INF-0615 – Aprendizado de Máquina Supervisionado I

Trabalho 3 - Árvores de Decisão e Florestas Aleatórias

Status de paciente com COVID-19

Data de Entrega: 26/09/2021

1 Descrição do Problema

A pandemia do vírus COVID-19 afetou diretamente e indiretamente todas as sociedades do mundo ao longo de 2020 e 2021. Muitas pessoas de diferentes nações, idades e classes sociais foram contaminadas, apresentando diferentes quadros clínicos de reação ao vírus. A doença, que iniciou-se na cidade Wuhan na China, em três meses, já havia se alastrado por grande parte dos países do globo, levando a Organização Mundial da Saúde a decretá-la como uma pandemia em 12/03/2020 [1].

Nesse trabalho, **vocês irão inferir o possível estado do paciente diagnosticado com o vírus COVID-19 dentre as três possíveis classes: em tratamento, falecido ou recuperado.** A base de dados contém os seguintes atributos:

- **Age**: Idade do paciente;
- **Sex**: Gênero do paciente;
- **Country**: País referente ao caso reportado;
- **Latitude**: Latitude da região onde o caso foi reportado;
- **Longitude**: Longitude da região onde o caso foi reportado;
- **Date Onset Symptoms**: Data em que o paciente começou a sentir os sintomas;
- **Date Admission on Hospital**: Data em que o paciente deu entrada no hospital;
- **Date Confirmation**: Data de confirmação da presença do COVID-19 no paciente;
- **Lives in Wuhan**: Valor binário. Informa se o paciente mora em Wuhan;
- **Travel History Dates**: Data aproximada de quando foi feita a viagem (não informa se esse atributo refere-se à data de partida ou chegada);
- **Travel History Location**: Informa se está disponível a informação do local da viagem;
- **Chronic Disease Binary**: Valor binário. Informa se o paciente é portador de doença crônica;
- **Date Death or Discharge**: Data em que o paciente deixou o hospital, tanto se ele veio a falecer ou recuperado;
- **Travel History Binary**: Valor binário. Indica se o paciente fez alguma viagem;
- **Label (target): Indica o estado do paciente no dia que foi feito o report sobre seu estado clínico. Os possíveis valores são: "OnTreatment", "dead"ou "recovered"**.

As datas foram todas transformadas em uma única sequência contínua de inteiros com referência em 01/01/2020, considerado como dia 0. Assim, por exemplo, dia 31/01/2020 é o dia 30, 01/02/2020 é o dia 31, 02/02/2020 é o dia 32 e assim sucessivamente. Dias anteriores a essa data de referência são negativos, assim dia 30/12/2019 é o dia -2 por exemplo. Dessa maneira, todas as data são valores contínuos.

Todos os atributos contínuos (datas e idade) tiveram seus valores NA's substituídos pelo valor médio do país considerado. Se todos os valores do país são NA, então toma-se a média global.

Os atributos categóricos também passaram por um processo similar de substituição de dados faltantes, no entanto tomando-se a moda ao invés da média. E, ao invés de tomar-se o valor global no caso de um país em que apresenta NA para a feature em todos seus exemplos, cria-se uma nova categoria informando a ausência do valor da feature. Por exemplo, para a feature *lives in Wuhan*, criou-se a categoria *not informed* para o caso em que não sabia-se se o paciente vivia ou não em Wuhan.

Este dataset provém originalmente de um conjunto de dados reunidos por diversos países no mundo e colocados à disposição para fins acadêmicos e de pesquisa.

2 Tarefas

Neste Trabalho, pedimos que vocês:

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

2. 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.

3. 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 também a matriz de confusão relativa e a acurácia balanceada no **conjunto de teste**.

4. Explorem pelo menos 2 possíveis subconjuntos de features (*feature selection*) para treinar uma árvore de decisão. Tomem o melhor modelo e reportem a matriz de confusão relativa e a acurácia balanceada do **no conjunto de teste**.

5. 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**.

6. 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.

Agora **vocês definem o conjunto de validação**. Tomem a base *train_val_set_patient_status_covid19.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 a base de treinamento e de validação estarão disjuntas** quando vocês desenvolverem seus modelos. Isso será também critério avaliativo.

Reparem que a Árvore de Decisão (e por consequência as Florestas Aleatórias) lidam naturalmente com problemas multi-classe, logo não é necessário aplicar o protocolo *Um-contra-Todos*.

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 no *ensemble* 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 três 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 disponível.
3. 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:

- *train_ val_ set_ patient_ status_ covid19*: conjunto de dados processados para serem utilizados como treinamento e validação;
- *test_ set_ patient_ status_ covid19*: dados de teste serão **disponibilizados 2 dias antes do prazo final de submissão**;
- *trabalho03_ nomes_ dos_ membros.r*: Código de apoio à leitura das bases. Vocês devem fazer o trabalho a partir dele. Coloquem no nome do arquivo o primeiro nome de cada membro e **no código o nome completo de cada membro do grupo**.

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 sexta-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 e na Opcional, se decidirem por fazê-la.

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 reportados 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;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;

6 Referências

1. *WHO announces COVID-19 outbreak a pandemic.* World Health Organization. `http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic`