Supervised Learning

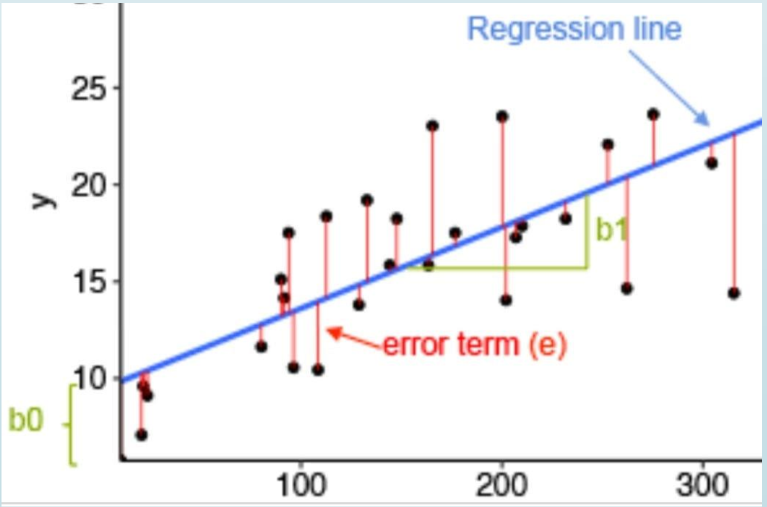
Regression

Sadegh Sulaimany

University of Kurdistan

www.bioinformation.ir

Spring 2024



Outline

- Introduction
- Background
- Example
- Linear regression
- Calculating best coefficients
- Multiple linear regression
- Gradient Descent
- Logistic Regression

Introduction

- Most commonly used machine learning method
- Fitting data with functions or **function fitting**
  - predict the value (or class) of a dependent attribute $\mathbf{y}$,

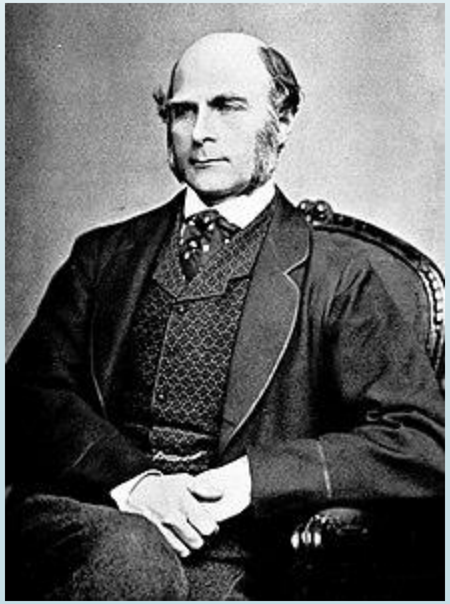
    by combining the predictor attributes $\mathbf{X}$ into a function

$$y = f(\mathbf{X})$$

  - There are several function fitting techniques
    - Prevalent one:
      1. **Linear regression for numeric prediction**
      2. **Logistic regression for classification**

Background

- Relatively old technique dating back to the Victorian era
  - 1830s to the early 1900s
- Pioneer
  - Francis Galton
    - concept of **regressing toward the mean**
    - systematically comparing children's heights against their parents' heights



Introduction

- Theoretical framework for the simplest of function-fitting methods:
  - the linear regression model
- main focus will be on a case study that demonstrates
  - how to build regression models

- First challenge
  - curse of dimensionality
  - As the number of predictors $\mathbf{X}$, increases,
    - ability to obtain a good model reduces
    - computational and interpretational complexity increases

