Introdução a Técnicas de Regressão

Especialização em Ciência de Dados Aplicada

Paulo Henrique Ribeiro Gabriel

Faculdade de Computação
Universidade Federal de Uberlândia

2024

Sobre a Disciplina

Objetivos

- Apresentar os principais conceitos relacionados às técnicas preditivas relacionadas a regressão
- O conteúdo será desenvolvido de modo prático com aplicações em diferentes área do conhecimento

Sobre a Disciplina

Aulas

- 18/05 (Manhã) - Introdução
- 25/05 (Tarde) - Regressão Linear
- 02/06 - Feriado
- 08/06 (Tarde) - Regressão Polinomial
- 15/06 - Recesso
- 22/06 - Recesso
- 29/06 (Tarde) - Regressão Logística
- 06/07 (Tarde) - Outras técnicas de regressão
- 13/07 (Manhã) - Apresentação de exercícios

Conceitos

- A análise de regressão é uma técnica simples de aprendizado supervisionado
- Trata-se de uma das técnicas mais básicas da área de aprendizado de máquina
- Utilizada para encontrar a melhor tendência para descrever um conjunto de dados (*dataset*)

Conceitos

Origens

- Métodos de regressão surgiram no Século 19
  - Legendre (1805) e Gauss (1809)
  - Objetivo inicial: prever órbitas ao redor do Sol
- Atualmente, possui papel central em estatística
  - Estimação de uma função de regressão
- Avanços computacionais recentes permitem que novas metodologias sejam exploradas

Conceitos

Desafios

- Capacidade cada vez maior de armazenamento de dados
- Métodos com menos suposições sobre o verdadeiro estado da natureza ganham cada vez mais destaque
- Métodos tradicionais não são capazes de lidar de forma satisfatória com bancos de dados em que há mais variáveis que observações
  - Situação muito comum nos dias de hoje
- São frequentes as aplicações em que cada observação consiste em uma imagem ou um documento de texto
  - Objetos complexos que requerem metodologias mais elaboradas

Conceitos

Objetivos

- O objetivo de um modelo de regressão é determinar a relação entre variáveis
- Estatisticamente, queremos relacionar a variável aleatória

$$Y \in \mathbb{R}$$

e um vetor

$$\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$$

Conceitos

Objetivos

- Mais especificamente, queremos estimar uma função de regressão

$$r(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

- Com isso, vamos descrever a relação entre as variáveis
- Quando Y é uma variável quantitativa, temos um problema de regressão
- Quando Y é uma variável qualitativa temos um problema de classificação

NOTAÇÃO

- Nomes da variável Y:
  - Variável de resposta
  - Variável dependente
  - Rótulo (*label*)

Notação

- Nomes do vetor **x**
  - Observações
  - Variáveis explicativas
  - Variáveis independentes
  - Características
  - Atributos (*features*)
  - Preditores
  - Covariáveis

NOTAÇÃO

- Precisamos de técnicas para estimar $r(\mathbf{x})$
- Ou, no jargão de aprendizado de máquina: treinar o modelo de regressão

EXEMPLO NUMÉRICO

- A primeira técnica de análise de regressão que examinaremos é a regressão linear
- Literalmente: utiliza uma linha reta para descrever um conjunto de dados
- Para isso, utilizamos a equação da reta:

$$y_i = \alpha + \beta x_i$$

onde:

- y_i é a variável alvo
- α e βx_i são coeficientes calculados pela regressão
- α é o intercepto no eixo y
- βx_i é inclinação da reta

Exemplo Numérico

Conjunto de dados

x	y
1	3
2	4
1	2
4	7
3	5

	x	y	xy	x^2
1	1	3	3	1
2	2	4	8	4
3	1	2	2	1
4	4	7	28	16
5	3	5	15	9
Total	11	21	56	31

Exemplo Numérico

Algoritmo

- Agora, vamos calcular os coeficientes α e β da seguinte maneira:

$$\alpha = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$\beta = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

onde:

- n é o total de linhas (“observações”)
- $\sum x$ é a soma de todos os valores da coluna x
- $\sum y$ é a soma de todos os valores da coluna y
- $\sum xy$ é a soma de todos os valores da coluna xy
- $\sum x^2$ é a soma de todos os valores da coluna x^2

