06

A regressão por debaixo dos panos e medidas de erro

Transcrição

[00:00] O que nós precisamos ver agora é entender como que a regressão funciona debaixo dos panos. Voltando lá para o que eu falei no começo das aulas, vou atualizar aqui o GeoGebra de novo, pra nós limparmos. A função, o principal objetivo da regressão linear é encontrar a equação da reta, que é essa daqui.

[00:28] Vamos escrever essa equação. Então eu vou ter um, 10x + 7. Vamos diminuir um pouco esse valor 2x. E aqui nós jogamos um -7. E o que acontece? Se nós trocarmos esse 2x, pra nós entendermos o que está acontecendo, é como se esse fosse o nosso M, que eu mostrei, e esse daqui fosse o nosso B, que eu também mostrei. Na verdade é o -7, que é o B.

[01:05] Vamos trocar esse 2 por 1, 3, 4. Está muito grande, vamos deixar um pouco mais suave. 1.9, .8, .7, 1, 1.1, .2, .3, .4, .5, .6, .7, .8, .9, ou seja, o que nós conseguimos ver? Que eu mudar esse valor, eu estou mudando a reta nesse sentido, eu estou mudando o ângulo dela. Então eu vou deixar aqui no 2 e vou fazer a mesma coisa com esse processo. Tem o -7, o -6.9, .8, .7, .6, .5, .4, .3, .2, .1, ou seja, quando eu mudo o meu B, eu estou mudando assim, eu estou mudando nesse sentido, a posição da reta em relação ao eixo.

[02:05] Na prática o que acontece? Será que o nosso modelo encontrou isso pra nós? Nós conseguimos encontrar esses valores pelo .coef, de coeficiente, porque aqui o nosso M nada mais é do que o coeficiente angular, e o B é o ponto que intercepta os eixos. Então se eu tenho aqui o modelo.coef, esse daqui nada mais é do que o meu alfa. E se eu tenho o modelo.intercept, esse daqui é meu beta.

[02:36] Logo a equação da reta, vamos chamar isso daqui de M, ele recebe modelo.coef e o meu B recebe o modelo.intercept. Então a nossa equação da reta é alguma coisa assim.

[02:56] Logo, se eu quiser fazer alguma previsão, vamos pegar dois dados aqui, pra nós ver se nós justamente estamos fazendo certo. Aqui não vai dar nada, vai dar um erro. O meu M é um array. Aqui também. Então nós vamos pegar, fazer essa equação e ver o que dá de resultado. Vamos pegar um valor da nossa planilha. Eu peguei esse valor do filme Os Doze Macacos. Vou chegar no meu dado, troquei meu X por esse dado, vamos ver. Deu 5.6 milhões.

[03:34] Agora nós vamos ter o modelo.predict. Deu exatamente o mesmo dado, ou seja, nós encontramos a equação da reta que nós queremos. E qual que é essa medida de erro? Como é que nós estamos? Como que esse dado é bom? Será que esse dado é bom? Como que nós avaliamos isso?

[03:54] A ideia mais simples é nós tentarmos pensar nisso como uma medida de diferença entre o dado que nós previmos e o dado real. Então vamos construir aqui de novo. Nós vamos ter vários pontos, esses pontos são os pontos dos filmes e, vamos lá, aqui também mantendo essa correlação positiva, pontos dos filmes. Nós vamos ter aqui a reta de previsão, essa é uma linha, essa reta aqui.

[04:33] E qual que é a ideia? Eu querer calcular a diferença entre esse ponto e a reta que eu gerei. Entre esse ponto e a reta que eu gerei. Entre esse ponto e a reta que eu gerei. Entre esse ponto e a reta que eu gerei. Entre esse ponto e a reta que eu gerei. Isso é pra todos os pontos que eu tenho aqui.

[05:05] Vamos digitar aqui, como se eu estivesse somando todos esses pontos, o dado que eu previ, o dado real, na verdade, menos o dado que eu previ, pela minha equação. Mas você não concorda que se eu simplesmente só somar esses dados e, lembra que eu tenho valores positivos e negativos pra esses dados? Na verdade é o I pra cada ponto do meu dado, então é como se fosse um I aqui, da i-ésima linha, dado real e o dado previsto. Aqui seria P, I, porque é de P e I, então vou colocar uma vírgula para o dado previsto em relação aquilo.

[05:49] Na verdade, em estatística nós usamos o chapéu I, pra dizer que é a nossa variável estimada e não a variável real. Eu vou copiar aqui de fora, vamos ver se aparece. Y, não aparece. Se eu digitar assim. Não tem aqui no Chrome, mas é outra alternativa. Então, só pra deixar aqui, vamor deixar esse P.

