PROCESSO DE DESENVOLVIMENTO DOCUMENTAÇÃO DE PROJETO		
Nome do curso: *Introdução à Machine Learning*	Overfitting e Underfitting	Responsável: Paulo Jarbas Camurça

Overfitting e Underfitting

Olá, seja bem-vindo!

Durante esse curso, você aprendeu que os modelos de Machine Learning são utilizados para encontrar padrões nos dados e que a partir desses elementos, é possível elaborar novas previsões. Você também descobriu que a qualidade das previsões de um modelo, depende de quão bem ele pode generalizar as suas respostas, à medida que ele é submetido a dados nunca vistos na etapa de treino.

Nesta aula, você desvendará dois tipos de problemas de generalização que pioram a qualidade das previsões dos modelos de Machine Learning: o **overfitting**, e **underfitting.** Você irá aprender quais medidas devem ser tomadas para evitar que esses problemas ocorram. Ficou curioso? Vamos lá!

Bons estudos!

Em primeiro lugar, para que você entenda melhor a ideia da generalização que um modelo pode ter, imagine que você é motorista de uma empresa e recebe a tarefa de ir buscar um novo funcionário no aeroporto da cidade. Como referência, você vê duas fotos da pessoa que deve encontrar e terá que reconhecê-la em meio a um conjunto de pessoas que transitam pelo local.

O novo funcionário pode ter uma aparência mais velha, ou estar vestido com roupas diferentes daquelas mostradas inicialmente nas fotos, dificultando o

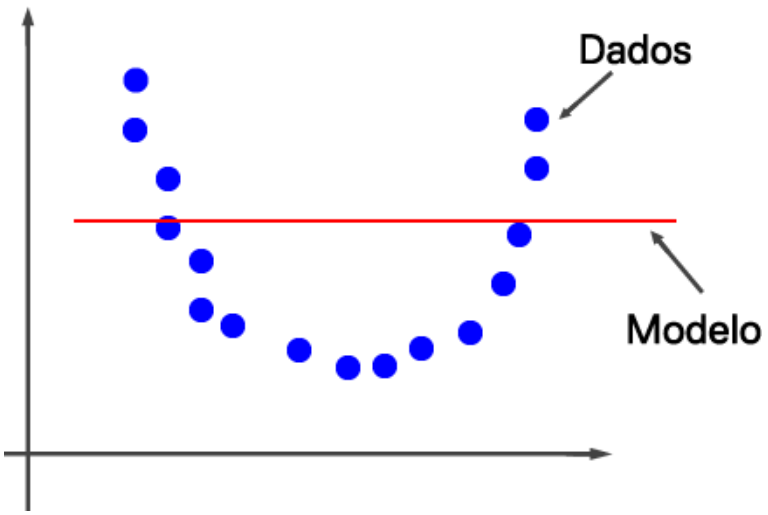
reconhecimento, não é mesmo? Deste modo, sua capacidade de reconhecer essa pessoa, mostra o quanto você consegue generalizar uma imagem.

Agora que você já entendeu que essa ideia acontece de forma semelhante nos modelos de Machine Learning, pois um bom modelo deve ser capaz de generalizar suas previsões, quando submetido a novos dados, você está pronto para conhecer a definição do conceito de **Underfitting**.

Underfitting, também chamado de sub-ajuste, é um problema que ocorre quando o modelo tende a subestimar os resultados, ou seja, o modelo escolhido para resolver o problema é muito simples, no entanto, o problema tem uma complexidade maior.

Como exemplo, idealize um modelo que possua como característica uma reta. Lembre-se que para um conjunto de dados, você precisa escolher um modelo que melhor se adeque a esses dados, certo?

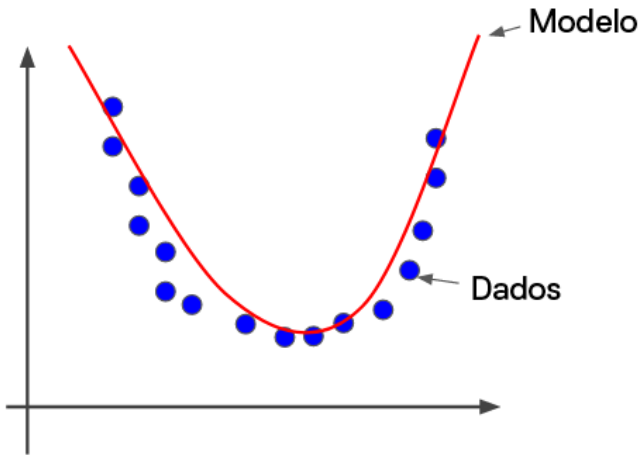
Agora, observe a representação dos dados em um gráfico que possui um eixo vertical e um eixo horizontal. Nele, há uma reta vermelha, que representa o modelo. Acima e abaixo da reta, há vários pontos azuis espalhados, que representam os dados e a distribuição dos pontos azuis se assemelha a forma da letra “U”.



Você percebeu que o modelo, no entanto, não se ajustou corretamente aos dados, caracterizando um underfitting, pois o modelo usado é muito simples e não captura os padrões existentes nos dados, que tem uma forma de “U”.

Outro aspecto que pode causar underfitting é quando se usam poucos dados na etapa de treinamento, fazendo com que o modelo não aprenda o suficiente. Ou seja, usando um modelo que possua uma forma mais complexa, como o de uma parábola, ele conseguirá se ajustar melhor a esses dados e, assim, não apresentará underfitting.

