01

A regressão linear múltipla

Transcrição

Você pode baixar o arquivo CSV movies_multilinear_reg.csv (https://s3.amazonaws.com/caelum-online-public/machine-learning-aprendizado-supervisionado/movies_multilinear_reg.csv).

Seguem também os dados do filme Zootopia:

```
movieId,Titulo,Documentary,Sci-Fi,Mystery,Horror,Romance,Thriller,Crime,Fantasy,Comedy,Animatio
9999999,Zootopia,0,0,0,0,0,0,0,1,1,1,1,0,1,110,27.74456356,?????
```

[00:00] Na verdade não está muito bom. Como nós podemos ver, essa métrica que nós vimos, a nossa reta explica quase metade dos nossos dados. Então cerca de 50% são explicados e trazem uma informação realmente relevante. 50% das nossas previsões seriam consideradas assertivas, se nós jogarmos uma moeda, a diferença é que o nosso modelo opera 5% melhor do que uma moeda, é estranho isso até.

[00:32] E nós podemos chegar pro nosso chefe e falar: “talvez nós precisemos de mais dados”. E chega uma pessoa do departamento de Relações Públicas, coincidentemente, e fala que descobriu, estava dando uma explorada nos arquivos antigos e descobriu dados mais completos, dados como eles realmente eram, tanto pro caso de Zootopia quanto pra todos os outros filmes que nós temos. Então nós ficamos: “nossa, cara, que animal. Sensacional isso”. E ele chega e nos mostra, ele fala: “esses são os dados que nós temos”. Estão abrindo aqui os dados.

[01:02] Se nós vermos, vou até dar um zoom aqui, agora nós não temos mais só o investimento. No caso de Toy Story, nós temos como se aqui eles fossem as categorias e elas já estão codificadas em dummy variables, ou seja, eu teria uma única tabela, uma única coluna chamada categoria e essa coluna poderia ter múltiplos valores, eu isolei, de forma que cada valor das categorias fosse uma única coluna, e eu só me pergunto se tem ou não tem essa determinada categoria do filme.

[01:39] No caso documentário, sim ou não? Sci-Fi, sim ou não? Mistério, sim ou não? Aqui no caso de Toy Story é fantasia, sim. Comédia, sim. Animação, sim. Infantil, sim. Aventura, sim. O resto, tudo não. E além disso, tem a duração. No caso aqui foi 103 minutos e alguns quebrados. Investimento, 11 milhões e a bilheteria que ele teve.

[02:02] Infelizmente nós não conseguimos ver esses dados, não conseguimos botar essas dados num gráfico, por quê? Porque lembra que nós tínhamos aquele gráfico de pontos, onde eu teria um eixo X, um investimento e uma bilheteria associada?

[02:13] É como se eu tivesse agora, por exemplo, um investimento, uma bilheteria e um eixo Z de duração, e um eixo, talvez uma quarta dimensão, num quarto hiperplano aqui que nós não conseguimos visualizar, infelizmente, pra aventura, e mais um pra drama, e nós vamos expandindo esses dados de forma que nós não conseguimos ver essa relação.

[02:35] Mas a ideia agora é nós conseguirmos traçar essa reta, calcular essa regressão de forma que nós tenhamos uma variável dependente, mas ela depende, na verdade, de várias variáveis que são independentes entre si. Então eu vou ter um M1 associado a investimento, um M2 associado a duração.

[02:58] Voltando aqui pro nosso caso do GeoGebra, nós temos um Y, nós teríamos um MX + B. Agora, quando nós estamos pensando assim, de regressão, no caso de uma regressão múltipla, justamente porque nós estamos lidando

com múltiplas variáveis, nós vamos ter um M1 vezes um X1, um investimento, mais um M1. E aqui, vamos setar um M1, mais um M2. Esses dados estão grandes. Então um MX e esse M mais outro MX, mais outro MX e assim por diante.

