04

Entendendo os seus dados

Transcrição

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

[00:00] O desafio que nosso chefe nos passou foi: o próximo filme que nós vamos receber para fazer a distribuição é o Zootopia. E qual deve ser a bilheteria dele? Você consegue prever mais ou menos quantas pessoas vão assistir ao filme?

[00:20] Se nós entrarmos aqui no Google, vamos entrar aqui no Google e ver um pouquinho melhor sobre ele. Vamos um pesquisar aqui: Zootopia, que é um filme. Vamos ver se conseguimos algumas informações dele, entrando aqui no IMDB dele. O que nós conseguimos ver?

[00:38] É uma animação, aventura e comédia. Tem uma hora e pouco de duração, quase duas horas. É um filme da Disney. Isso é interessante, uma informação interessante.

[00:51] Venceu um Oscar, ganhou vários prêmios. Essas animações da Disney, geralmente tendem a fazer bastante sucesso. Nós temos o indicativo que a chance desse filme fazer sucesso e ter uma qualidade boa, é alta. Mas se nós chegarmos para o nosso chefe e falarmos: vai vir muita gente.

[01:10] Ele vai falar: mas isso é meio óbvio. É um filme da Disney. Essas coisas geralmente fazem sucesso. Eu preciso de um dado mais concreto. E você pensa: poxa, eu preciso de um pouco mais informação. Explorar um pouco melhor os nossos dados.

[01:28] Se eu entrar aqui na minha pasta, eu vou ver abria a minha pasta. Estou andando aqui pelos diretórios. Aqui, meu curso. Essas aqui, no caso, essas são as primeiras informações que nós temos. Vou dar um zoom, aproximar duzentos por cento. Aproximei os dados.

[01:48] Qual é informação que nós temos? A única informação que nós temos para prever a bilheteria é investimento. Como que esses conjuntos de dados ou datasets estão distribuídos? Vamos dar uma explorada neles.

[02:02] A primeira linha é do filme Toy Story. Nós podemos ver que o investimento de foi de mais ou menos onze milhões de reais. No caso, dólares. Porque é um filme do exterior e trouxe pra cá. Teve uma bilheteria aproximada de 5.7 milhões de pessoas.

[02:18] Jumanji também foi outro filme famoso, teve mais ou menos 15 milhões e a bilheteria foi 5.8. Grumpier foi um numero maior. Teve 30 milhões e 9.5. Aqui, 5 milhões. Seis.

[02:32] Conforme nós podemos ver dos filmes, nós podemos ver: poxa, tem uma informação que são de investimentos que ele teve e a bateria associada. Será que esses dados tem alguma ligação? Será que eu consigo correlacionar esses dados?

[02:48] Primeiro, nós precisamos explorar esses dados de um ponto de vista um pouco melhor. Qual a ideia que nós temos? E se nós pegássemos e considerássemos um gráfico onde um eixo fosse o investimento e o outro eixo, ou seja, o eixo Y, por exemplo, é a bilheteria. Então, eixo X é o investimento e o eixo Y é a bilheteria.

[03:11] Eu vou ter dois valores e um ponto um ponto X, investimento. Bilheteria, Y. E eu consigo fazer um gráfico de pontos. Se eu pegar esses gráficos de pontos e dispersar, eu vou tem um gráfico de dispersão. Vamos dar uma olhada em

como eu faria isso.

[03:30] Vou entrar aqui no Atom, que é um editor de texto, e eu quero escrever o script. Vou dentro do meu curso eu vou salvar aqui o script. Que nome nós podemos dar para ele? Exploração, ponto, py. Nós queremos que seja aqui dentro. E aí nós vamos salvar em um arquivo python.

[03:55] Num primeiro momento, nós queremos usar a biblioteca de análise de dados do python. Quer dizer, não é do python, é a biblioteca que nós conseguimos fazer a análise de dados, que é o pandas. Import pandas as pd. Certo. Depois, durante os exercícios, eu vou fazer um vídeo separado como que nós instalamos todos esses módulos essenciais.