Comprovando que quanto mais dados utilizados na etapa de treinamento, menor a ocorrência de underfitting. Repare os dados representados em um gráfico, formado por um eixo vertical e um eixo horizontal. Nele, há uma curva vermelha em forma de “u”, que representa o modelo. Há vários pontos azuis próximos à curva, alguns acima dela e a maioria abaixo. A distribuição dos pontos azuis também tem a forma da letra “u”,

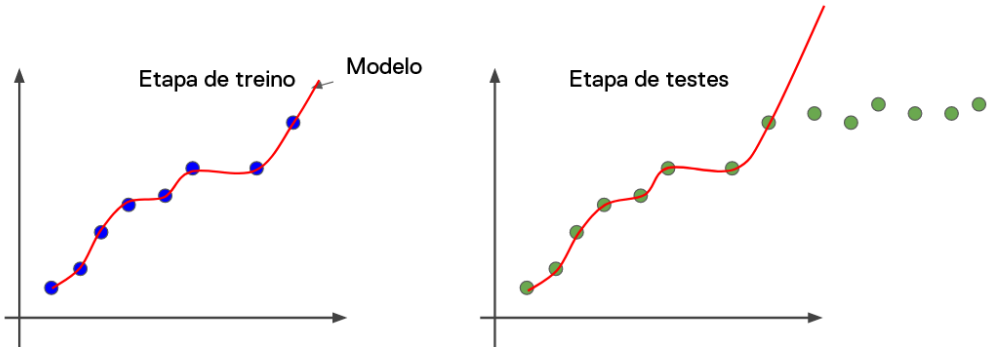


Já o **Overfitting**, também chamado de sobre-ajuste, ocorre quando o modelo consegue se adaptar muito bem aos dados de treinamento, porém, na etapa de testes, na presença de novos dados, o modelo não consegue generalizar e pode apresentar resultados incorretos.

Observe nos gráficos o exemplo de overfitting, em que a curva do modelo consegue representar, fielmente, os dados que foram usados para treinar o algoritmo.

Existem duas representações de dados: À esquerda, há um gráfico formado por um eixo vertical e um eixo horizontal intitulado “Etapa de treino”. Nele, há uma curva ondulada vermelha, que representa o modelo. Sobre essa curva, há vários pontos azuis enfileirados, que representam os dados.

À direita, há um gráfico formado por um eixo vertical e um eixo horizontal intitulado “Etapa de testes”. Nele, existe uma curva ondulada vermelha, que representa o modelo. Sobre ela há alguns pontos verdes enfileirados, que representam os dados, e vários outros pontos que estão fora da curva e mais distantes dela.



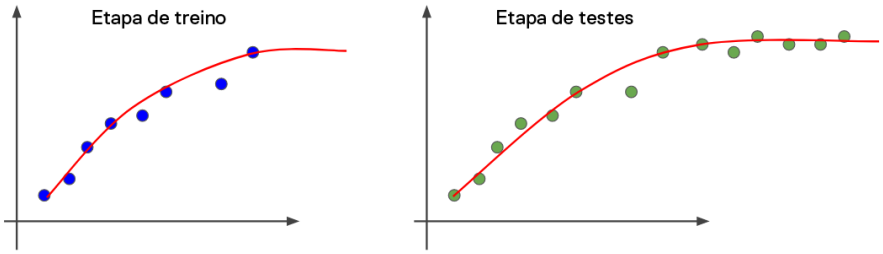
Assim, o modelo consegue “memorizar” todo o conjunto de dados de treino, mas não consegue aprender como os dados se comportam, e quando for submetido aos dados de testes, ele não conseguirá prever corretamente.

Isso geralmente ocorre quando o modelo usado é muito sofisticado para resolver um problema simples, ou quando é usada toda a base de dados para o treinamento. Como solução para este problema, pode-se utilizar um modelo mais simples para representar os dados, como indica os seguintes gráficos:

À esquerda, há um gráfico formado por um eixo vertical e um eixo horizontal intitulado “Etapa de treino”. Nele, há uma curva vermelha, que representa o modelo.

Sobre essa curva, há alguns pontos azuis enfileirados, que representam os dados, e dois pontos que estão fora da curva, mas ainda assim próximos dela.

À direita, há um gráfico formado por um eixo vertical e um eixo horizontal intitulado “Etapa de testes”. Nele, existe uma curva vermelha, que representa o modelo. Sobre ela há vários pontos verdes enfileirados, que representam os dados, e apenas três pontos que estão fora da curva, mas ainda assim próximos dela.



Nesse caso, embora o modelo não represente fielmente todos os pontos na etapa de treino, ele consegue reproduzir o padrão dos dados, fazendo previsões corretas na etapa de testes.

Bem, até aqui você aprendeu como os modelos de Machine Learning podem apresentar problemas que surgem de acordo com a abordagem escolhida para solucionar um obstáculo, ou pela forma que o modelo é treinado, o underfitting e overfitting. Você também estudou quais estratégias podem ser utilizadas para tentar evitar esse tipo de problema.

Para um melhor aprendizado, é essencial que você coloque em prática os comandos que foram abordados nessa aula e resolva os exercícios.

Bons estudos e até mais!