Regressões

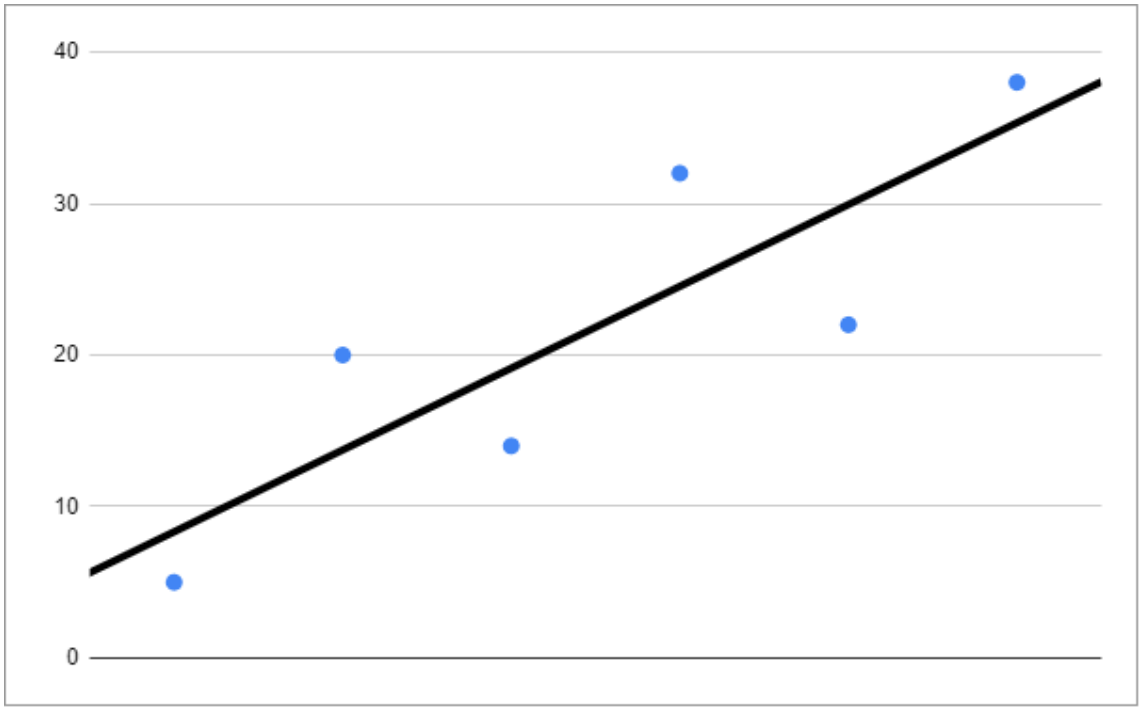
Introdução

Uma das principais formas de aprendizado de máquina supervisionado é por meio da formação de equações lineares, que aproximam de forma satisfatória a predição de um resultado y a partir de um ou mais inputs (x1, x2, x3), esse tipo de modelo, é comumente chamado de modelo de regressão linear, e é um dos modelos mais comuns no aprendizado de máquina.

Vamos entrar mais a fundo neste assunto, começando com o que é regressão. Em matemática, as regressões são utilizadas principalmente para encontrar relacionamentos entre variáveis aparentemente distintas. Imagine o cenário onde diversos pacientes com diabetes foram estudados e os médicos responsáveis por este estudo anotaram diversas características como idade, peso, massa muscular, pressão arterial média, entre outros, utilizando um modelo de regressão linear, podemos inferir determinadas características patológicas de cada um dos pacientes a partir destes dados.

Geralmente, na maioria das análises de regressão, estabelecemos um fenômeno de interesse e um número de observações (que correspondem no exemplo anterior, ao tipo de diabetes e os dados de saúde dos pacientes, respectivamente), e então, utilizamos os mecanismos de aprendizado de máquina para descobrir uma função que realiza o mapeamento entre as variáveis independentes, ou inputs (x, ou os dados médicos dos pacientes) e as variáveis dependentes, ou outputs (tipo de diabete)

A imagem abaixo exemplifica visualmente um modelo de regressão linear, que com base em uma dispersão de dados, consegue traçar uma possível interação entre as variáveis independentes e o resultado final.



