Exercício-Programa 4: Previsão de Estágio da Infecção por COVID-19

Entrega: 12/07/2020

Motivação

Recentemente foi lançado um banco de dados público e anônimo sobre pacientes com covid-19[1], cujas fontes são três 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 pelo exame PCR com resultado positivo. Na segunda fase, o corpo do paciente já começou a reagir e a produzir anticorpos do tipo igm, 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 que 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 PCR seja positivo e o corpo já 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 periodicamente nos pacientes infectados pela covid-19, prever quais exames PCR, igm e igg estão 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 dos exames de detecção PCR, igg e igm como resultados binários; ou seja, cada um desses resultados deve ser visto apenas como "detectado" ou "não detectado".

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.
2. Resumo

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

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: 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.

A seguir detalharemos elementos importantes do relatório.

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.

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.

Eu já dei uma pequena transformada e normalizada nos dados fornecidos, mas aqui vocês deverão fazer diversas outras atividades. Em particular vocês devem decidir quais serão os dados de entrada para rede neural, que devem ser resultados de exames. É importante notar que não há nenhuma garantia e que todos os pacientes tenham realizados os mesmos exames em todas as datas, então vocês devem indicar como estarão tratando dados inexistentes/ faltantes. É razoável que , tendo três fontes diferentes de exames, esses não tenham exatamente os mesmos nomes em todas as fontes, e é importante verificar se as unidades de resposta de cada um dos exames são as mesmas em todos as fontes.

Similarmente, é muito improvável que todos os pacientes tenham realizado os três exames de detecção de saída em todos os exames, então é importante dizer como você vai lidar se não houver algum dado de saída numa determinada instância. No entanto para não alongar muito esta parte do relatório, você não deve levar mais do que duas páginas nesta descrição. Use figuras se isto facilitar a explicação.

Esta fase de pré-processamento deverá produzir uma planilha (csv) ou o banco de dados (sql) que sirva 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.

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.

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 $k = 10$. Em cada treinamento, você deverá usar 90% do conjunto de dados para treino, e 10% para validação. Descreva como essas partes foram obtidas.

Neste caso é esperado que você repita o treinamento 10 vezes. Se o tempo de treinamento for excessivo (muitas horas) você pode decidir por utilizar $k = 5$ ou $k = 3$, mas terá de apresentar o tempo de treinamento como justificativa para este encurtamento do experimento.

Resultados

O mínimo que você deve apresentar nesta fase de resultados são os valores de acurácia para cada uma das três medidas para cada um dos k-treinamentos. No final, você deve apresentar a média e desvio padrão, o melhor e o pior valores de acurácia para cada uma das três medidas de saída.

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

Equipe

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

Instruções para entrega

Você deve submeter via eDisciplinas apenas um arquivo `ep4.zip` contendo a sua solução até às 23:59 do dia 12/07/2020. Este arquivo compactado deverá conter três diretórios.

Um diretório `PRE` com 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.

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.