Exemplo Numérico

Algoritmo

i	x	y	xy	x^2
1	1	3	3	1
2	2	4	8	4
3	1	2	2	1
4	4	7	28	16
5	3	5	15	9
Total	11	21	56	31

- $n = 5$
- $\sum x = 11$
- $\sum y = 21$
- $\sum xy = 56$
- $\sum x^2 = 31$

Exemplo Numérico

Algoritmo

Voltando para as fórmulas:

$$\alpha = \frac{(21)(31) - (11)(56)}{5(31) - (11)^2} = \frac{35}{34} = 1.029$$

$$\beta = \frac{5(56) - (11)(21)}{5(31) - (11)^2} = \frac{49}{34} = 1.441$$

Exemplo Numérico

Algoritmo

- Inserindo α e β na equação da reta:

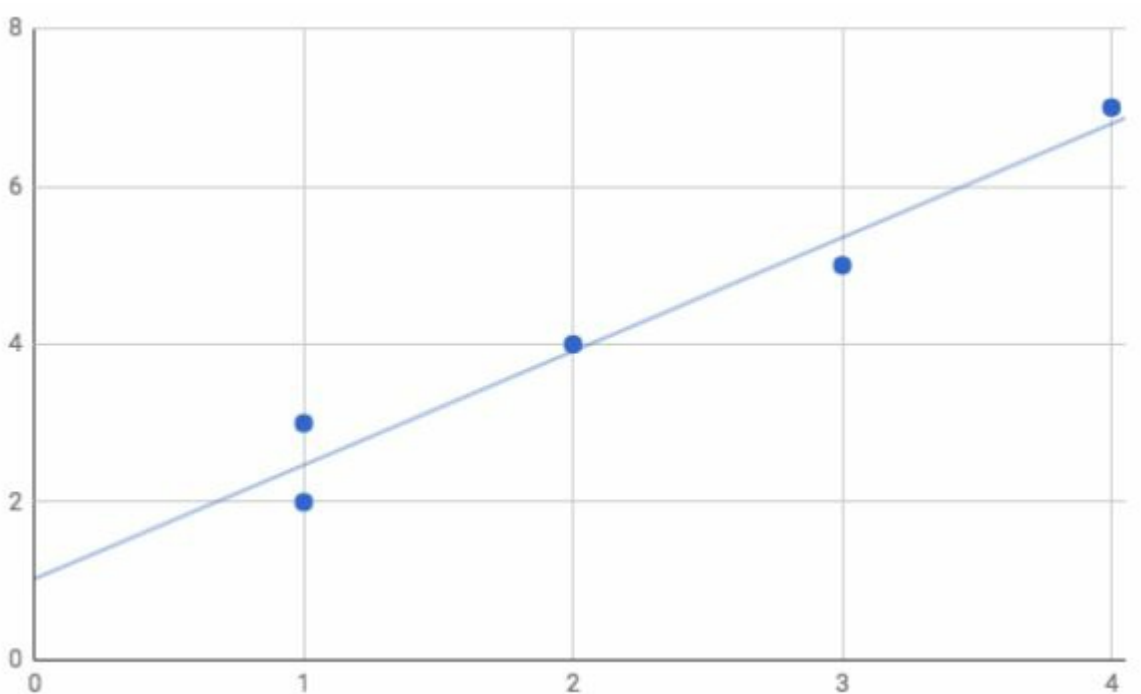
$$y_i = \alpha + \beta x_i$$

$$y_i = 1.029 + 1.441 x_i$$

- Temos, portanto, uma reta que “descreve” nosso conjunto de dados
- E agora?

Exemplo Numérico

Testando o Modelo



Exemplo Numérico

Testando o Modelo

- Vamos fazer um teste, para ver se nosso modelo está bem ajustado aos nossos dados
- Para isso, vamos considerar uma das nossas observações, por exemplo, $x_2 = 2$
- Nesse caso, o valor esperado é $y_2 = 4$
- Pela nossa equação, temos:

$$y_2 = 1.029 + 1.441x_2 = 1.029 + 1.441(2) = 3.911$$

um valor bem próximo

MAIS ALGUNS CONCEITOS

- A análise de regressão vem em muitas formas:
  - linear
  - não-linear
  - logística
  - multilinear
- A regressão linear compreende uma linha reta que divide seus pontos de dados em um gráfico de dispersão
- O objetivo da regressão linear é dividir seus dados de forma a minimizar a distância entre a linha de regressão e todos os pontos de dados no gráfico de dispersão

