INF-0615 – Aprendizado de Máquina Supervisionado I

Trabalho 2

Predição de Produção de Anticorpos para Desenvolvimento de Vacinas

Data de Entrega: 12/09/2021

1 Descrição do Dataset

O Sistema Imunológico humano apresenta células específicas para produção de anticorpos os quais são utilizados no combate aos invasores do organismo (antígenos). Essa produção de anticorpos é baseada na leitura da cadeia proteica do antígeno, ou seja, se o invasor apresenta uma determinada sequência de átomos em sua molécula de proteína, ela pode ativar a produção de anticorpos. Reconhecer se uma dada parte da molécula do invasor irá estimular a produção de anticorpos é de grande valia para a produção de vacinas contra o invasor.

No caso atual do COVID-19, seria interessante saber se alguma parte de sua molécula estimula a produção de anticorpos no organismo humano. Esta parte específica de sua cadeia proteica poderia ser utilizada na produção de vacinas, e assim o Sistema Imunológico já iniciaria previamente a produção de anticorpos aumentando sua capacidade de combate previamente à invasão do antígeno.

Neste trabalho, iremos disponibilizar uma base de dados com diversas informações sobre cadeias de proteínas para vocês predizerem se ela estimula ou não a produção de anticorpos no organismo humano. Assim, o *target* é binário, sendo 0 para representar que a cadeia proteica não estimula a produção de anticorpos, e 1 indicando a produção de anticorpos. Os atributos disponibilizados são listados abaixo:

- **Start_Position**
- **End_Position**
- **Chou_Fasman**
- **Emini**
- **Kolascar_Tongaonkar**
- **Parker**
- **Isoelectric_Point**
- **Aromaticity**
- **Hydrophobicity**
- **Stability**
- **Antibody Valence (target)**

Todos eles são atributos contínuos mensurando características sobre uma determinada cadeia proteica.

Após vocês desenvolverem seus modelos, iremos disponibilizar duas bases de teste: uma contendo sequências proteicas para vocês avaliarem o modelo, e outra contendo todo sequenciamento proteico do vírus SARS, um vírus da mesma família do COVID-19, identificado em 2003. Assim, vocês poderão mensurar a performance do modelo de vocês caso fosse usado na identificação de moléculas proteicas na produção de vacinas para o vírus SARS, e assim propondo um caminho para a pesquisa de vacinas para o COVID-19.

2 Tarefas

Pedimos que vocês:

1. Inspecionem os dados. Quantos exemplos vocês tem? Há exemplos com features sem anotações? Como vocês lidariam com isso?
2. Inspecionem a frequência de cada classe. A base de dados está balanceada ? Se não, como vocês lidarão com o desbalanceamento ?
3. Apliquem alguma técnica de normalização de forma a deixa os dados mais bem preparados para o treinamento.

4. Como *baseline*, treinem uma regressão logística com todas as features para predizer se haverá ou não a produção de anticorpos. Reportem a matriz de confusão relativa, o TPR, o TNR e a acurácia balanceada nas bases de treinamento, validação e teste (apenas arquivo *proteins_teste_set.csv*).

5. Implementem soluções alternativas baseadas em regressão logística através da combinação das features ou modelos polinomiais para melhorar o resultado do *baseline*. Comparem suas soluções reportando a matriz de confusão relativa e a acurácia balanceada no conjunto de validação. Tomem **apenas a melhor solução**, baseada na acurácia balanceada no **conjunto de validação**, e reportem a matriz de confusão relativa, TPR, TNR e acurácia balanceada no conjunto de teste (apenas arquivo *proteins_teste_set.csv*).

6. Tomem um dos modelos do item anterior e varie o fator de regularização (λ). Criem a curva viés/variância colocando os diferentes valores de λ no eixo das das abscissas. Identifiquem as regiões de *underfitting*, ponto ótimo e *overfitting*, e então tomem o modelo com o melhor fator de regularização e reporte a matriz de confusão relativa, o TPR, o TNR e a acurácia balanceada no conjunto de teste (apenas arquivo *proteins_teste_set.csv*). **Lembrem-se que a curva viés/variância é criada utilizando apenas os dados de treinamento e de validação!**.

7. Escrevam um relatório de no máximo 5 páginas:

   (a) Descrevam o que foi feito, bem como as diferenças entre o seu melhor modelo e o seu baseline;

   (b) Após desenvolverem todos os modelos, tomem o melhor modelo de todos (melhor performance no conjunto de validação). Reportem a matriz de confusão relativa e acurácia balanceada nos conjuntos de teste *proteins_teste_set.csv* e *SARS_test_csv*. Há uma diferença significativa entre eles ? Se sim, qual explicação você daria para essa diferença ?

   (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 Arquivos

Os arquivos disponíveis no Moodle são:

- *proteins_training_set.csv*: conjunto de dados para treinamento;
- *proteins_validation_set.csv*: conjunto de dados para validação;
- *proteins_test_set.csv* e *SARS_test_set.csv*: **(serão disponibilizados no sábado anterior ao prazo final da submissão)**: conjuntos de dados de teste retidos pelo professor;

4 Avaliação

O dataset foi previamente dividido em quatro conjuntos: treino, validação e teste, e um outro conjunto de teste adicional relacionado especificamente ao vírus SARS. Apenas os dois primeiros serão disponibilizados para que vocês implementem as suas soluções.

No sábado anterior ao prazo final de submissão, iremos disponibilizar no Moodle os conjuntos 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 reportados e as conclusões tomadas.

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 (arquivo .R) e o relatório (formato .pdf) 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;