[04:17] Além disso, eu também vou usar a biblioteca para nós podermos plotar os gráficos, nós a conseguimos por meio do matplotlib.pyplot. Aqui eu vou usar um alias para me referir a ela. Além disso, o que nós precisamos também? Nós precisamos agora escrever esses dados. Num primeiro momento, o que eu quero? Eu quero ler esse csv que eu tenho desses dados. Esse csv que eu acabei de abrir.

[04:45] Como que eu faço isso? O pandas tem um método que é justamente isso, ler csv, read csv. Eu tenho o meu movies, que são os filmes que eu tenho. Eu botei aqui um "pd.read_csv()". Eu sei que está dentro do meu datasets, certo? Eu posso justamente aqui escrever "datasets/" o nome do arquivo que eu quero, que é "regrassao_linear_alura.csv". Se formos ver aqui, é exatamente esse nome.

[05:20] Aqui já apareceu um pouco de informação. Ao invés de eu sair escrevendo o script, vamos entender o que são esses dados e o que está acontecendo aqui até então?

[05:34] Já vou salvar esse dado, que nós vamos usá-lo. Vamos abrir aqui o terminal. Expandir aqui. Cd, documents. Ls. Cd, André. Ls. Curso, que é o meu curso. Voltei para cá. Vamos entrar aqui no python. Vamos importar as bibliotecas que eu falei. Estamos aqui, felizes e contentes, importando as duas bibliotecas. Importamos as duas bibliotecas.

[06:08] Eu vou até reescrever para ficar um pouco melhor "pd.read_csv("datasets/regrassao_linear_alura.csv")". O que é isso aqui, pessoal? É basicamente o mesmo csv que nós acabamos de ler aqui, só que numa estrutura dentro do python, que é um data frame.

[06:47] O data frame é justamente uma estrutura que o pandas tem, onde nós conseguimos fazer essa análise de dados um pouco mais fácil. Se eu limpar aqui, se eu digitar aqui o type movies é um data frame, certo? Qual é a vantagem do data frame? Nós conseguimos manipula-lo muito mais facilmente. Se eu quiser pegar só os cinco primeiros dados, eu consigo pegar só os cinco primeiros dados.

[07:15] Aqui ficou meio grande, mas vou pegar, por exemplo um pouco melhor, só o primeiro dado "movies.head(1)". Eu tenho aqui o Toy Story, o investimento que ele teve, as bilheterias e o número de pessoas. E se eu quiser acessar só os meus dados sobre investimentos. Eu não estou interessado no título, eu não estou interessado na bilheteria. Só quero olhar os investimentos.

[07:40] É só fazer assim: investimento em milhões. Ops, não deu certo por conta desse acento aqui. Vamos limpar aqui. Investimento em milhões. Olha só que legal. Se eu pegar o tamanho, eu quero ver o número de linhas que tem, eu só preciso fazer "len(movies)". No caso, é o número de linhas que ele tem. Se nós olharmos aqui, coincide totalmente. São nove mil, cento e vinte e cinco filmes. Divididos em quatro colunas, que é justamente o ID do filme, o nome, o investimento e os bilhões de pessoas.

[08:20] O que nós queremos fazer agora? Nós queremos plotar isso nesse gráfico de pontos que eu tinha falado. Lembra que eu falei que nós queremos um eixo X e nós queremos um eixo Y. O nosso eixo X, nesse gráfico, nós queremos que

seja o investimento. É só fazer assim. Eu quero que o X, uma variável X receba o investimento em milhões. E o eixo Y recebe a bilheteria. Bilheteria, pessoas.

[09:03] Vou printar aqui. X, investimento em milhões, exatamente como nós tínhamos visto. Se eu chegar da mesma forma, eu vou ter um Y. Mesma coisa. O que nós precisamos fazer para plotar esse gráfico de pontos, esse gráfico de dispersão, que do inglês é scatter plot. Quando nós fazemos essa tradução, nós temos um plt scatter. Por que plt? O plt é da biblioteca. É o alias que nós temos. É um método do matplotlib.

[09:34] Se nós chegarmos aqui e fizermos "plt.scatter(x,y)". Gerou o nosso gráfico. Eu vou até anotar aqui "plt.scatter(x,y)". Como é que nós vemos esse gráfico? Nós vemos com o "plt.show()". Justamente, o show de mostrar. Nós vamos aqui. Voltou pra cá. Printou. Apareceu.

