Exercício-Programa 2: Previsão de Estágio da Infecção por COVID-19 – Dados do HC

Entrega: 30/06/2025

1 Motivação

Durante a pandemia foi lançado um banco de dados público e anônimo sobre pacientes com COVID-19[1], cujas fontes são cinco hospitais da cidade de São Paulo que tratam este tipo de infecção. Os bancos de dados possuem informações sobre o paciente e informações de resultados de exames. Os arquivos presentes no diretório `dados` contém uma versão levemente pré-processada destes dados.

A infecção pela COVID-19 possui três fases detectáveis. Na primeira fase o paciente está infectado mas a reação de seu sistema imunológico pode ainda não ter-se iniciada, o que é detectado por algum exame que detecta COVID-19 com resultado positivo. Na segunda fase, o corpo do paciente já começou a reagir e a produzir anticorpos do tipo IgM (imunoglobulina M), o que é detectado por um exame que verifica a presença deste tipo de anticorpo. Na terceira fase, o paciente está curado da infecção e o corpo já produziu anticorpos do tipo IgG (imunoglobulina G) que tendem a revelar que o paciente teve a doença e em alguns casos conferem imunidade ao paciente para uma nova infecção, o que é detectado pela presença desse tipo de anticorpo. Estas fases podem ter sobreposição, ou seja, é possível que o paciente esteja infectado e o corpo igm esteja produzindo anticorpos IgM; e é possível que a produção de IgG se inicie enquanto corpo ainda está combatendo a infecção com anticorpos IgM.

O objetivo desse trabalho é construir uma rede neural que, dado um conjunto de exames realizadas em um determinado paciente de COVID-19, se o exame `COVID-19 IgG` está positivos, e desta forma prever em que fase da doença se encontra o paciente. É possível que ao considerar o histórico de exames do paciente, obtenha-se uma previsão mais acurada; no entanto, neste exercício, apenas consideraremos dados estáticos. Desta forma, por *entrada* estamos considerando os exames realizados por um determinado paciente numa determinada data. Por *saída* estaremos considerando o resultado do exame de detecção de COVID-19 IgG como resultado binário; ou seja, o resultado deste exame deve ser visto apenas como "detectado" ou "não detectado".conjunto filtrado de pacientes

Neste trabalho vocês deverão entregar um arquivo do tipo Zip contendo programas em Python, dados e um relatório geral descrevendo suas atividades. Nas seções a seguir descrevemos quais programas e dados devem ser entregues. Em particular, o relatório deve possuir a seguinte estrutura.

1. Título e nome dos autores.

[1] `https://repositoriodatasharingfapesp.uspdigital.usp.br/`

2. Resumo.

3. Introdução, que deve conter os seguintes elementos: motivação para o trabalho, contendo um ou dois parágrafos; objetivos do trabalho, contendo um parágrafo; estrutura do relatório a seguir, contém um parágrafo.

4. Metodologia, que descreve os seguintes itens: dados pré-processamento dos dados; arquitetura da rede neural; descrição dos experimentos.

5. Resultados, em que são apresentados os valores obtidos nos experimentos, apresentando tabelas e gráficos.

6. Discussão. No nosso caso esta seção deverá centrar-se na viabilidade e utilidade de usar arquitetura neural proposta como uma previsor da fase da infecção.

7. Bibliografia. Aqui você deve inserir todas as fontes utilizadas na realização desse trabalho; as fontes devem estar citadas em alguma parte anterior do trabalho.

2 Metodologia

A metodologia deve abordar o pré-processamento dos dados; a arquitetura do software e, em particular, da rede neural; e a descrição dos experimentos.

2.1 Dados

O repositório de dados montado pela USP e FAPESP possui informações sobre pacientes de cinco hospitais da cidade de São Paulo, a saber, o Hospital das Clínicas (HC), a Beneficência Portuguesa, o Hospital Sírio Libanês, o Hospital Albert Einstein e o Grupo Fleury.

Neste exercício, para evitar problemas de heterogeneidade entre os dados, diferença de nomes e outras divergências entre as diversas instituições, vamos nos concentrar apenas nos dados do Hospital das Clínicas da Faculdade de Medicina da USP (HC).