[06:24] Mas o ponto é que nós podemos sair somando todos esses dados e aqui pode dar um 0, e nós podemos chegar nosso erro ao 0. Então nós não erramos. Mas esses pontos são todos espalhados, na verdade é porque nós tivemos muitos pontos pra cima da reta e muitos pontos pra baixo da reta, e quando nós somamos esses dois, deu 0.

[06:41] Como que nós saímos disso? Uma ideia muito simples é justamente nós elevarmos esses valores ao quadrado, e quando nós tivermos aqui, ao invés de nós termos esses dados, que foi o dado previsto menos o dado real.

[06:59] Então dá pra ver que na verdade tanto faz, eu posso deixar assim ou justamente ser o dado previsto menos o dado real, vamos deixar yr, eu posso deixar assim e eu posso elevar esse valor ao quadrado. Então posso fazer assim, vou até o final e elevei ao quadrado, ou seja, deu um valor positivo.

[07:32] E o que acontece? Esse valor vai indo cada vez mais alto, conforme for maior essa diferença, e eu só me preocupo com a diferença. E eu vou elevando ao quadrado e eu vou ter a soma novamente desses termos ao quadrado, então ele não é mais assim, eu vou ter esse valor aqui, esses valores internos. Aqui é um pouco difícil de digitar. Eu vou ter esse valor ao quadrado.

[08:05] Então eu tenho esse valor ao quadrado, a somatória de todos esses pontos. Exatamente aquilo que eu estava escrevendo ali, eu somei pra todos esses pontos e elevei ao quadrado.

[08:14] Mas esse valor deu muito grande, deu muito pequeno? Ainda não é muito bem aquilo que nós queremos. O que seria mais interessante era eu entender, na verdade, quantos por cento da variação que eu tenho no meu Y ela pode ser explicada por uma variação que eu tenho no meu X, quantos por cento a minha variável dependente varia, ela muda em função de uma mudança que eu tenho na minha variável independente, no caso que nós estamos trabalhando, no investimento.

[08:43] Como que nós descobrimos isso? Num primeiro momento nós tentamos entender a variação total desse meu Y, qual que é? Vamos pegar essa reta, a reta mais simples possível que nós poderíamos traçar pra fazer nossas previsões, qual seria ela? Se pra cada valor de investimento que eu tivesse, o meu Y sempre fosse o mesmo, não é? Então uma reta exemplo disso é justamente se eu pegasse esse segmento de reta e fosse exatamente alguma coisa mais ou menos assim.

[09:11] Ou seja, pra cada X que eu tenho, pra cada X que eu venho nessa direção, o meu Y não varia. E se eu for pegar e fazer exatamente essa conta de novo, ou seja, pra cada Y que eu tenho, menos é o valor que eu tenho, ou seja, só que como que eu calculo essa reta? Já que justamente o valor vai ser único, qual que vai ser esse valor? É se nós pegarmos todos esses dados e calcular a média deles.

[09:48] No caso aqui que nós estamos trabalhando, seria alguma coisa do tipo. Nós temos os nossos Y representados pelo filmes_bilheteria, e nós vamos lá e calculamos a média. Como que nós calculamos a média? Isso daqui pode ser interpretado como um array do dataframe, então uma operação de array nós temo o np.mean lá do numpy filmes_bilheteria. É como se pra todo investimento que nós tivéssemos, a bilheteria sempre fosse a mesma, no caso, esse valor, fosse aproximadamente fosse 6.7 milhões de pessoas.

[10:2 1] Nós temos esses dados exatamente nessa média, nesse valor de medida. E o que nós fazemos? Nós simplesmente calculamos a diferença como nós fizemos aqui. Então nós vamos ter mais uma somatória, vamos chamar de sum, justamente pra explicar também pra nós. Eu vou até apagar e fazer de novo.

[10:46] Nós vamos ter aqui um sum do nosso Y-Y, a média de todos os Y pegando esses valores e elevando ao quadrado, não é esse daqui que está sendo elevado ao quadrado, é o que está fora que está sendo elevado ao quadrado. Somei tudo

isso.

[11:04] Se nós lembrarmos de estatística, o que é isso daqui? A soma dos termos ao quadrado é pegando ali, tirando a média. É a variância total que eu tenho da minha variável independente da bilheteria, é o total que ela varia. E o que nós queremos? Nós queremos a variação que é explicada, que Y varia em função de quanto o X varia, ou seja, eu tenho meu total do Y varia, eu falo, quanto o X impacta nessa mudança?