[10:05] Está legal? Nós tínhamos nove mil, cento e vinte e cinco pontos, certo? E aí ficou quase uma rajada. Nós conseguimos enxergar que os pontos estão andando em uma direção, mas isso está muito poluído. Porque a ideia é que isso é um ponto, isso aqui também é um ponto. Mas está muito sujo. Nós temos uma noção, mas nós queremos ver isso de um ponto de vista um pouco mais claro.

[10:35] Como nós conseguimos ver isso? Vamos pegar uma amostra desses dados. Uma amostra aleatória, porque ela pode se comportar mais ou menos com base em nosso conjunto total. E nós conseguimos interpretar essa amostra. Eu vou fechar aqui e voltei pra cá. Qual a ideia, então? Nós queremos pegar uma amostragem desses dados. Amostra, do inglês, sample. Vamos mandar aqui um "plt.sample()". Ele recebe o número de dados igual a duzentos. Vamos pegar duzentos dados da nossa amostra.

[11:09] E eu vou fazer isso algumas vezes para mostrar que esses dados são totalmente aleatórios. E nós repetimos o processo. Vou aqui, estou copiando o meu sample e vamos lá. Eu limpei aqui, sample. Ele não tem o atributo sample. Na verdade, também não é sample, porque não é do plt, ele é nosso data set. Nós queremos o sample do nosso data set. Na verdade, é um método do nosso data frame, que é o pandas. É o movies, que foi o que nos criamos lá e é ponto, sample mesmo.

[11:45] Aqui nós temos o sample. Criou duzentos dados, nós podemos ver. Uma forma mais simples de vermos é se pegarmos simplesmente o primeiro cara. E veremos que não é o Toy Story. Life in a Day. Se eu repetir esse processo, eu tenho aqui sample e "sample.head(1)". Velozes e Furiosos seis. Esse é um caso mais recente. Se eu fizer mais uma vez, Assassins. Três vezes que eu fiz, três dados totalmente diferentes. Ele aleatoriza. Ele pega uns duzentos dados aleatórios para nós podemos fazer isso.

[12:23] Qual é o processo agora? É repetirmos isso. É fazermos o nosso X e o nosso Y, e repetir. Vou voltar para o meu Atom, na plt novamente. No método do nosso data frame, que no caso é o Movies. E agora, vou reescrever o meu X e o meu Y. Eu vou ter o meu sample. Isso daqui retorna um data frame também. Como pudemos ver, podemos acessar. Para acessarmos as suas colunas, é da mesma forma como acessamos as colunas do primeiro conjunto de dados. É só digitar investimento em milhões e um Y que é sample. Bilheteria. Pessoas.

[13:21] Se eu chegar aqui e copiar esses dados, e novamente fizer um "plt.scatter()", ele gerou outro dado "plt.show()". Olha só que legal. Pronto. Agora está muito mais visível para nós. Nós conseguimos visualizar exatamente aquilo que visualizamos antes, só que de uma forma um pouco mais clara, porque aqui cada ponto é literalmente um filme.

[13:48] E o que nós conseguimos ver? Que os dados fluem nessa direção, ou seja, eles têm uma associação positiva. Por que associação positiva? Porque quanto mais eu aumento o meu investimento, mais a bilheteria e o número de pessoas tende a aumentar. Ou seja, eu tenho uma associação linear, uma correlação positiva entre esses dados, porque esses dados quase formam uma linha.

[14:24] O que eu quero dizer com correlação? É que esses dados tem uma relação. Não é que um causa o outro. Nós não podemos chegar para o nosso chefe e falar: se eu aumentar o número de investimento, quanto mais eu aumentar meu investimento, mais pessoas vão ver o filme por uma relação de casualidade. Porque uma coisa não tem necessariamente a ver com a outra.

[14:49] Esses dados, às vezes, podem estar ligados. E o que nós queremos entender é justamente investigar se de fato essa correlação impacta. Será que nós conseguimos extrair alguma informação útil desses dados? Isso é exatamente o que vamos ver no próximo vídeo.