Coeficiente de determinação

Na imagem anterior, temos uma linha central que representa o nosso modelo, e tenta traçar uma tendência central de acordo com os dados obtidos das nossas observações, mas como podemos ter certeza de que este modelo é satisfatório? Para isso temos o coeficiente de determinação, que nada mais é do que um valor entre zero e um, no qual zero indica que o nosso modelo não representa os

dados observados, e um, que indica que o modelo em questão representa perfeitamente os dados obtidos.

É importante ressaltar que os modelos de regressão na maioria das vezes não trabalham buscando valores de R^2 muito próximos de 1, já que isso pode indicar que o modelo é capaz de predizer apenas as condições específicas daquele experimento.

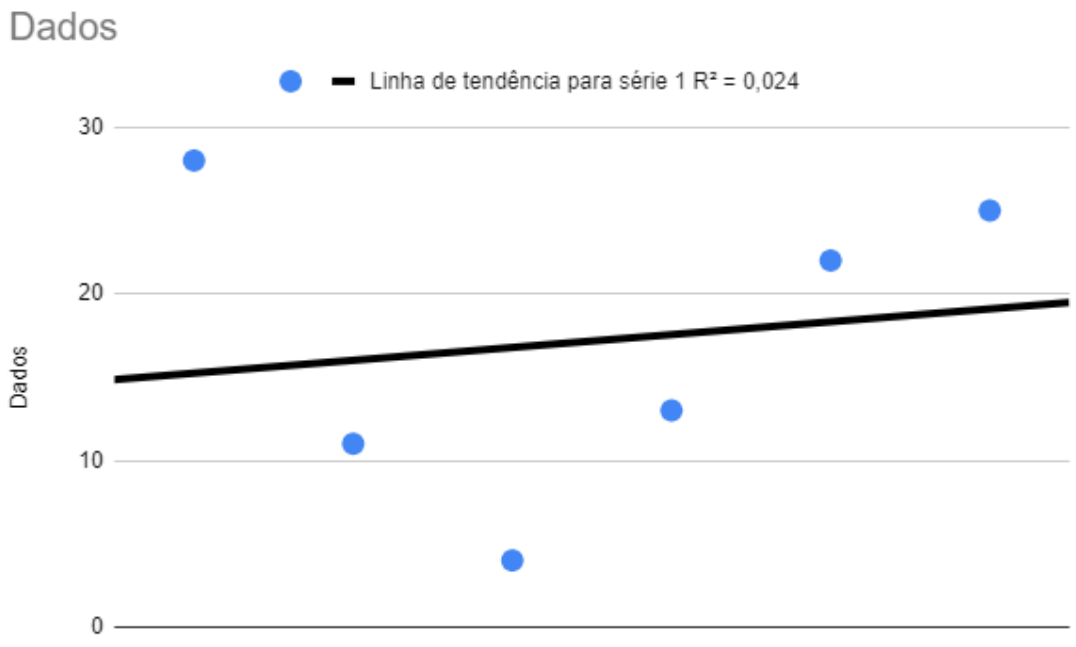
Outros tipos

Além das regressões lineares, existem outros tipos de regressões matemáticas, como as polinomiais, que não reproduzem graficamente uma linha reta, mas sim uma curva complexa, e é especialmente útil em modelos nos quais o espalhamento de dados das observações são muito distantes quando plotados em um gráfico.