MAIS ALGUNS CONCEITOS

- O termo técnico para a linha de regressão é hiperplano
  - Um hiperplano, de maneira informal, é uma linha de tendência
- Uma característica importante da regressão é a inclinação (*slope*)
- A inclinação é muito útil na formulação de previsões

MAIS ALGUNS CONCEITOS

CORRELAÇÃO

- As relações entre as variáveis dependentes e independentes são feitas através de algum coeficiente de correlação
- Uma das métricas de correlação mais utilizadas é o coeficiente de Pearson , que mede a associação linear entre duas variáveis
- Esse coeficiente de correlação pode ser definido pela equação a seguir:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

onde

- n é o total de amostras
- $\bar{x}$ e $\bar{y}$ são as médias aritméticas de ambas as variáveis

Mais alguns conceitos

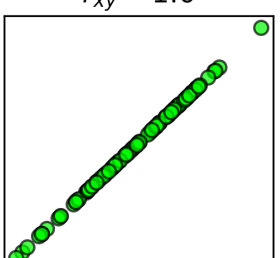
Correlação

- Os valores do coeficiente de Pearson variam entre -1 e 1
- Assim, quanto mais próximos desses extremos, melhor correlacionado estão as variáveis
- Alguns exemplos com gráficos de dispersão de variáveis com diferentes correlações são mostrados a seguir

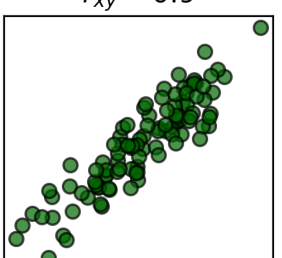
MAIS ALGUNS CONCEITOS

CORRELAÇÃO

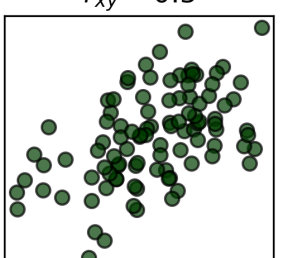
Correlação positiva perfeita
$r_{xy} = 1.0$



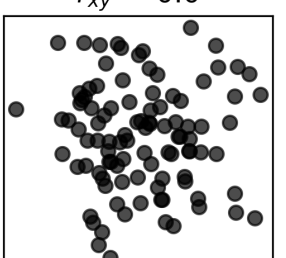
Correlação positiva alta
$r_{xy} = 0.9$



Correlação positiva baixa
$r_{xy} = 0.5$



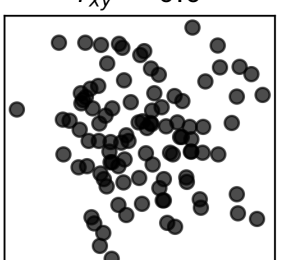
Sem correlação
$r_{xy} = \text{-}0.0$



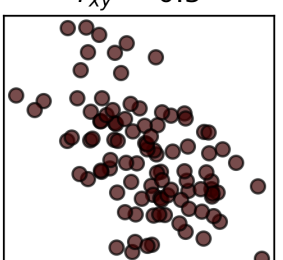
MAIS ALGUNS CONCEITOS

CORRELAÇÃO

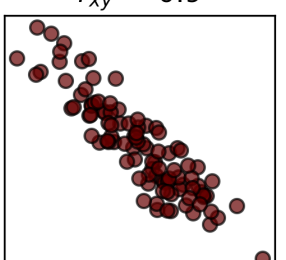
Sem correlação
$r_{xy} = -0.0$



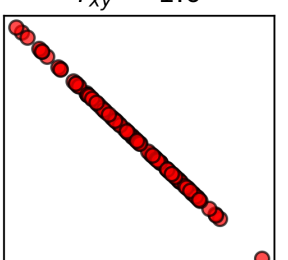
Correlação negativa baixa
$r_{xy} = -0.5$



Correlação negativa alta
$r_{xy} = -0.9$



Correlação negativa perfeita
$r_{xy} = -1.0$



Outro exemplo

Previsão do Preço do Bitcoin

Data	Preço	Dias transcorridos
19/05/2015	234.31	1
14/01/2016	431.76	240
09/07/2016	652.14	417
15/01/2017	817.26	607
24/05/2017	2358.96	736