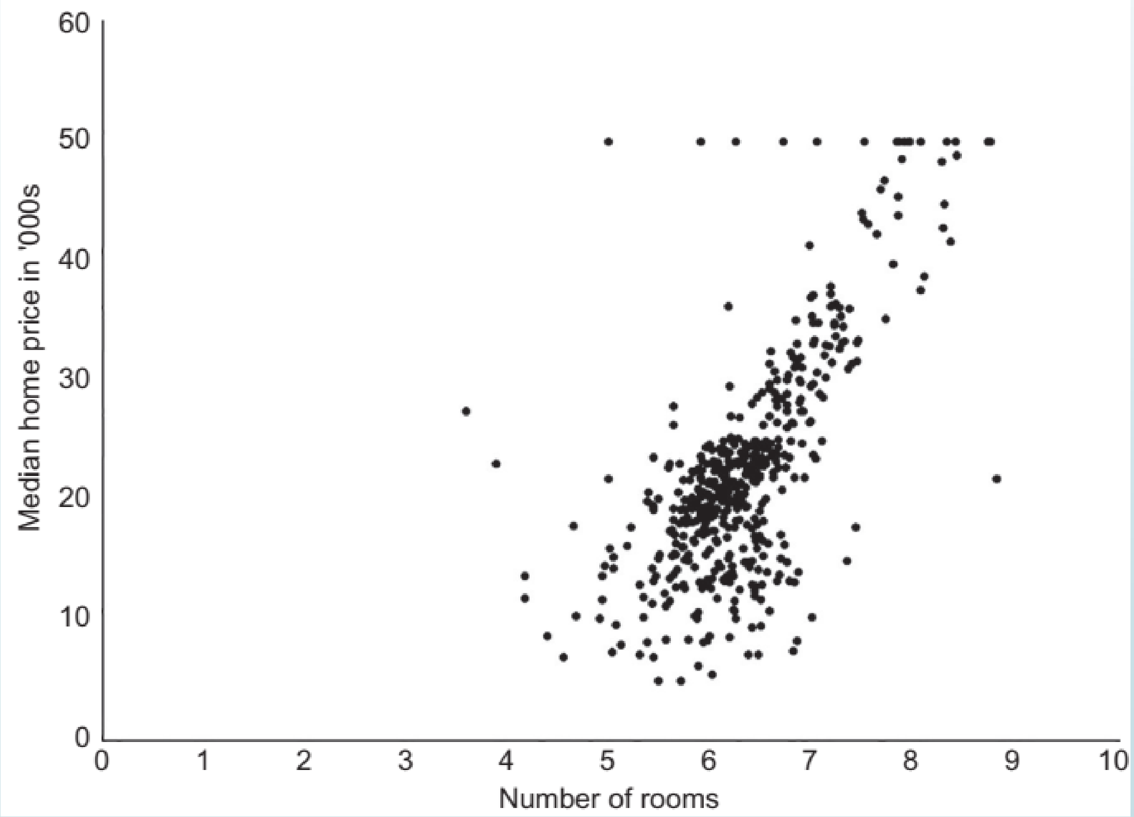
Feature Selection

Introductory example

- Knowing the effect of
  - the number of rooms in a house (**predictor**)
    on its median sale price (**target**)

- Overall trend
  - increasing the number of rooms
    tends to also increase median price

- **Linear regression**
  - finding a line (or a curve)
  - that best explains this tendency



Linear regression

- Visualization is difficult for more than two predictor
  - General statement
    - the dependent variables are expressed as a linear combination of independent variables

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

    - Problem statement
      - **Finding the best fitting line (best fitting function)**
      - Evaluation?
        - **Error function**

Linear regression

- Minimizing the error function

  - $\hat{y}$ is actual target
  - y is predicted target
  - $\boldsymbol{e}$ is error
  - $\mathbf{J}$ is the total squared error

$$\hat{y} = b_0 + b_1 x$$

$$e = y - \hat{y} = y - (b_0 + b_1 x)$$

$$\frac{J}{n} = \frac{\sum e^2}{n} = \frac{\sum (y_i - \dot{y}_i)^2}{n} = \frac{\sum (y_i - b_0 - b_i x_i)^2}{n}$$

Calculating the best coefficients

- Finding the best combination of $(b0, b1)$
  - Minimizing the total error
  - Taking partial derivatives of $\mathbf{J}$ with respect to $\boldsymbol{b1}$ and $\boldsymbol{b0}$ and set them equal to zero

Calculating the best coefficients

$$\frac{J}{n} = \frac{\sum e^2}{n} = \frac{\sum (y_i - \dot{y}_i)^2}{n} = \frac{\sum (y_i - b_0 - b_i x_i)^2}{n}$$

$$\partial J/\partial b_1 = \partial J/\partial \hat{y} \quad \partial \hat{y}/\partial b_1$$
$$\Rightarrow \partial J/\partial b_1 = 2(\Sigma(y_i - b_0 - b_1 x_i))\partial \hat{y}/\partial b_1 = 0$$
$$\Rightarrow \Sigma(y_i - b_0 - b_1 x_i)(-x_i) = 0$$
$$\Rightarrow -\Sigma(y_i x_i) + \Sigma(b_0 x_i) + \Sigma(b_1 x^2{}_i) = 0$$
$$\Rightarrow \Sigma(y_i x_i) = b_0 \Sigma(x_i) + b_1 \Sigma(x^2{}_i)$$

