Tradeoff Viés Variância

Olá, seja bem-vindo!

Ao utilizar um modelo de Machine Learning, é fundamental observar se as previsões realizadas correspondem aos resultados esperados para o contexto do problema que está sendo resolvido, e isso tem muito a ver com os conceitos de Underfitting e Overfitting, que você já deve estar familiarizado, não é mesmo?

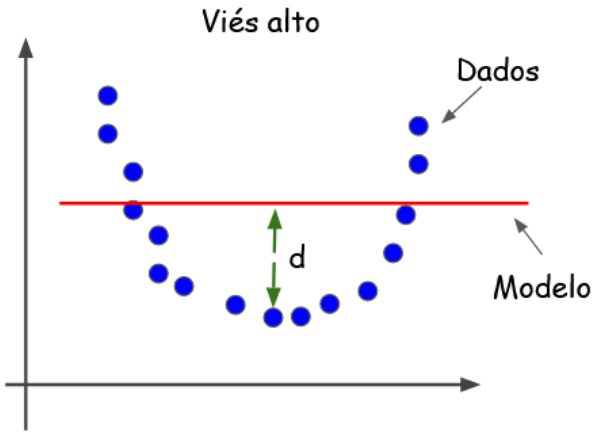
Pois bem, caso não esteja recordando essas duas categorias, aqui vai uma breve contextualização: o Underfitting ocorre quando o modelo é muito simples para resolver o problema proposto e não consegue aprender o suficiente, apresentando erros tanto na fase de treino quanto na fase de testes. Já o Overfitting acontece quando o modelo é mais complexo do que o problema exige, e acaba conhecendo tudo sobre os dados de treino, sem generalizar corretamente.

Porém, o problema enfrentado por quem trabalha com Machine Learing é justamente a dificuldade em encontrar um equilíbrio para que os conceitos de Underfitting e Overfitting não aconteçam. Mas calma, não é preciso se desesperar. Neste curso, você aprenderá sobre o Tradeoff Viés/Variância, que está relacionado aos problemas de previsões mencionados.

Para que você entenda melhor essa ideia, vou lhe lembrar uma expressão bastante utilizada no dia a dia e que representa bem esse conceito, “perder de um lado para ganhar de outro”. Agora, deve haver um questionamento na sua mente, “certo, mas e o que isso significa?” É simples! De maneira geral, o Tradeoff é uma escolha que pode causar um custo, mas que consegue trazer um benefício por outro lado.

Então, até aqui tudo entendido? Ótimo! Contudo, antes de estudar o Tradeoff, é importante que você compreenda as subcategorias Viés e a Variância: a primeira é um tipo de erro que representa a diferença entre os valores presentes na base de dados, assim como os valores que foram previstos pelo modelo. Então, quando o Viés é elevado, o modelo não irá aprender corretamente, fazendo previsões incorretas. Entretanto, quando o Viés é baixo, os desvios entre as observações e as previsões diminuem, fazendo com que o modelo seja mais preciso.

Para que fique mais fácil a compreensão, acompanhe um modelo, que foi escolhido para fazer previsões sobre um conjunto de dados, exibido no seguinte gráfico:

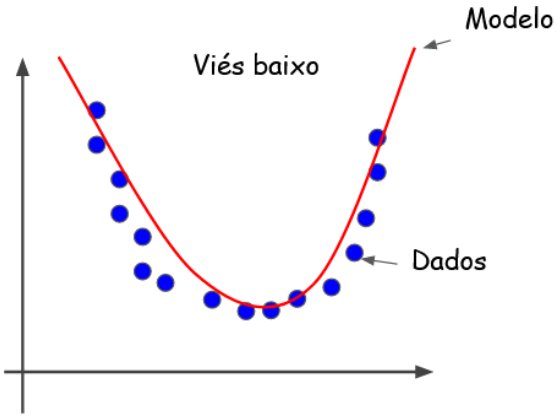


Esse gráfico representa o Viés Alto. Nele, há uma reta vermelha, que representa o modelo escolhido inicialmente para tentar modelar os dados. Acima e abaixo da reta, existem vários pontos azuis espalhados, que representam o conjunto de dados, ou observações. Abaixo da reta, há uma seta bidirecional, que aponta para os pontos azuis e para a reta, indicando a distância entre o conjunto de dados e a reta.

