**Overfitting versus Underfitting**

**Uma conversa com o professor de GDE**

- Oi Fernando, tudo bem?

-Oi João, tudo ótimo e você?

-Tudo bem, também. Você está se preparando para sua próxima aula de GDE?

-Não! Estava aqui lendo uma reportagem sobre a comparação entre modelos para um cenário de aplicação de Machine Learning. Sabe como é, gosto de estar atualizado nessas informações... João, aproveitando que você está aqui, teria como você tira uma dúvida? Nesse texto aqui, o autor utiliza muito os termos de Overfitting e Underfiiting só que eu não estou entendendo direito o que eles querem dizer, teria como você me explicar?

-Claro Fernando, são termos muito importantes nessa área mesmo. Bem, seguinte, você sabe que Machine Learning é uma ferramenta que nós, seres humanos, tentamos implementar em computadores para que eles sejam capazes de prever acontecimentos futuro. Mais ainda, Machine Learning trabalha sem nenhum auxílio e regras explícitas, trabalha apenas com modelos estatísticos e inferências e a utilização de banco de dados construídos por dados ao longo da história. Dessa forma, Para desempenhar essa função, como eu disse, é necessário um modelo estatístico. Existem diversos modelos, Regressão Linear, Árovre de Decisão, Random Forrest e etc, para serem aplicados em diferentes tipos de cenários. Porém, como escolher o modelo ideal para o cenário que estou vendo?

Existem diversas medidas e uma delas é, por exemplo, o RMSE, o erro médio quadrático. Bem, como você já deve saber, no processo de aplicação de um modelo, é realizado uma separação no banco de dados entre treinamento e teste. Essa separação serve para treinarmos o modelo com parte do banco de dados, para que ele aprenda como deve desempenhar no cenário introduzido. É essa parte que molda todo o modelo, incluindo as features necessárias para descobrir o que o modelo precisa e mexermos também nos parâmetros que o modelo precisa para prever o que ele tem que prever. Depois de treinado, aplica-se o modelo no banco de dados teste. Em ambos os casos, pode-se usar o RMSE, para ver se o valor predito pelo modelo, se iguala ao valor real que aconteceu no passado. Nesses casos, os cientistas de dados esperam sempre ter um RMSE baixo, já que um erro baixo implica que o modelo acertou a previsão, portanto é um bom modelo.

Porém, e se o RMSE não for baixo em nenhum dos casos, no treinamento e teste? Bem ai entramos em um cenário de Underfitting, ou seja, o modelo é simples demais, ou seja, ele não desempenha bem o seu papel de prever dados a partir do que foi lhe concedido. É provável que em um cenário de Underfitting, o modelo possua um erro de treinamento alto, mas o de teste ainda mais alto. As causas disso devem se dar á um mal ajuste dos parâmetros do modelo e uma escolha precária das features que o modelo precisava para prever o futuro.

Já o Overfitting, não é exatamente o que parece, já que se o Under é ruim, o Over deve ser bom ne? Só que na realidade, o Overfitting é um cenário que o erro de treinamento é minúsculo, mas o de teste é altíssimo. Isso acontece porque o modelo é complexo demais e foi pensado para responder exclusivamente o banco de dados de treinamento e não qualquer banco.

Um exemplo prático desses conceitos é o seguinte. Imagine que eu jogue em duas sessões (treinamento e teste) bolinhas azuis e vermelhas misturadas. Agora imagine que eu pegue uma corda (nosso modelo) que agrupe o máximo de bolinhas vermelhas, primeiro no treinamento. Infelizmente algumas vermelhas ficam para fora da corda e algumas azuis para dentro, mas no final, eu consegui agrupar muitas, tendo uma precisão alta, ou seja, um erro baixo. Agora imagine que eu pegue essa corda e coloque na sessão de teste, na exata mesma área que o de treinamento. Você concorda que agora, o erro provavelmente vai ser bem maior, porque pensamos apenas para o treinamento, e deixamos a corda muito complexa, com várias curvinhas para deixar fora o que tinha que fica fora e dentro o que tinha que ficar dentro. Isso, é o Overfitting. O Under, você consegue imaginar que é basicamente colocar a corda de um jeito pouco cuidadoso que não vai ficar bom nem no teste ou no treinamento. Deu pra entender?

-Deu, brigado João!