[03:49] Então nós teríamos aqui um M2 vezes outra variável aqui, que no caso é a duração. E aqui um M3 vezes uma das categorias, e assim por diante.

[04:07] E agora qual que é a nossa ideia? Nós precisamos ler os nossos dados, quebrá-los em dados de treino e teste, e simplesmente repetir o processo, passando mais variáveis. E o scikit-learn já automaticamente detecta se nós estamos passando múltiplas variáveis ou uma variável só.

[04:23] Como? Pela forma do nosso data frame, pela disposição dos nossos dados, porque aqui nós passamos 6843 linhas em uma única coluna. Mas se nós passarmos 6843 linhas em várias colunas, ele já entende que está se tratando de uma regressão múltipla.

[04:38] Como é que nós fazemos aqui? Nós vamos primeiro ler o nosso arquivo, vou chamar de movies, ele recebe pd.read_csv("datasets/movies_multilinear_reg.csv"). Vamos ler pra ver o que aconteceu? Nós abrimos aqui, vamos no terminal, vamos ler. Fizemos a leitura.

[05:29] O que nós queremos agora? Nós queremos da segunda coluna em diante pra ser as nossas variáveis independentes. E então, num primeiro momento, vamos só pegar, nós estamos pegando as colunas, então movies[movies.columns[2:17]]. Se tem 18 lá, a penúltima é até a 17. Agora copiei esse dado, vamos pra esse terminal, "filmes_independente".

[06:01] O que acontece agora? Ele está pegando justamente da segunda em diante, passando só aquilo que nós queremos. Então justamente não é mais investimento que é só esses dados. Então se nós olhamos o tipo desse dado, vou limpar aqui, ele também vai ser um dataframe, nesse caso, "independente", olha só, é o data frame. E o scikit já conseguiria entender isso pra nós.

[06:36] Vamos agora fazer a mesma coisa com o nosso filmes dependente, no caso, que é a nossa variável dependente. E como é que nós a pegamos? Nós estamos pegando justamente a última coluna, então "movies.columns", da 17 até o final, também conhecida como última coluna. Então vamos vir aqui, colei. Olha só, qual é o tipo dela? Olhando em filmes.

[07:17] Eu agora vou vê-la em si, vamos vê-la em si? Olha só, 9125 linhas pra uma coluna. Se nós virmos o tipo dela, da forma como nós pegamos aqui, não é mais um series, vai ser já um dataframe, ele já está separado pra nós.

[07:34] E nós vamos passar esses dados pra fazer o nosso treino, então outra forma de nós treinarmos, passando justamente os dataframes, ao invés de ter que fazer o reshape e tudo mais com o array que nós vimos no primeiro momento.

[07:47] No primeiro caso aqui, vamos dividir em treino e teste. Como nós fizemos já anteriormente, treino e teste, "train_bilheteria", "test_bilheteria", ele vai receber "train_test_split". Filmes dependente, filmes independente. "filmes_independente", "filmes_depentende".

[08:20] Agora aqui nós temos nossas variáveis de treino, temos aqui 15 colunas, 6843 linhas. Se nós lembrarmos desse número, bate também com o tamanho dos dados que nós queremos, 9125, a mesma quantidade de dados então a proporção é a mesma, 75%, 25%. Agora o que nós precisamos fazer é justamente treinar o modelo de novo, vamos até recriá-lo aqui.

[08:44] O modelo vai receber "LinearRegression" e agora nós vamos aplicá-lo, o "fit" que nós temos, o "train" e o "train_bilheteria". Agora nós pegamos esse dado, copiamos, e ele já treinou. Então ele já entende que nós estamos

tratando de uma regressão múltipla, já que eu passei múltiplos dados na minha variável de treino, na minha variável independente.

[09:18] Vamos ver agora como são as métricas que nós conseguimos ter. Primeiro de tudo, vamos fazer uma previsão? Vamos à previsão, pra ver como que esse dado está se comportando. Vamos pegar aqui o Toy Story, que nós já tínhamos ali a informação. Aqui vai copiar, copiei esse cara, "Command + V". E aqui deu problema.