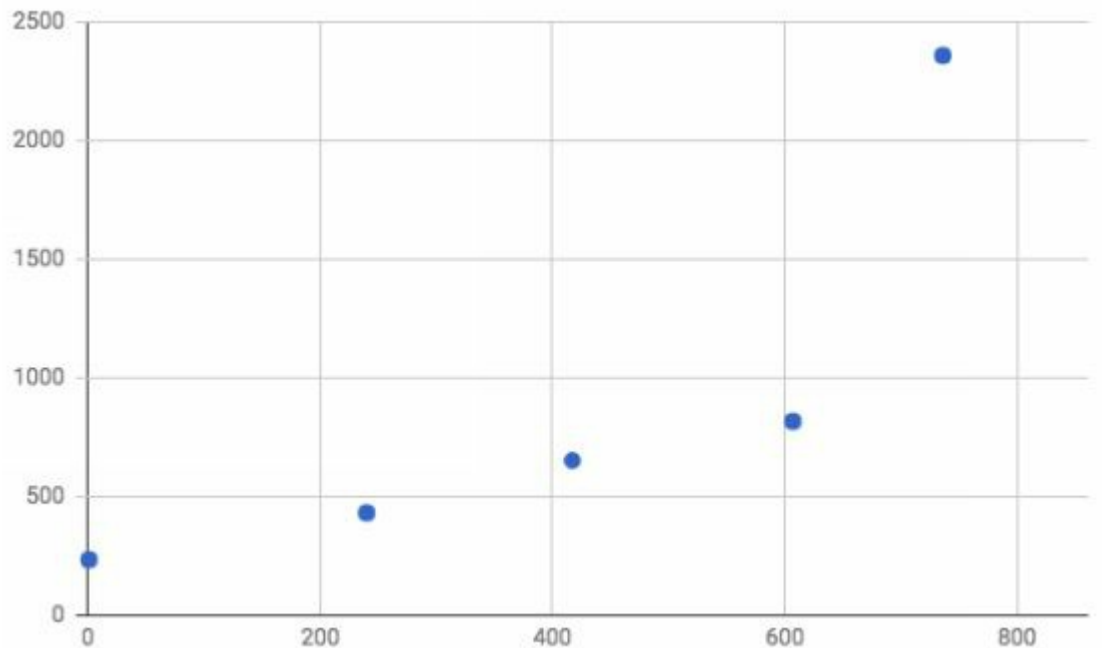
Outro exemplo

Previsão do Preço do Bitcoin

- Note que temos três variáveis (atributos)
- Vamos construir um gráfico de dispersão correlacionando o total de dias com o preço do Bitcoin
- Valores numéricos (segunda e terceira colunas) são fáceis de inserir em um gráfico e não requerem conversão especial
- Além disso, a primeira e a terceira colunas contêm a mesma variável “tempo”
  - Logo, a terceira coluna por si só é suficiente

Outro exemplo

Previsão do Preço do Bitcoin



Outro exemplo

Previsão do Preço do Bitcoin

- Nosso objetivo é estimar qual será o valor do Bitcoin no futuro
- Nesse caso, o o eixo y traça a variável dependente, que é o "Preço do Bitcoin"
- A variável independente (x_i), neste caso, é o tempo
- O "Número de dias transcorridos" é assim plotado no eixo x.

Outro exemplo

Previsão do Preço do Bitcoin

- Depois de traçar os valores x e Y no gráfico de dispersão, podemos ver uma tendência na forma de uma curva ascendente, com um aumento acentuado entre os dias 607 e 736
- Suponha que você deseja comprar Bitcoins
- Com base na trajetória da curva, o que você faria?