Esse modelo não é adequado para representar esses pontos, justamente pela grande distância entre a reta e alguns pontos dos dados, ou seja, quanto maior essa distância, menor é a semelhança entre o modelo e os dados. Assim, essa grande diferença significa que o modelo escolhido possui um Viés Alto. Por isso, preste muita atenção!

Outro ponto a ser ressaltado, é que esse modelo tem a forma de uma reta que pode ser dada por uma equação do tipo y = ax +b, e tal equação não existe nenhum termo ao quadrado, que seria típico de uma curva que representaria melhor esses dados. Desse modo, o modelo não consegue aprender a identificar o padrão da forma dos dados. Portanto, quando isso acontece, você tem o Underfitting. Logo, é possível concluir que: quando houver um Viés Alto, acontece o Underfitting no modelo.

Porém, se você usar um outro modelo um pouco mais complexo, ou seja, com uma equação mais sofisticada, do tipo y = x2, que representa uma parábola, as distâncias entre os pontos e a curva do modelo diminuem, como está representado neste outro gráfico, que caracteriza um Viés Menor, não ocorrendo o Underfitting. Além de ter uma parábola formada por uma curva vermelha em forma de “U”, que representa o modelo. Ademais, existe também vários pontos azuis próximos à parábola, isso significa que o modelo é mais adequado para esses dados, pois a distância entre a parábola e os pontos é menor que no gráfico apresentado anteriormente.



Por outro lado, se o viés for muito baixo, o modelo representará tão bem os dados de treino que ocorrerá outro tipo de problema, que está associado à sensibilidade do modelo, quando ele for prever um novo valor. Neste ponto, você precisará compreender melhor a definição de Variância, a segunda subcategoria mencionada anteriormente.

Bom, a Variância é uma medida de dispersão, ou seja, ela tenta explicar o grau de variação dos valores em relação a sua média, e mostra o quão distante cada valor está do produto central em um conjunto de dados. Assim, quando o modelo "memorizar" toda a base de dados usada para treino ao ser submetido a novos dados na etapa de testes, e como ele não estará preparado para dados muito diferentes dos que ele conhece, as distâncias dos pontos em relação à curva do modelo aumentará, ocasionado uma alta variabilidade ou alta variância.

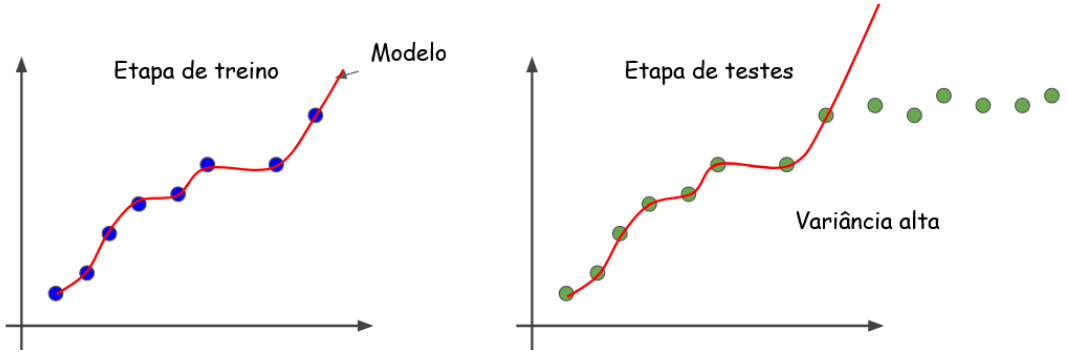
Portanto, quando o modelo se especializa nos dados de treino, ou seja, quando ele consegue descrever exatamente como os dados se comportam, mas não é capaz de fazer generalizações, ocorre o Overfitting. Todavia, o que de fato são essas generalizações? Simples, elas não passam de um jargão, utilizado em Machine Learning, para se referir ao modelo que não consegue prever corretamente quando surgem novos dados que ele não conhece. Então, com essa definição de Variância, você pode concluir que modelos com Variância Alta apresentam Overfitting.