$$\partial J/\partial b_0 = 2(\Sigma(y_i - b_0 - b_1 x))\partial \hat{y}/\partial b_0 = 0$$
$$\Rightarrow \Sigma(y_i - b_0 - b_1 x_i)(-1) = 0$$
$$\Rightarrow -\Sigma(y_i) + \Sigma(b_0.1) + \Sigma(b_1 x_i)1 = 0$$
$$\Rightarrow -\Sigma(y_i) + b_0 \Sigma(1) + b_1 \Sigma(x_i) = 0$$
$$\Rightarrow \Sigma(y_i) = b_0 N + b_1 \Sigma(x_i)$$

Calculating the best coefficients

- two equations in two unknowns b_0 and b_1
- can be further simplified and solved to yield the expressions

$$b_1 = (\Sigma x_i y_i - \bar{y}\Sigma x_i)/(\Sigma x^2{}_i - \bar{x} - \Sigma x_i)$$

$$b_0 = (\bar{y}\Sigma x^2{}_i - \bar{x}\Sigma x_i y_i)/(\Sigma x^2{}_i - \bar{x}\Sigma x_i)$$

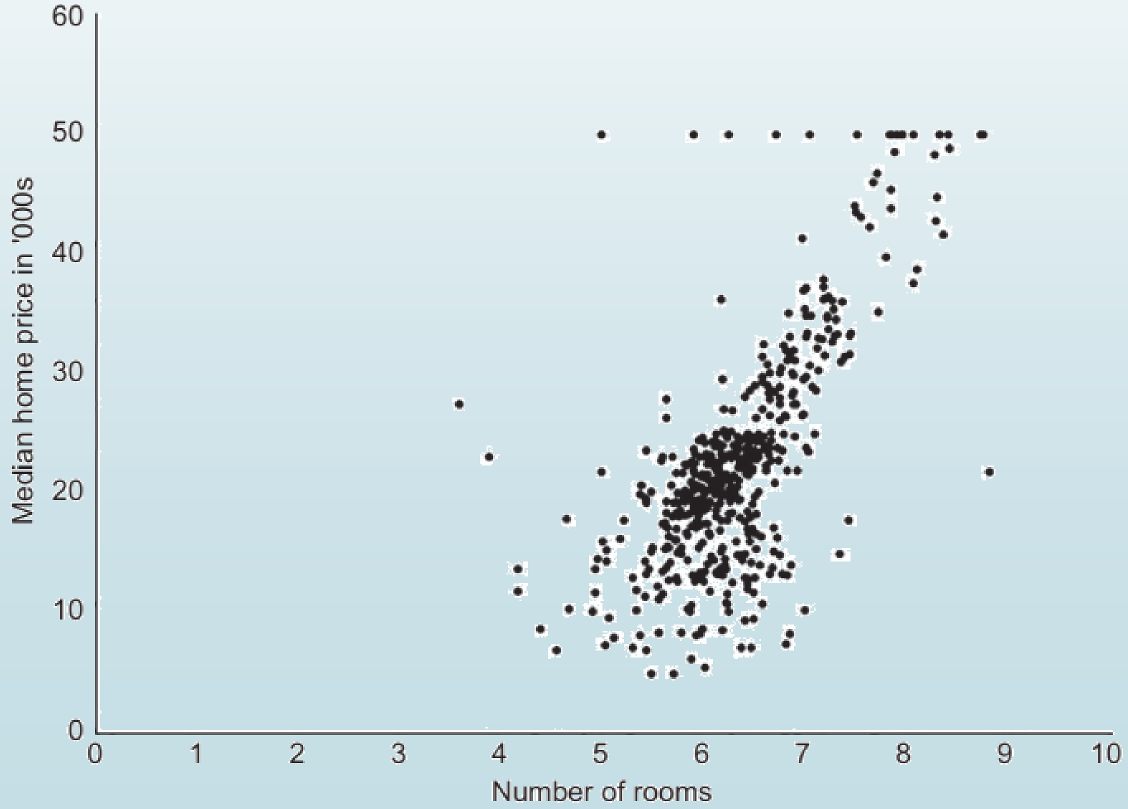
$$b_1 = \text{Correlation } (y, x) \times \frac{s_y}{s_x}$$

?

