

Curso de Data Science



Aula 10 - Regressão Linear

O que você irá aprender nesta aula?

- Boston Housing dataset
- Regressão Linear

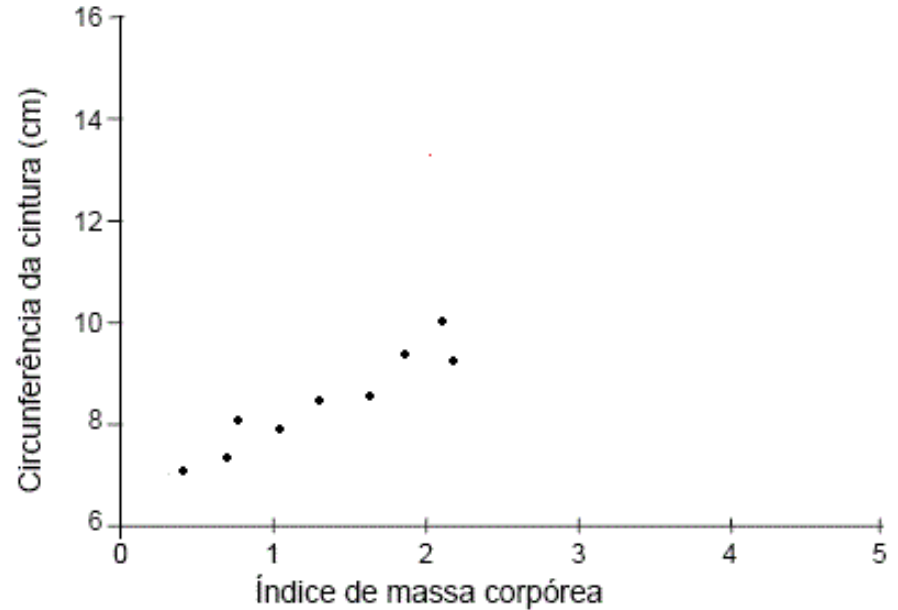
Regressão Linear

Regressão: queremos encontrar um valor

Classificação é sobre prever um rótulo; regressão é sobre prever uma quantidade

Regressão Linear

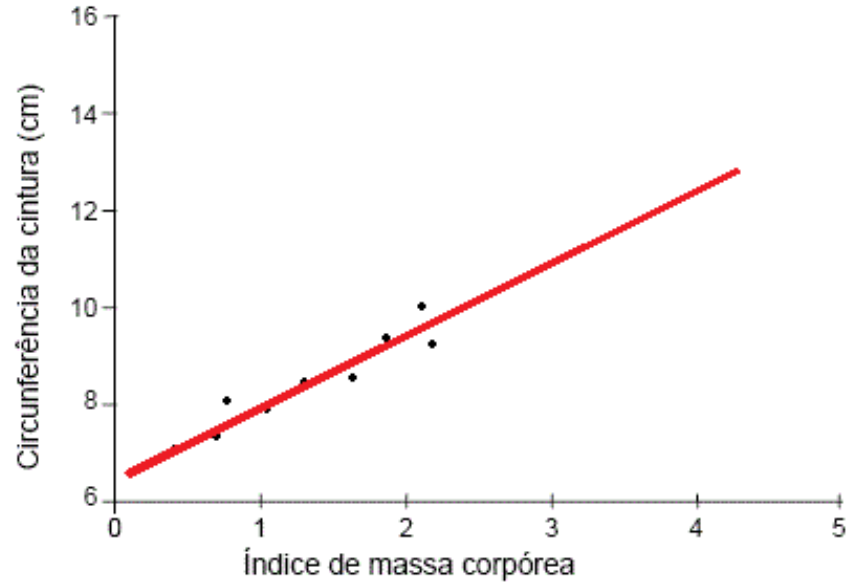
Qual seria o valor de circunferência esperado para um índice de massa corpórea de 4?



Regressão Linear

Qual seria o valor de circunferência esperado para um índice de massa corpórea de 4?

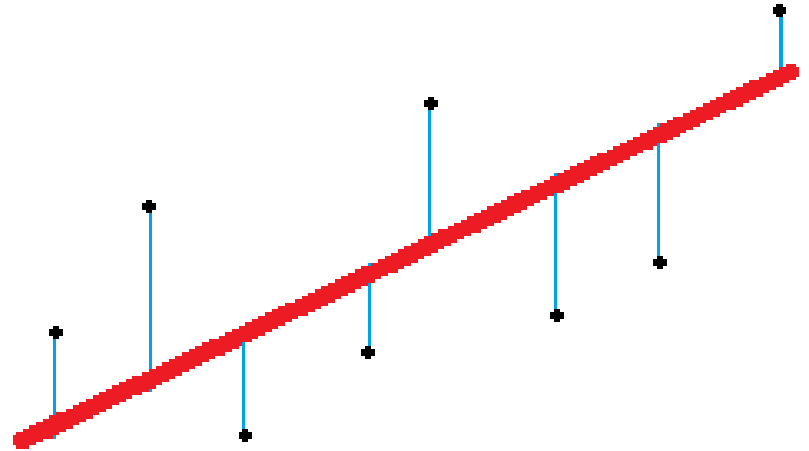
Uma reta parece acomodar bem os dados



Regressão Linear

O computador usa um algoritmo para determinar a equação da reta que nos permite fazer a previsão do valor.

Método dos mínimos quadrados ordinários: encontra reta que minimiza as distâncias até os pontos:
https://pt.wikipedia.org/wiki/M%C3%A9todo_dos_m%C3%A9todos_quadradados



Regressão Linear

Fórmula:

$$y = b_0 + b_1 * x_1$$

Diagram illustrating the components of the linear regression formula:

- y is labeled as the **Dependent variable (DV)**.
- b_0 is labeled as the **Coefficient**.
- b_1 is labeled as the **Coefficient**.
- x_1 is labeled as the **Independent variable (IV)**.

Em Python (scikit-learn):

In []:

```
1 lm.coef_
```

In []:

```
1 lm.intercept_
```


Regressão Linear

- Coeficiente de determinação, também chamado de R^2 .
- Medida estatística de quão próximos os dados estão da linha de regressão ajustada
- Varia entre 0 e 1

Em Python (scikit-learn):

```
In [ ]: 1 lm.score(X,y)
```

Sobre o problema

Boston Housing Dataset

- UCI Machine Learning Repository (University of California, Irvine)
- Coletado em 1978
- 14 informações sobre 506 casas na cidade de Boston, Massachusetts, nos Estados Unidos.



Sobre o problema

Queremos encontrar alguma forma de prever o preço de uma casa, baseado nas suas características.

Sobre o problema

- CRIM: per capita crime rate
- ZN: proportion of residential land zoned
- INDUS proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per \$10,000
- PTRATIO: pupil-teacher ratio by town
- B: the proportion of blacks by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in \$1000's

Então, nesta aula vimos:

- Boston Housing dataset
- Regressão Linear

Muito obrigado!