `pandas` e `numpy`. Para os gráficos, podem ser utilizadas as bibliotecas `matplotlib` ou `seaborn`, conforme preferência da equipe. Outras bibliotecas complementares também podem ser utilizadas, especialmente na etapa de análise exploratória.

Para aquelas equipes com restrições de *hardware*, recomenda-se o uso do Google Colab, disponível em: <http://colab.research.google.com/>. O notebook produzido ao final deve ser incluído no repositório GitHub. Para as demais equipes, recomenda-se o uso do gerenciador de pacotes Anaconda e a utilização de ambientes virtuais `conda env`.

3. Critérios de Avaliação

Os critérios de avaliação levarão em conta a organização do repositório, a qualidade do código produzido, a completude das tarefas solicitadas, a documentação, a qualidade textual do relatório em termos de utilização da norma culta, coesão, coerência, o respeito aos prazos e a colaboração da equipe na elaboração do projeto.

4. Links Úteis

- Repositório com Exemplos
- Template da SBC
- Panorama da COVID-19 no Brasil

Referências

1. ORGANIZAÇÃO MUNDIAL DE SAÚDE (Organização das Nações Unidas). Director-General's opening remarks at the media briefing on COVID-19. 2020. Disponível em https://bit.ly/30jiUXY. Acesso em 17 de Julho de 2020.

2. GOVERNO FEDERAL (Brasil). Ministério da Saúde. Coronavírus. Brasília, 17 jul. 2020. Disponível em: https://coronavirus.saude.gov.br/. Acesso em: 17 jul. 2020.

3. GOVERNO FEDERAL (Brasil). Ministério da Saúde. Ministério da Saúde declara transmissão comunitária nacional. Brasília, 20 mar. 2020. Disponível em: https://www.saude.gov.br/noticias/agencia-saude/46568-ministerio-da-saude-declara-transmissao-comunitaria-nacional. Acesso em: 17 jul. 2020.

4. GOVERNO FEDERAL (Brasil). Ministério da Saúde. COVID-19 - Painel Geral. Brasília, 16 jul. 2020. Disponível em: https://covid.saude.gov.br/. Acesso em: 17 jul. 2020.