$$b_0 = y_{\text{mean}} - b_1 \times x_{\text{mean}}$$

Multiple Linear Regression

- For some of the points (houses) at the top of the chart, where median price=50
  - the median price appears to be independent of the number of rooms!
  - This could be because there may be other factors that also influence the price
  - more than one predictor will need
  - to be modeled
  - **Multiple linear regression (MLR),**
  - an extension of
    - simple linear regression



$$\text{Median price} = 9.1 \times (\textit{number of rooms}) - 34.7$$

Multiple Linear Regression

$$\hat{y}_i = b_0 + b_1 x_1 + \ldots + b_D x_D \qquad x = [x_0, x_1, \ldots, x_D]$$

$$E = \sum_{i=1}^{N} (y_i - B^T x_i)^2$$

- B^T is the vector of weights $[b_0, b_1, \ldots, b_D]$
- take a derivative of **E** with respect to each weight **B**
  - will end up with **D** equations to be solved for **D** weights
    - one corresponding to each feature

Multiple Linear Regression

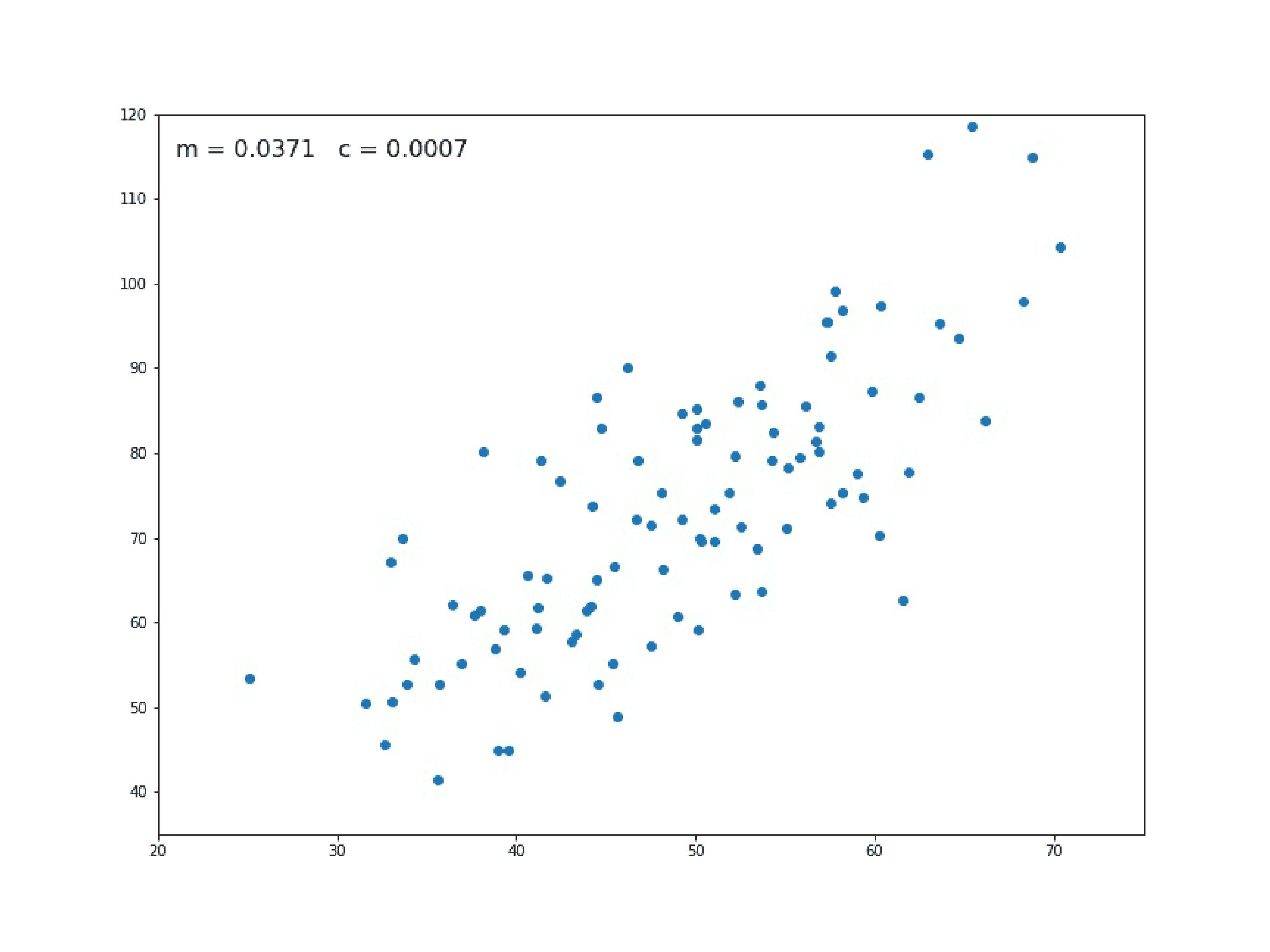