Agora, para exemplificar, acompanhe os dois gráficos seguintes: o primeiro representa a etapa de treino, já o segundo a etapa de teste com uma Alta Variância, onde a reta não acompanha a realidade dos dados a partir de um determinado ponto.

À esquerda, o gráfico representa a "Etapa de treino". Nele, há uma curva ondulada vermelha, que indica o modelo. Sobre essa curva, há vários pontos azuis enfileirados, ilustrando os dados.

À direita, o esquema representa a "Etapa de testes", em que há uma curva ondulada vermelha, indicando o modelo; e pontos verdes, que ilustram os dados. Ademais, há alguns

pontos enfileirados sobre a curva e vários outros pontos que estão fora da curva e mais distantes dela, sinalizando a Variância Alta.

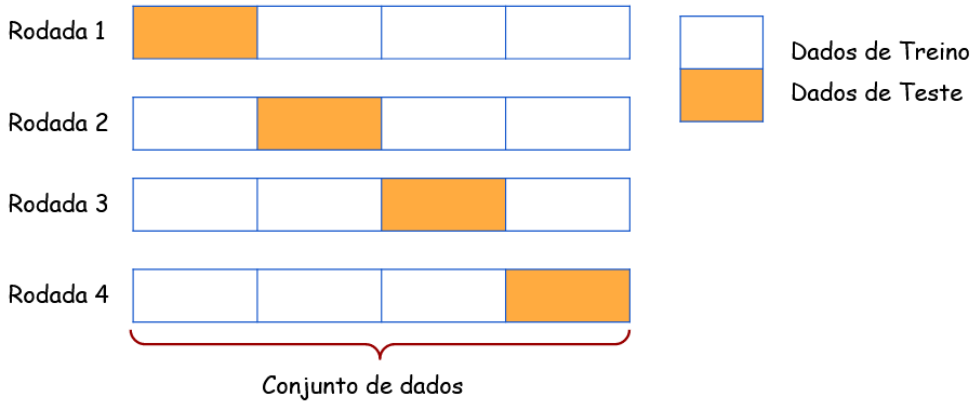


Até aqui, você pôde aprender como os conceitos Viés e Variância afetam o comportamento dos modelos, percebendo que seria ideal um modelo que conseguisse descrever bem os dados de treino e que fosse capaz de fazer uma boa generalização para dados futuros. Portanto, esse ponto, onde se tem que ponderar na escolha de um modelo baseado no Viés e na Variância, chama-se Tradeoff Viés/Variância, também conhecido de Dilema Viés/Variância.

Portanto, essa escolha se torna difícil, pois se você diminuir a complexidade do modelo, pode ocorrer Underfitting; e se ela for aumentada muito, pode ocorrer Overfitting. Por esses motivos, existem algumas técnicas para tentar melhorar esse Tradeoff.

Uma delas é a redução da dimensionalidade dos dados, que se constitui na tentativa de descrever os dados através da maioria dos atributos, mas não todos, ou seja, elimina-se apenas os atributos que têm pouca, ou nenhuma, relação com o atributo de saída, além de reduzir a quantidade de atributos desnecessários.

Outra técnica utilizada é a validação cruzada. Nela, o conjunto de dados é submetido ao treino e ao teste várias vezes. Porém, usando uma parte diferente dos dados para cada rodada de treino e teste. Desse modo, como geralmente utiliza-se cerca de 20% a 30% dos dados para treino, e o restante para teste, suponha que, na primeira rodada, seria usado a primeira parte dos dados como 25% para teste, e os 75% restante para treino. Na segunda rodada, seria utilizado a segunda parte como 25% de testes, e o restante seria usado para treino, e assim sucessivamente. Lembrando que esse procedimento pode ser repetido por um número determinado de rodadas.



Muita coisa não é mesmo?

Neste curso, você pôde entender o que é o Tradeoff Viés/Variância e como deve ponderar na escolha de modelos para que não ocorra tanto Underfitting quanto Overfitting. Além disso, foram conhecidas também as técnicas de redução de dimensionalidade dos dados, assim como a validação cruzada, que são as mais usadas para tentar evitar esses problemas. Por isso, não esqueça de resolver os exercícios propostos e praticar o que você aprendeu para consolidar sua prática em construir modelos cada vez mais completos de Machine Learning.

Bons estudos e até mais!