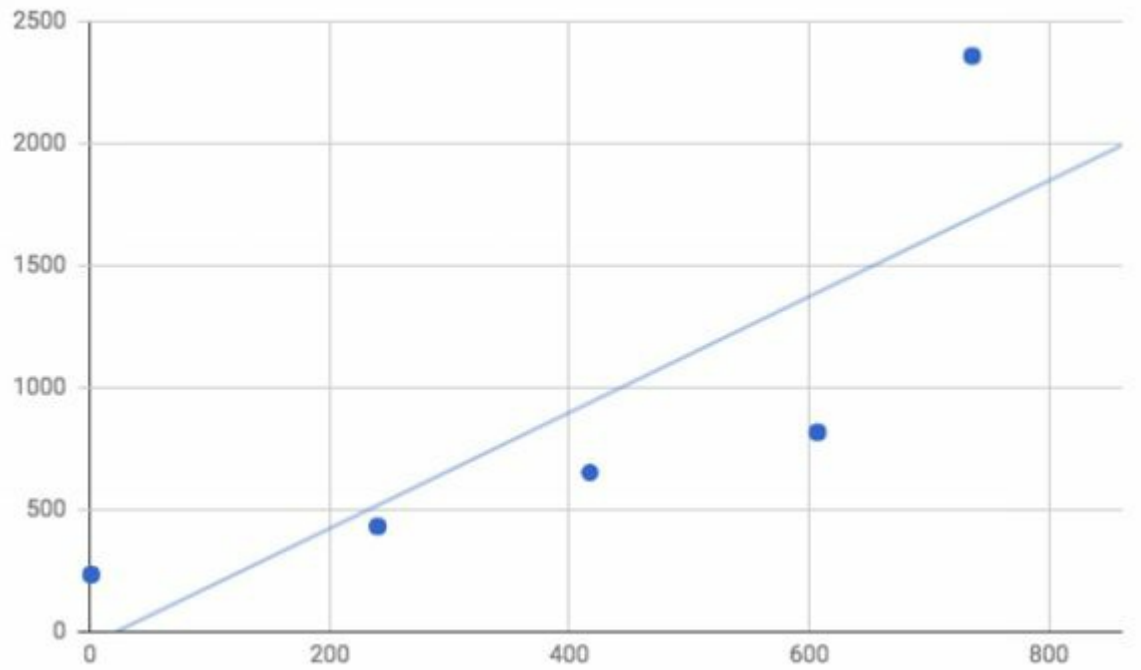
Outro exemplo

Previsão do Preço do Bitcoin

- Vamos supor que você deseja comprar e revender Bitcoins
- Assim, se você comprar agora (dia 736), quando o valor do Bitcoin aumentar ainda mais, você poderá recuperar seu investimento e ter um lucro
- Para avaliar essa decisão, precisamos primeiramente estimar quanto podemos ganhar de lucro potencial
- Ou seja: precisamos descobrir se o retorno do investimento será adequado no curto prazo