O repositório de dados montado pela USP e FAPESP possui informações sobre pacientes de cinco hospitais da cidade de São Paulo, a saber, o Hospital das Clínicas (HC), a Beneficência Portuguesa, o Hospital Sírio Libanês, o Hospital Albert Einstein e o Grupo Fleury.

Neste exercício, para evitar problemas de heterogeneidade entre os dados, diferença de nomes e outras divergências entre as diversas instituições, vamos nos concentrar apenas nos dados do Hospital das Clínicas da Faculdade de Medicina da USP (HC).

No repositório de dados do HC há três planilhas. Uma delas, no formato XLSX, é um dicionário de dados, explicando os campos das outras duas planilhas. A segunda planilha, no formato CSV, trata de informações do paciente, da qual podemos utilizar apenas as informações sobre gênero/sexo e ano de nascimento. A terceira planilha, também no formato CSV, é a planilha principal que contém os exames de um paciente. As duas planilhas CSV, na realidade, tem seus campos separados por uma barra vertical '|', e as vírgulas nesse caso não devem ser consideradas como separadores de campos.

2.2 O Pré-processamento

O pré-processamento é uma fase importantíssima que nem sempre recebe a devida atenção. No nosso caso, para ressaltar a importância dessa fase, metade da nota vai vir para processamento.

Neste exercício o pré-processamento deve transformar as planilhas csv recebidas em uma única planilha que será a entrada do processamento neural. Nesta planilha de entrada, a chave de identificação deve ser o identificador do atendimento. O identificador do paciente não precisa ser um campo desta planilha, mas ele deve ser usado para inserir dois campos importantes: a idade e o sexo do paciente. Os demais campos são todos os exames feitos por um paciente e a coluna deve vir preenchida com o resultado do exame; o resultado depende do exame, e alguns exames tem um resultado numérico enquanto outros têm o resultado binário, que deve ser codificado como zero ou um. É importante salientar que nem todos os atendimentos fazem todos os exames possíveis e quando um atendimento não tiver um determinado exame, o resultado desse exame deverá ser zero se for um exame numérico, ou 0,5 se for um exame binário.

Como o nosso objetivo principal é prever se o IgG é detectado ou não, atendimentos que não contenham este exame devem ser excluídos da planilha de entrada. Alternativamente, para aumentar o volume de dados, você pode realizar algumas hipóteses que inserem novos pacientes na lista. Por exemplo, se um paciente possui algum teste relacionado a detecção de COVID-19, mas não o teste de IgG, pode-se incluir o paciente assumindo que o IgG é falso. Desta forma, seriam admitidos no *conjunto filtrado de pacientes* todos os pacientes com algum exame relacionado a detecção de COVID-19 e seus anticorpos.

Esta fase de pré-processamento deverá produzir uma planilha (CSV) ou o banco de dados (SQL) que servirá de entrada para o treinamento da rede neural. Você deverá entregar tanto a planilha/banco de dados quanto os programas em Python que foram utilizados para gerar esta planilha.

Arquitetura da rede neural

Nesta seção você deverá descrever arquitetura da rede neural utilizada e incluir os programas em Python utilizados para treinar e rodar a rede. Não há qualquer problema que esses programas sejam modificações dos programas fornecidos no exercício em sala, mas se na descrição da arquitetura você simplesmente disser algo como “a mesma arquitetura do exercício”, vou entender que você não tem ideia do que está sendo utilizado. Em particular, o seu relatório deve explicitar Qual a função que está sendo otimizada.

É necessário descrever o número de entradas, o número de camadas, o tamanho de cada uma das camadas escondidas; é necessário descrever a função utilizada para calcular a perda (loss) e eventuais otimizações do usado no treinamento. É necessário descrever como a rede é utilizada no treinamento e durante sua execução, quando iremos prever o estágio de infecção de um paciente, mencionando formato das entradas, o tamanho dos batches, e o formato da saída.