- Partial derivative for each weight

$$\partial E/\partial b_j = \partial E/\partial \hat{y} * \partial \hat{y}_i/\partial b_j$$
$$\Rightarrow \partial E/\partial b_j = 2\Sigma(y_i - B^T x_i)\partial \hat{y}_i/\partial b_j$$
$$\Rightarrow \partial E/\partial b_j = 2\Sigma(y_i - B^T x_i)(-x_i)$$
$$\Rightarrow \Sigma y_i(-x_i) - B^T \Sigma(x_i)(-x_i)$$

- Simple writing for D weights in matrix form
  - B is a $1 \times D$ matrix or vector
  - set this derivative to zero, to solve for the weights

$$\partial E/\partial B = -(Y^T X) + B(X^T X)$$
$$-(Y^T X) + B(X^T X) = 0$$

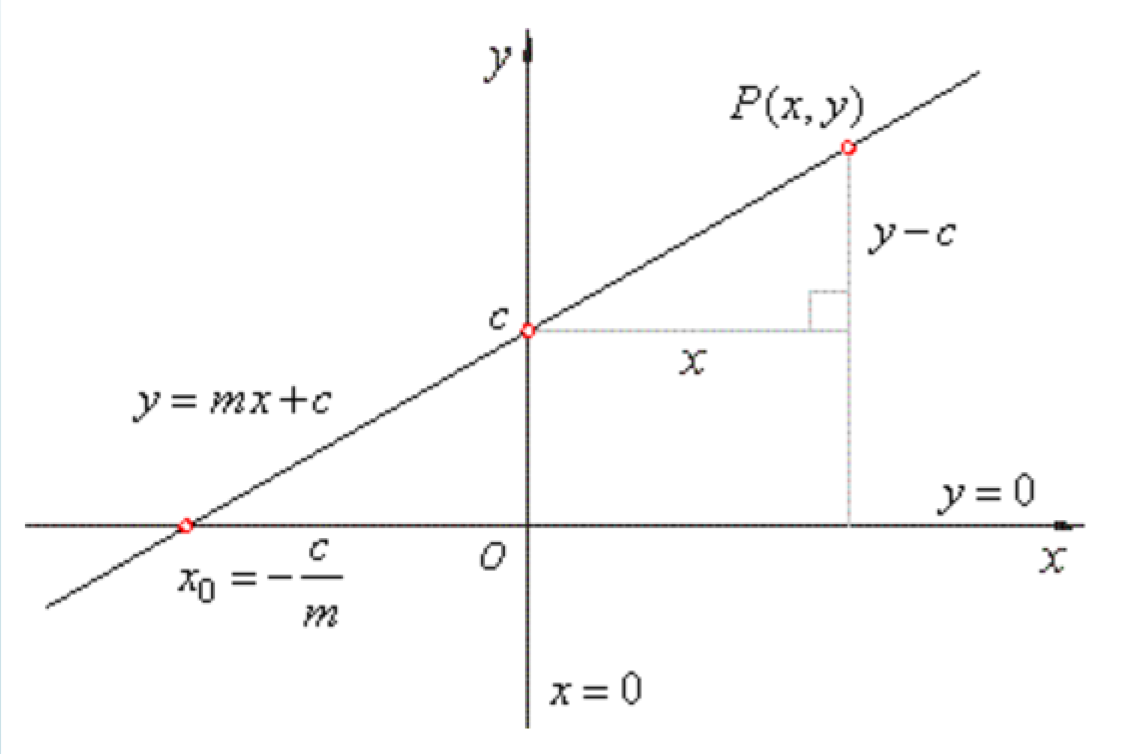
Linear Regression with Gradient Descent



Review

- In statistics, **linear regression** is
  - a linear approach to modelling the relationship between a dependent variable and one or more independent variables.
    - Let **X** be the independent variable and **Y** be the dependent variable

$$Y = mX + c$$



- **m** is the slope
- **c** is the y intercept.