[09:50] Por que ele deu problema? É um motivo muito simples, porque nós não estamos mais agora passando só o investimento, nós precisamos passar todo esse vetor. Nós temos que passar um conjunto de vetores, que são todos os nossos dados que nós temos, todas as informações que nós temos pra conseguir fazer a previsão, lembra? Porque justamente nós moldamos essa reta, que na verdade é mais essa reta, nós criamos essa regressão aqui pra nós.

[10:17] Eu estou copiando esses dados à mão justamente pra nós vermos com cuidado esses caras e pegá-los na ordem, então aqui no caso é 1, dá até um pouco de trabalho. Agora aqui é 0, aqui é 1.

[10:36] Esses dados estão no csv. Csv é separado por vírgula, então talvez se nós entrarmos aqui, nós podemos abrir aqui, multilinear, reg, vamos ver, será que nós temos sorte? Olha só, os dados já estão aqui organizadinhos pra nós, já separados por vírgula. Muito mais fácil. Então nós vamos fazer essa previsão. Vamos até voltar aqui.

[10:59] Então "modelo.predict", nós passamos um array, que é justamente o formato dos dados que nós temos. Olha só, ele está reclamando, porque eu precisaria fazer um reshape dos dados. Não, na verdade é porque eu estou passando array de 1d, então é só eu fazer o quê?

[11:15] Se eu fizer um modelo.predict aqui e passando aqui e aqui, e aqui dentro, eu tenho os dados que eu quero, vamos até fazer de novo, aproximadamente 5 milhões, 5.8 milhões. Se nós olharmos no nosso gráfico, 5.6 milhões, é quase lá. Uns 200 mil de diferença, um valor muito diferente, muito menor do que aqueles quase um milhão que nós vimos da primeira vez.

[11:43] Lembra que novamente nós tínhamos aquele coeficientes e o interceptador? Vamos ver quais são os dessa reta? Se nós viermos aqui, modelo.coef, é um pra cada um daqueles caras. Então aquele M1 que nós vimos, que é o caso se nós olharmos pro documentário, ele é esse valor. Se nós olharmos pro nosso último c ara, que era o investimento, o investimento novamente aqui, é agora esse valor. E agora o intercept é o mais B que eu coloquei ali no final. Interessante.

[12:18] Se eu pegar esse dado e multiplicar por 0, esse dado e multiplicar por zero, e esse dado, multiplicar por esses 11 milhões e somar esse B no final, eu vou ter o mesmo valor que eu acabei de fazer a previsão. Aqui daria um pouquinho mais de trabalho pra fazer na mão, mas também daria pra fazer.

[12:33] Agora nós definimos nessa curva que nós queríamos encontrar, vamos ver o quão bom é esse dado? O quão boa é essa métrica? Então novamente nós vamos calcular o R quadrado dela, o coeficiente de determinação usando score.

[12:46] Vamos pensar primeiro nos nossos casos de treino, como que nós digitamos? Por isso que eu escrevi o script, eu comentei antes. Eu tenho aqui o meu dado de treino, teste, então vou treinar treino e treino bilheteria, então é train e train bilheteria. Então vamos treinar em cima do nosso caso de treino pra ver como é que ele se saiu. Olha só, 83%. Então em 83% dos casos a nossa previsão é assertivamente boa. Nós explicamos 83% das nossas informações.

[13:19] Vamos treinar agora, vamos ver, na verdade, como que esses dados se comportam em cima de dados que nós nunca vimos, em cima dos nossos dados de teste? Vamos repetir o processo, "modelo.score(test, test_bilheteria)". 81%. Um pouquinho menos, mas como nós podemos ver, dá aqueles quase 82%, que é praticamente a mesma coisa. Mas mesmo assim, acima de 80% que nós podemos afirmar, com certeza, que os nossos dados estão bons.