Entregar os programas Python usados no treinamento e na execução da rede, bem como um minúsculo manual do usuário. No relatório essa parte não deverá ter mais do que três parágrafos.

2.3 Descrição dos experimentos

Nessa parte você deve descrever como foram feitos os treinamentos, e o que está sendo medido (apresentando uma fórmula de preferência) no experimento. Como a quantidade de dados é relativamente pequena, espera-se que você realize um processo de "k-fold treinamento' ', com $k \geq 5$. Em cada treinamento, você deverá usar 80% do conjunto de dados para treino, e 20% para validação, mas se você alterar esta proporção, informe iss no relatório. Descreva como essa divisão dos dados foi obtidas.

Neste caso é esperado que você repita o treinamento 5 vezes. Se o tempo de treinamento for excessivo (muitas horas) você pode decidir por utilizar $k = 3$, mas terá de apresentar o tempo de treinamento como justificativa para este encurtamento do experimento.

3 Resultados

Primeiramente, vamos nos concetrar sobre as medidas a serem feitas, que são todas baseadas na matriz de confusão. Uma *matriz de confusão* é uma tabela em que cada linha da representa instâncias de uma classe real enquanto cada coluna representa instâncias da classe prevista. No caso de problemas de classificação binária, temos uma matriz de 4 células, cada uma com seu nome:

	P	N
T	TP	FN
F	FP	TN

TP *Verdadeiros Positivos:* número de instânicas classificadas corretamente como positivas no teste.

FN *Falsos Negativos:* número de instânicas verdadeiras erroneamente classificadas como negativas no teste.

FP *Falsos Positivos:* número de instânicas falsas erroneamente classificadas como positivas no teste.

TN *Verdadeiros Positivos:* número de instânicas classificadas corretamente como negativas no teste.

Note que o número de testes $n = TP + TN + FP + FN$. A partir dos valores da matriz de confusão, definimos as seguintes medidas básicas.

$$\text{Acurácia} = \frac{TP+TN}{n} \qquad \text{Precisão} = \frac{TP}{TP+FP}$$

$$\text{Cobertura} = \frac{TP}{TP+FN} \qquad \text{Medida-F} = \frac{2*\text{Precisão}*\text{Cobertura}}{\text{Precisão}+\text{Cobertura}}$$

Na literatura médica, a medida de Cobertura é chamada de Sensibilidade.

Você deve calcular nesta fase de resultados pelo menos os valores de destas quatro medidas para os dados de teste de detecção do IgG para cada um dos k-treinamentos. No final, você deve apresentar a média e desvio padrão, o melhor e o pior valor para cada uma das medidas de acurácia, precisão, cobvertura e medida-F.

Se você propuser alguma variação da rede neural, deverá realizar um experimento para cada variante e apresentar os dados relativos a cada uma das variantes.

4 Equipe

Este exercício poderá ser realizado individualmente ou em duplas. Duplas possuem exatamente dois elementos, e nenhum desvio desta regra será contemplado.

Informar na capa do seu relatório quem são os membros da equipe e seus números USP. Só há necessidade de realizar uma entrega por dupla.

5 Instruções para entrega

Você deve submeter via eDisciplinas apenas um arquivo `ep2.zip` contendo a sua solução até às 23:59 do data de entrega. Este arquivo compactado deverá conter três diretórios. Um diretório `PRE` com os programas de pré-processamento e a planilha a ser usada no treinamento gerada durante o pré-processamento; um diretório `NN` com os programas de Treinamento execução da rede neural; e um diretório `REL` com o relatório contendo os itens solicitados. Este relatório Deverá estar no formato PDF; você pode usar qualquer editor que achar mais conveniente para editá-lo.

Certifique-se que o arquivo foi submetido sem problemas (baixando e executando os programas) e que ele consiste em um script executável escrito em Python 3.

6 Avaliação

O pré-processamento e sua descrição no relatório valem 50% da nota, o resto do relatório 45% e o código da rede neural vale só 5%.

Espero sinceramente que você aproveite esse trabalho, que é pertinente à disciplina, atual, socialmente relevante e muito divertido.