**EXERCÍCIOS: PRÉ-PROCESSAMENTO DE DADOS, REDUÇÃO DE DIMENSIONALIDADE E APRENDIZADO SUPERVISIONADO**

|  |  |  |
| --- | --- | --- |
| **João Lucas**  *Graduando*  [*joaolucasc@ucl.br*](mailto:joaolucasc@ucl.br) | **Cleres Caetano**  *Graduando*  [*cleresjunior@ucl.br*](mailto:cleresjunior@ucl.br) |  |

|  |  |  |
| --- | --- | --- |
| **Nathan Andrade**  *Graduando*  [*nathanag@ucl.br*](mailto:nathanag@ucl.br) | **Davi Ferlin**  *Graduando*  [*daviferlin@ucl.br*](mailto:daviferlin@ucl.br) |  |

**4. Explique o dilema entre bias e variância e o seu relacionamento com underfitting e overfitting.**

O dilema entre bias e variância se resume em uma tentativa de equilibrar a simplicidade e a complexidade de um algoritmo. O algoritmo com alto bias é simplista, o que leva ao underfitting, que é a não captação dos padrões que foram estimulados durante o treinamento. Já o algoritmo com baixo bias tem um alto grau de variância, consequentemente, ocorre quando o modelo é muito complexo, se ajustando minuciosamente aos dados de treino. O excesso de variância gera o overfitting, que seria a coleta desses dados em exagero e, consequentemente, o algoritmo perde a capacidade de "generalizar" para novos dados.

Dessa forma, o equilíbrio entre bias e variância importa, visto que o algoritmo deve ser flexível para captar padrões reais sem se apegar a detalhes específicos. Essa busca entre reduzir bias e controlar a variância orienta a construção e o ajuste do processo, tentando evitar tanto o underfitting quanto o overfitting.

**5. Comente sobre a veracidade das afirmações:**

**a) “Quanto mais variáveis de entrada forem usadas em um modelo de aprendizado de**

**máquina, melhor será a qualidade do modelo”.**

Nem sempre, o uso de mais variáveis de entrada vai resultar em algo bom para o modelo, pois essa adição de variáveis pode ocasionar o overfitting. A técnica de machine learning sugere o uso de variáveis relevantes para evitar que o modelo se confunda com dados irrelevantes e perca a capacidade de generalizar. A "grosso modo", se entregar muitas variáveis ao algoritmo, o modelo fica sobrecarregado de informações e ao avaliar os dados, vai encontrar diversos padrões em um objeto, considerando objetos que não são relevantes.

**b) “Independente da qualidade, quanto mais amostras forem obtidas para uma base de**

**dados, maior a tendência de se obter modelos mais adequados”.**

Essa afirmação está errada. Afirmando que a frase "(...) quanto mais amostras forem obtidas para uma base de dados, maior a tendência de se obter modelos mais adequados. " está correta, pois as amostras de qualidade maior ocasionam no melhor desenvolvimento da análise do modelo, capturando padrões e reduzindo variâncias. Porém, se a qualidade dos dados for baixa ou com muitos erros, o modelo pode aprender informações incorretas.

**c) “Às vezes com simples manipulações na base de dados (limpeza, conversão de**

**valores, etc.) pode-se conseguir melhoras significativas nos resultados, sem fazer**

**nenhuma alteração na técnica de aprendizado de máquina usada”.**

Essa afirmação é verdadeira, pois "limpar" a base de dados, como limpeza de outliers, remoção de valores incorretos, facilita a leitura pelo modelo, gerando resultados mais precisos. Dessa forma, a base da dados devem estar "transparente", podendo ser mais eficaz, essa boa prática, do que modificar o algoritmo.

**6. Em uma empresa é adotado um método de Aprendizado de Máquina para detectar defeito de fabricação de peças mecânicas, sendo que raramente acontece este tipo de problema na fábrica. Um funcionário anuncia empolgado que o sistema alcançou uma acurácia de 99%,**

**porém seu gerente não achou o resultado tão relevante. Responda:**

**a) Por que o gerente não ficou empolgado com o resultado achado?]**

O gerente não ficou empolgado, pois a acurácia de 99% não necessariamente, significa que a leitura foi correta. A acurácia pode ser enganosa, se um modelo que classifique quase todas as peças como "sem defeito" ainda pode atingir uma acurácia alta, mesmo que não seja detectado os casos com defeito. O modelo pode estar falhando na identificação dos defeitos e passando uma sensação de que são poucas peças com falha**.**

**b) O que o funcionário poderia fazer para confirmar se o método empregado é**

**adequado para o problema?**

O funcionário poderia usar algumas técnicas para verificar a veracidade da acurácia, dentre elas a precisão, que indica a quantidade de peças que foram identificadas como defeituosas entre as que o modelo identificou como defeituosas. A seguir, a técnica de recall, que mostra a quantidade de peças realmente defeituosas, que foram corretamente identificadas pelo algoritmo. A matriz de confusão, irá auxiliar juntamente a esses métodos, oferecendo uma visão mais detalhada da capacidade do modelo de detectar as peças defeituosas, confirmando se o método utilizado está adequado.

**7. Como pode ser usada uma árvore (de regressão ou de decisão) para avaliar uma amostra**

**quando ela possui uma ou mais variáveis faltantes?**

Quando uma árvore de decisão ou regressão encontra uma amostra com dados faltantes, ainda assim é possível avaliar essa amostra com o uso de ferramentas que lidam com essas lacunas de dados de forma prática e eficiente. Por exemplo, a imputação que é a substituição dos valores faltantes por uma média estatística daquele dado. A divisão por probabilidade também é uma técnica interessante para esses casos, pois quando temos uma variável ausente, a árvore segue por todos os ramos possíveis, usando probabilidade, combina os resultados dos caminhos.

Assim é possível lidar com dados incompletos e seguir o propósito do algoritmo.

**REFERÊNCIAS**

1. Árvore de decisão: entenda esse algoritmo de Machine Learning - [Árvore de decisão](https://blog.somostera.com/data-science/arvores-de-decisao#:~:text=Uma%20decision%20tree%20pode%20ser,importantes%20e%20ajusta%20as%20condi%C3%A7%C3%B5es.)
2. O que é acurácia em machine learning - [Acurácia em machine learning](https://mariofilho.com/o-que-e-acuracia-em-machine-learning/#:~:text=Ela%20%C3%A9%20calculada%20dividindo%20o,sua%20acur%C3%A1cia%20%C3%A9%20de%2085%25.)
3. Bias, variância, underfitting e overfitting - [Bias, variância, underfitting e overfitting](https://juliocprocha.wordpress.com/2017/04/01/bias-e-varianciaunderfitting-e-overfitting/)
4. Aprendizado supervisionado x não supervisionado - [Aprendizado supervisionado x não supervisionado](https://medium.com/data-hackers/aprendizado-supervisionado-x-n%C3%A3o-supervisionado-ca79b522659d)