Outro exemplo

Previsão do Preço do Bitcoin



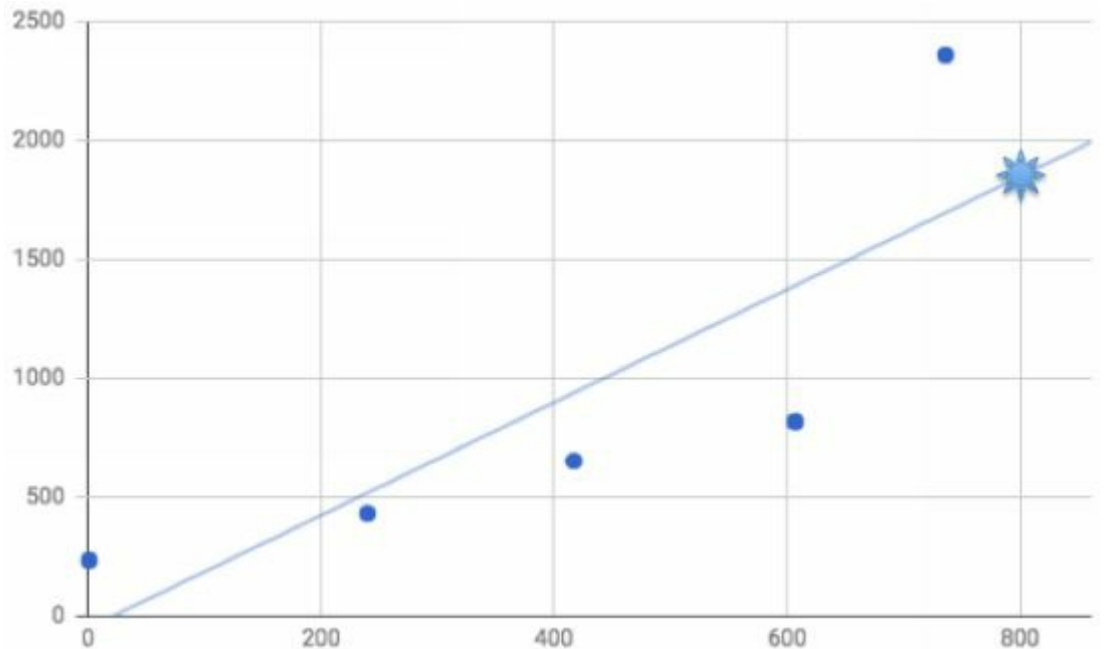
Outro exemplo

Previsão do Preço do Bitcoin

- Note que, à medida que uma variável aumenta, a outra variável aumentará no valor médio indicado pelo hiperplano
- Assim, se desejamos estimar o valor do Bitcoin em 800 dias, podemos inserir 800 como sua coordenada x e referenciar a inclinação encontrando o valor Y correspondente representado no hiperplano
  - Nesse caso, o valor de Y é \$1850.00

Outro exemplo

Previsão do Preço do Bitcoin



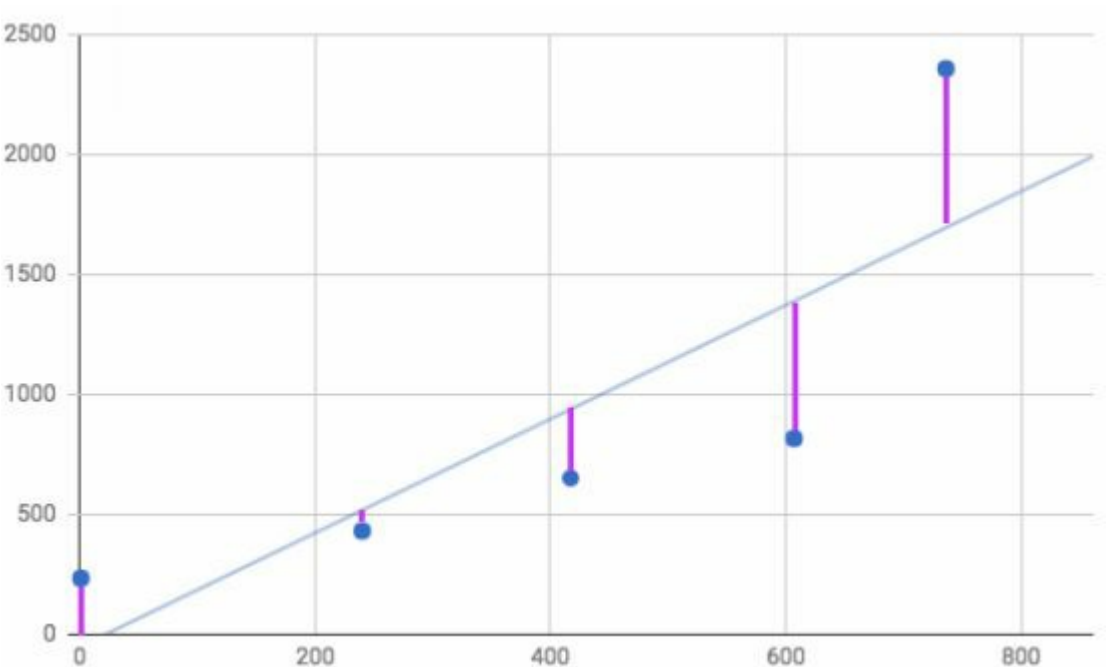
Outro exemplo

Previsão do Preço do Bitcoin

- para escolher tendências de investimento
- A linha de tendência oferece um ponto de referência básico para prever o futuro
- Se usássemos a linha de tendência como ponto de referência mais cedo, por exemplo, no dia 240, a previsão publicada teria sido mais precisa
- No dia 240 há um baixo grau de desvio do hiperplano, enquanto no dia 736 há um alto grau de desvio
  - O desvio refere-se à distância entre o hiperplano e o ponto de dados

Outro exemplo

Previsão do Preço do Bitcoin



Leitura Recomendada I

Esse material foi baseado, principalmente, nos trabalhos de Almeida et al., 2020; Izbicki and dos Santos, 2020; Theobald, 2017

- Almeida, A., Carvalho, F., & Menino, F. (2020). ***Introdução ao machine learning.*** Grupo DataAt.
- Izbicki, R., & dos Santos, T. M. (2020). ***Aprendizado de máquina: Uma abordagem estatística.*** Livro eletrônico.
- Theobald, O. (2017). ***Machine learning for absolute beginners.*** Livro eletrônico.