Problem definition

- we use this equation to train our model with a given dataset
- and predict the value of Y for any given value of X.
  - The challenge is to determine the value of **m** and **c**,
  - such that **the line** corresponding to those values **is the best fitting** line or gives the **minimum error**.

$$Y = mX + c$$

- **Loss Function**
- The loss is the error in our predicted value of m and c.
- Our **goal is to minimize this error** to obtain the most accurate value of **m** and **c**.
  - Mean Squared Error Equation
    - **y_i** is the actual value and **$\bar{y}_i$** is the predicted value

$$E = \frac{1}{n} \sum_{i=0}^{n} (y_i - \bar{y}_i)^2$$

- Substituting the value of **$\bar{y}_i$**

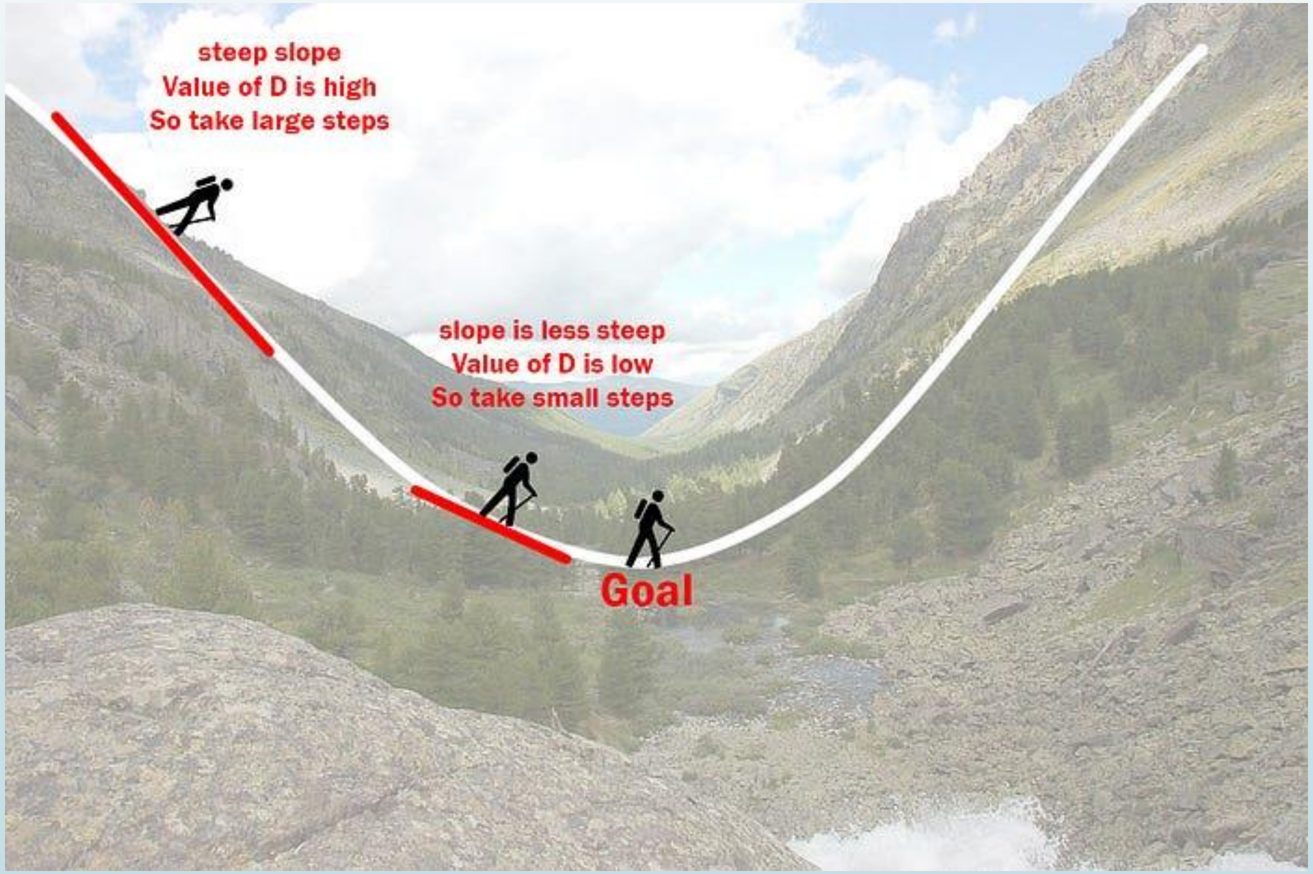
$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

1. defining the loss function
2. minimizing it
3. finding the associated **m** and **c**

Gradient Descent Algorithm

- is an iterative optimization algorithm
  - to **find the minimum of a function**.
  - Here: The Loss Function.