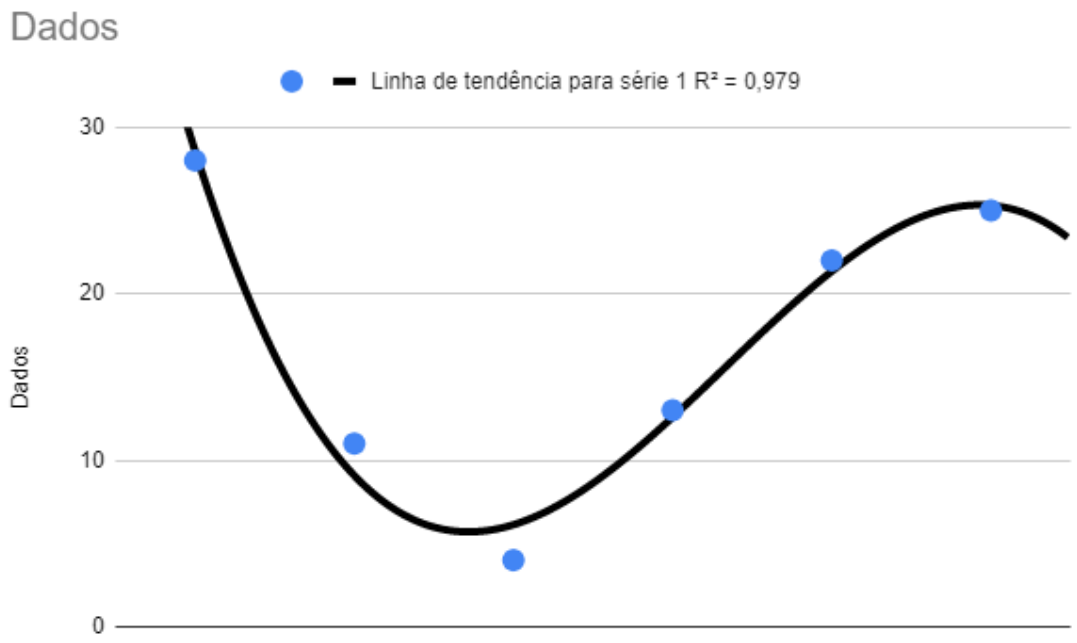
Outro problema mais comum decorrente de modelos formados a partir de regressões polinomiais são os casos de underfitting e overfitting. O primeiro acontece quando o coeficiente de determinação fica muito próximo de zero, enquanto que o segundo ocorre quando o coeficiente de determinação fica muito próximo de um.

Em um primeiro momento os casos de overfitting não aparentam ser problemáticos, afinal eles indicam que o nosso modelo aprendeu corretamente os padrões e foi capaz de chegar a uma função que bem representa as observações, no entanto, é importante ter em mente que para a análise de dados, um modelo que trabalha de maneira muito precisa, geralmente é ruim em realizar suposições sobre coleções de dados não presentes no data set de teste.

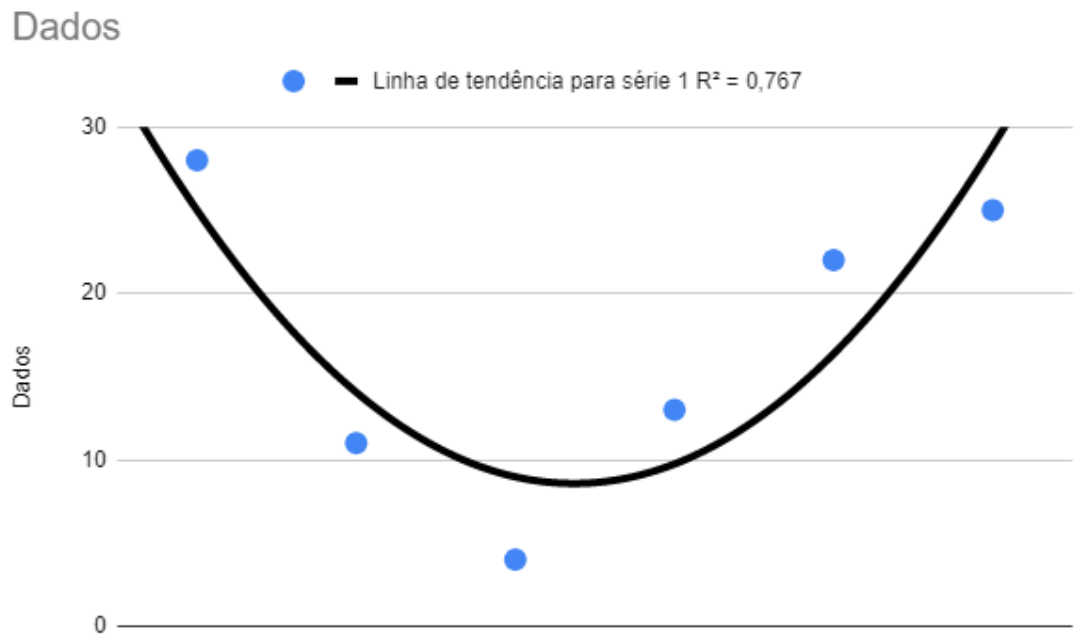
Repare nas imagens abaixo, observe que quanto maior o valor de R^2, mais próxima a curva se aproxima do resultado final.



R = 0.024



R = 0.979



R = 0.757

Esperamos que tenha gostado dessa explicação, lembrando que é importante conhecer a teoria primeiro antes de seguir com a prática, o que será muito mais simples graças as excelentes bibliotecas disponíveis na linguagem python.