INF-0615 – Aprendizado de Máquina Supervisionado I

Trabalho 3 - Árvores de Decisão e Florestas Aleatórias

Diagnóstico de COVID-19 a partir de análise sanguínea

Data de Entrega: 25/09/2022

1 Descrição do Problema

A pandemia do vírus COVID-19 afetou diretamente e indiretamente todas as sociedades do mundo ao longo dos últimos dois anos. Muitas pessoas de diferentes nações, idades e classes sociais foram contaminadas, apresentando diferentes quadros clínicos de reação ao vírus. Um dos principais desafios enfrentados pelas autoridades de saúde foi realizar de forma rápida e eficiente a detecção da presença do COVID-19 nas pessoas. Um resultado rápido permite um isolamento prévio e acompanhamento médico desde o estágio inicial da doença.

Nesse trabalho, **vocês irão predizer se um paciente apresenta o vírus do COVID-19 a partir dos níveis dos elementos e substâncias presentes no seu sangue**. Comparado com os exames mais acurados de COVID-19, a predição por amostra sanguínea permite uma detecção mais rápida, mais barata e mais escalável.

As bases de dados disponíveis foram originadas de amostras de pacientes do Hospital Beneficência Portuguesa de São Paulo, deixada a público para fins de pesquisa e estudo. Elas foram pré-processadas de forma que fosse possível aplica-las para o desenvolvimento deste trabalho.

Todas as três divisões (treino, validação e teste) apresentam 101 atributos, sendo um deles chamado **Resultado** (última coluna das bases de dados) o qual armazena o valor alvo que vocês devem predizer. Os demais atributos são utilizados para predição do valor alvo, e mensuram quantidades de diferentes elementos e substâncias no sangue como taxas de Hemoglobina, Hematócrito, Plaquetas, Magnésio, Fosforo, pressão, e etc.

2 Tarefas

Neste Trabalho, pedimos que vocês:

1. **Separem o conjunto de validação**. Tomem a base *covid_analysis_train_val_sets.csv* e façam o *split* em 80% para treinamento e 20% para validação. Lembrem-se de manter o mesmo conjunto de validação para todos os modelos. Além disso, notem que a base de dados de treinamento pode ter casos duplicatos. Assim **certifiquem-se que as duplicatas foram removidas antes da divisão em treino e vaidação**.

2. Inspecionem os dados de treinamento. Quantos exemplos há de cada classe? O dataset está desbalanceado? Se sim, como vocês lidarão com o desbalanceamento?

3. Treinem uma árvore de decisão como *baseline* e reportem a matriz de confusão relativa e a acurácia balanceada nos conjuntos de treinamento, validação e teste.

4. Treinem outras árvores de decisão variando o tamanho das árvores geradas. Plotem a acurácia balanceada no conjunto de treinamento e validação pela profundidade da árvore de decisão. Identifiquem as regiões de *underfitting*, ponto ótimo e *overfitting*. Tomem a árvore com tamanho ótimo e reportem sua matriz de confusão relativa e a acurácia balanceada no **conjunto de teste**.

5. Explorem pelo menos 2 possíveis subconjuntos de features (*feature selection*) para treinar duas (ou mais) árvores de decisão. Tomem o melhor modelo baseado na acurácia balanceada no conjunto de validação, e reportem a matriz de confusão relativa e a acurácia balanceada do no conjunto de teste (**Dica: observem a importância de cada feature por meio do atributo variable.importance - vejam os exercícios 07 e 08 na parte em que calcula-se a importância de cada feature. Vocês podem escolher as top-10 features mais importantes, menos importantes, misturar ambas, etc.)**

6. Treinem várias florestas aleatórias variando o número de árvores. Plotem a acurácia balanceada no conjunto de treinamento e validação variando o número de árvores geradas. Identifiquem as regiões de *underfitting*, ponto ótimo e *overfitting*. Reportem também a matriz de confusão relativa e a acurácia balanceada no teste para a floresta com o melhor número de árvores.

7. Escreva um relatório de no máximo 5 páginas reportando:

   (a) A diferença de desempenho entre o *baseline* e os outros modelos mais complexos gerados.

   (b) Houve overfitting ? Houve underfitting ? Analisem as curvas viés/variância geradas ao longo do trabalho.

   (c) uma Seção de conclusão do relatório explicando a diferença entre os modelos e o porquê que estas diferenças levaram a resultados piores ou melhores.

3 Opcionais

Como visto em aula, a técnica de *ensemble* pode auxiliar a realização da tarefa quando apresenta uma base de dados desbalanceada. Nestes casos, treinamos cada modelo do *ensemble* com as mesmas quantidades de exemplos de todas as classes presentes selecionados de forma aleatória. Isso aumenta a diversidade dos modelos e pode auxiliar no processo de predição. No entanto, a *Random Forest*, apesar de ser uma técnica de *ensemble*, não aplica este balanceamento por árvore, criando assim modelos que também pode sofrer com o desbalanceamento. Neste contexto, pedimos que vocês:

1. Implementem manualmente o protocolo *Random Forest* de forma que cada árvore na floresta tenha as mesmas quantidades de exemplos das duas classes. Note que, para cada modelo, vocês devem selecionar **com repetição** um subconjunto de exemplos de cada uma das classe para treiná-lo.

2. Variem o número de features consideradas no treinamento. Utilizando $\sqrt{m}$, $\frac{m}{2}$ e $\frac{3m}{4}$ atributos, em que m é o número total de atributos que vocês têm disponíveis.

3. Selecionem a *Random Forest* com o número ótimo de árvores e de atributos, e reportem a matriz de confusão relativa e a acurácia balanceada no conjunto de teste.

4. Reportem seus resultados e suas conclusões no relatório. Esses resultados foram melhores que os modelos treinados realizando o balanceamento *a priori*?

Se vocês optarem por fazer esta parte, o relatório pode conter até 7 páginas e a nota máxima passa a ser 12.

4 Arquivos

Os arquivos disponíveis no Moodle são:

- *covid_analysis_train_val_sets.csv*: conjunto de dados processados para serem utilizados como treinamento e validação;
- *covid_analysis_test_set.csv*: dados de teste serão **disponibilizados na quinta-feira antes do prazo final de submissão**;

5 Avaliação

O dataset foi previamente dividido aleatoriamente em dois conjuntos: treino+validação e teste. Relembrando que vocês devem fazer a divisão treino e validação neste trabalho.

Na quinta-feira anterior ao prazo final de submissão, iremos disponibilizar no Moodle o conjunto de teste e iremos avisá-los pelo canal da disciplina no Slack. No relatório, vocês devem reportar tudo que foi pedido na seção Tarefas.

A avaliação consistirá da análise do relatório e do código submetidos no Moodle. Iremos avaliar se as tarefas pedidas foram realizadas, como o treinamento e validação foram feitos, os resultados e as conclusões reportadas.

Observações sobre a avaliação:

- O trabalho poderá ser feito em duplas ou em trios, podendo haver repetição dos membros dos grupos a cada trabalho;
- O código e o relatório deverão ser submetidos no Moodle por **apenas um integrante do grupo**;
- Não se esqueçam de listar os nomes dos integrantes do grupo no início do relatório e no código;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;