  - one of the simplest and widely used algorithms in machine learning,
  - mainly because it can be applied to any function to optimize it.
  - Learning it lays the foundation to mastering machine learning.



Calculating gradient descent

1. Initially let **m = 0** and **c = 0**.
   - And Let **L** be our learning rate.
   - This controls how much the value of **m** changes with each step.
   - **L** could be a small value like 0.0001 for good accuracy.
2. Calculate the **partial derivative** of the loss function
   - with respect to **m**, and **c**

$$D_m = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$$

$$D_c = \frac{-2}{n}\sum_{i=0}^{n} (y_i - \bar{y}_i)$$

Calculating gradient descent

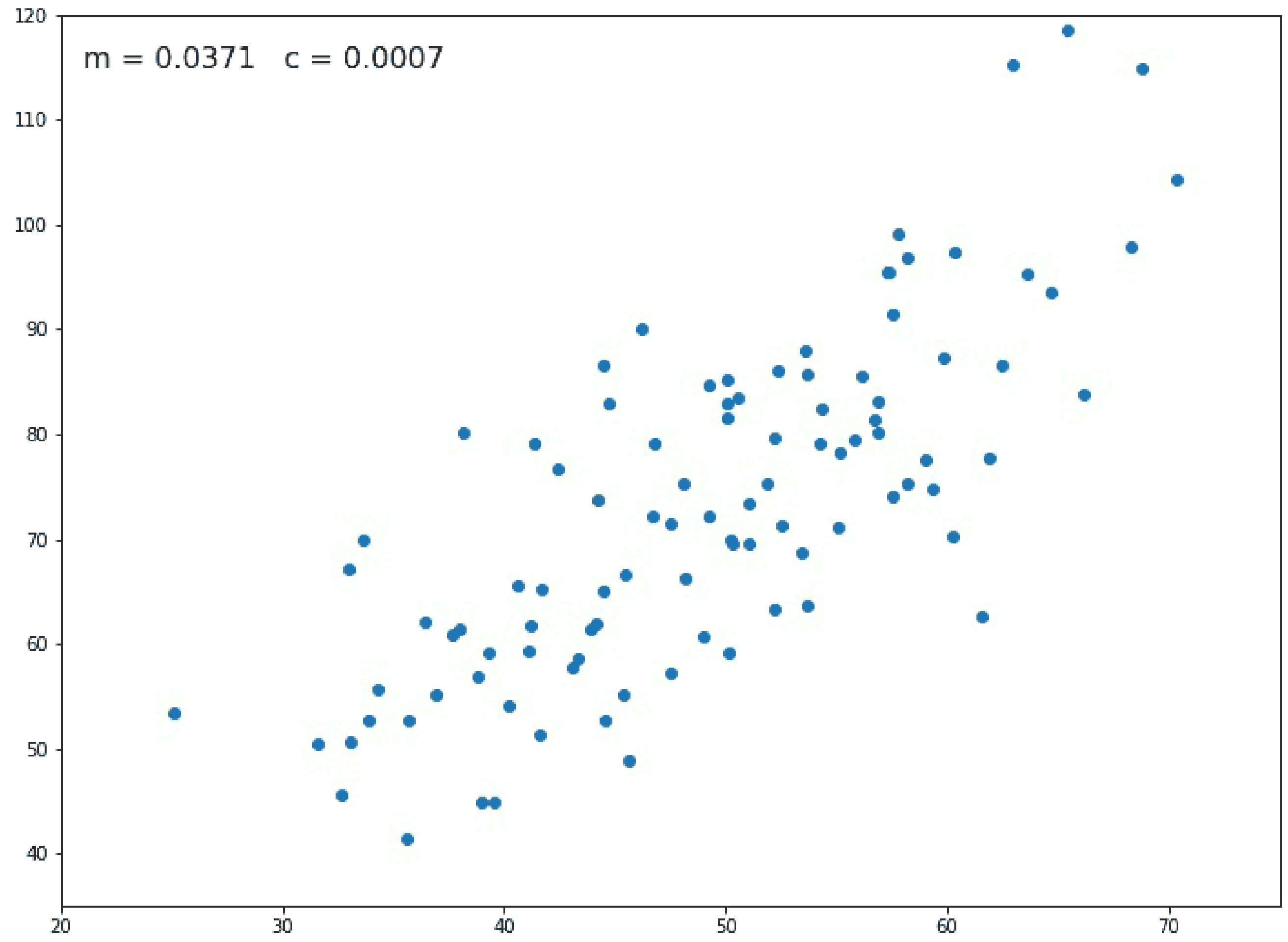
3. Update the current value of **m** and **c** using the following equation (affecting **L** coefficient):

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

3. Repeat this process until our loss function is a very small value or ideally 0
   - (which means 0 error or 100% accuracy).
   - The value of **m** and **c** that we are left with now will be the optimum values.