[11:29] Nós não temos essa informação, mas nós temos o tanto que não é explicado por essa mudança. Como que nós temos isso? Porque é justamente essa diferença que nós calculamos, essa soma desses erros quadráticos que nós calculamos.

[11:43] Então se nós dividirmos um pelo outro, e eu vou até escrever aqui em cima pra facilitar, eu vou ter aqui o sum, na verdade aqui ele está pegando como se fosse uma potência, não é esse caso que nós queremos.

[11:54] Eu vou ter, se eu for aqui pra fora – vamos ver se eu estou enxergando certo – esse valor aqui, eu vou ter outro sum, e o sum vai ser sum de soma, novamente, e aqui embaixo vai ser o sum que nós queremos, que é esse sum que eu acabei de calcular, então vou ter mais um dado aqui. Então nós podemos apagar esse dado aqui. Apagamos os dados. Nós temos essa somatória ao quadrado, que é justamente a variação total do nosso Y.

[12:39] E nós queremos entender do X, como que é do X, pro valor que nós temos, que nós previmos aquele nosso Y pred, de previsão, menos o valor real do nosso Y. Aqui não faz diferença a ordem dos elementos, porque eu estou elevando-os ao quadrado, então aqui eu vou ter uma elevação ao quadrado. Aqui está esse Y pequeno, mas na verdade é um Y grande, -Y, só que aqui o Y é normal e aqui é o nosso Y previsto.

[13:14] E eu vou ter justamente a quantidade de variação que não é explicada, então se eu quero, o que não faz parte dessa porcentagem, já que esse valor vai justamente variar de 0 até 1, já que esse daqui é a variação total, e essa é a variação não explicada pelo X, então eu preciso pegar um menos, ou seja, nós vamos ter, com essa métrica, quanto que o nosso X, a variação dos nossos dados de entrada, conseguem explicar o comportamento de dados geral, nossos dados gerais, a nossa reta de regressão geral, a nossa reta geral.

[13:48] E essa informação, essa métrica é uma medida que eu vou ter, no final das contas, essa medida vai de 0 até 1, ela é conhecida como o coeficiente de determinação, o R quadrado. Quanto mais perto do 0, menos explica, quanto mais perto do 1 ela mais explica, por quê?

[14:04] Se esse valor der 1 quer dizer que não tem nada, o X não explica nada, e vai dar 0 esse resultado. E quando aqui der 0, quer dizer que é um valor total que eu estou explicando. A quantidade que eu vario o meu Y é justificada por uma variação no X, é total.

[14:32] Se nós voltarmos pro nosso terminal, existe essa medida de pontuação, que é justamente esse dado, essa medida de pontuação é determinada pelo método score. Então vindo pro nosso script, nós podemos aplicar esse modelo.score nos nossos dados de treino pra ver o que aconteceu, mas não só isso, nós também podemos aplicar isso nos nossos dados de teste.

[15:03] E vamos aplicar primeiro nossos dados de treino pra ver o que acontece? Nós temos aqui nossos dados de treino, aplicamos, vamos aqui, modelo.score, aplicamos, deu 0.54% dos dados. Isso quer dizer que os nossos dados explicam mais ou menos 54% dos dados totais, isso é meio ruim porque se nós jogarmos uma moeda é quase a mesma coisa, a previsão é quase igual, pra fazer essa previsão. Então o treino ali não está indo muito bem.

[15:32] Vamos aplicar nos nossos dados de teste, que é uma situação um pouco mais genérica, já que nós nunca vimos os dados de treino e a reta é totalmente genérica pra esses caras. Vamos lá aplicar. 0.52, deu um pouco pior, deu 2% pior, mas deu um pouco pior, então isso não está legal.

[15:51] Nós não podemos chegar pro nosso chefe e falar: “nós criamos um modelo, chefe, que acerta 50% das vezes” ou 55% das vezes ou quase 55%. Ele vai falar: “mas então se eu jogar uma moeda eu vou fazer uma previsão melhor que a sua”, você vai ser demitido e ele vai usar uma moeda pra fazer as previsões, não é isso que nós queremos. Talvez nós precisemos de mais dados pra encontrar isso, só o investimento não é suficiente.

[16:15] Como que nós fazemos isso então? Como que nós descobrimos o trabalho de modelos um pouco melhores? É exatamente isso que nós vamos descobrir no próximo vídeo.