[13:49] E olha só que legal, nós podemos chegar então lá pro nosso chefe e falar: “pra Zootopia, com 83% de assertividade, desobri o modelo que acerta 83% das vezes, ele explica 83% dos nosso dados. Então nós temos certa confiança de que nós vamos fazer uma previsão legal pra Zootopia”. Então nós também temos esses dados pra Zootopia. Vamos ver como que é? Qual que seria esse caso?

[14:12] Nós temos o predict aqui e nós vamos passar pra ele, o que nós passamos pra ele? Os dados que nós temos. Novamente nós podemos abrir direto aqui no Atom, que ele já separa pra nós no preview, e nós já temos outro dataset, que é o “zootopia_completo_data”. Ele tem exatamente todas essas informações.

[14:33] Vamos abrir aqui um csv só pra nós visualizarmos? Nós abrimos um csv, olha só, só tem informação de Zootopia, mas tem todos os dados que nós queremos. Duração, 110 minutos; investimento, quase aqueles 27 milhões que nós vimos antes; categoria de fantasia, comédia, animação, criança e aventura.

[14:51] Nós vamos pro Atom novamente, vamos só copiar esses dados. Vamos aqui, copiamos os dados e agora vamos fazer a previsão, passar os dados aqui. Pra essas informações que nós temos da Zootopia, a nossa previsão é de aproximadamente 7.7 milhões, e nós afirmamos isso até mesmo com mais certeza.

[15:17] Nós podemos chegar pro nosso chefe e falar: “pra esses dados que nós vamos investir, pro dado que você for investir, a previsão quando esse filme chegar no Brasil é que vão ser de aproximadamente 7.8 milhões de pessoas”. Ele vai falar: “nossa, que legal, é um valor alto, então vale a pena investir nesse filme” e todo mundo fica feliz, e você ainda mais, porque você teve certeza, um pouco mais de confiança, um grau de confiabilidade alto na resposta que você está dando.

[15:39] Esse é o processo de regressão. Toda essa aula que nós vimos é uma aula de regressão e nós passamos desde o processo de análise dos dados, pra nós entendermos o nosso problema, até construir o modelo preditivo, refinar esse modelo, interar, e agora nós chegamos num modelo aceitável.

[15:59] E esse ponto, novamente, quando eu comentei no início do vídeo de regressão linear, a ideia é que a regressão é um método estatístico e a regressão vem num nome justamente da estatística, é porque ela cospe pra nós um número, diferente do que nós vimos em outros cursos de Machine Learning, que nós cuspíamos uma classificação, nós estamos preocupados em classificar sim ou não, positivo e negativo, agora nós estamos preocupados em prever um número.

[16:24] Mas nos dois casos nós sempre tivemos que aprender em cima de dados que nós já tínhamos antes. Aqui nós já tínhamos todos esses dados pra nós generalizarmos, nós aprendemos em cima de dados que nós já vimos e já conhecemos, pra fazer previsões em cima de dados que nós vamos ver ou não vimos, nós generalizamos pra esses casos. É o caso que nós vimos com o caso de classificação, em outro vídeo aqui da Alura, e é o caso que nós vimos agora, com o caso de regressão.

[16:57] Mas, além disso, nós podemos pensar nesse caso de regressão, não fazendo uma regressão, mas pra classificar, exatamente como nós vimos com Naive Bayes, mas, ou seja, outros métodos de classificação, é outro método de classificação.

[17:12] Mas esse caso que nós estamos vendo agora, que nós estamos aprendendo em cima de dados que nós já vimos, é o conhecido como aprendizado supervisionado. E a ideia é justamente isso, é nós supervisionarmos em cima de informações que nós temos, pra depois de generalizar pra outros casos.

[17:29] A seguir nós vamos entender como que funcionaria essa regressão, essa reta ou essa curva, pra nós trabalharmos com problemas de classificação também, essa é outra abordagem pro caso de aprendizado supervisionado, essa é a ideia da regressão logística, e é justamente isso que nós vamos aprender no próximo vídeo.