Sample Implementation



Logistic Regression

Logistic regression: 0 or 1?

- A method of **classification**:
  - **outputs the probability of** a categorical target variable **Y belonging to a certain class**.
  - often used for binary classification
- How to evaluate the model?
  - least squares error?
    - **No**
  - assigning a probability between 0% and 100%
    - that Y belongs to a certain class
  - Classification evaluation
    - Confusion matrix

How does Logistic regression works?

- a modification of linear regression
  - that makes sure to output a probability between 0 and 1
  - by applying the **sigmoid function**
  - when graphed
    - looks like the characteristic **S-shaped curve**

$$S(x) = \frac{1}{1 + e^{-x}}$$

    - Original form of linear regression

$$g(X) = \beta_0 + \beta_1 x + \epsilon$$

- Think of function that converting the model output
  - to a value in the [0,1] range

How does Logistic regression works?

- Calculating the probability
  - that the training example belongs to a certain class: P(Y=1)

$$P(Y=1) = F(g(x)) = \frac{1}{1+e^{-(\beta_0+\beta_1 * x)}}$$

  - Isolating **p** (probability that **Y=1**)

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x + \epsilon$$

  - Solving the linear regression is equal to calculating the
    - **odds ratio**, or **logit model**

How does Logistic regression works?

- The log-odds of **y** being **1**
  - is a linear combination of one or more predictor variables, according to the logistic model.
  - Let's say we have two predictors or independent variables, x1 and x2, and p is the probability of y equaling 1.
    - Then, using the logistic model as a guide:

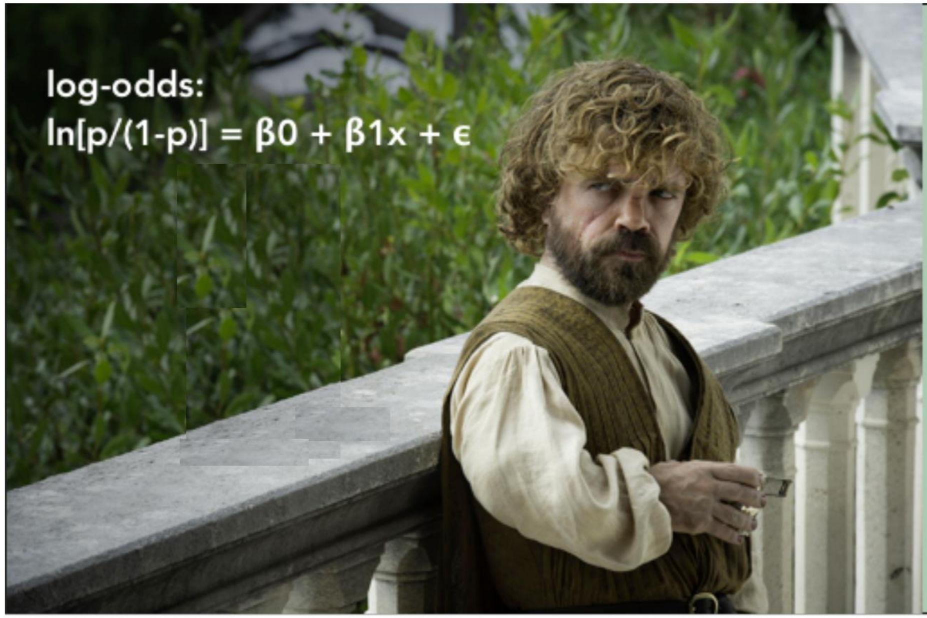
$$\ln \frac{p}{1-p} = a + bx_1 + cx_2$$

How does Logistic regression works?

- log-odds ratio
  - natural log of the odds ratio (**p/(1-p)**)

*"Yo, what do you think are the **odds** that Tyrion Lannister dies in this season of Game of Thrones?"*

*"Hmm. It's definitely 2x more likely to happen than not. **2-to-1 odds**. Sure, he might seem too important to be killed, but we all saw what they did to Ned Stark…"*



<— IS HE GONNA DIE?

p = P(Tyrion dies) = 2/3

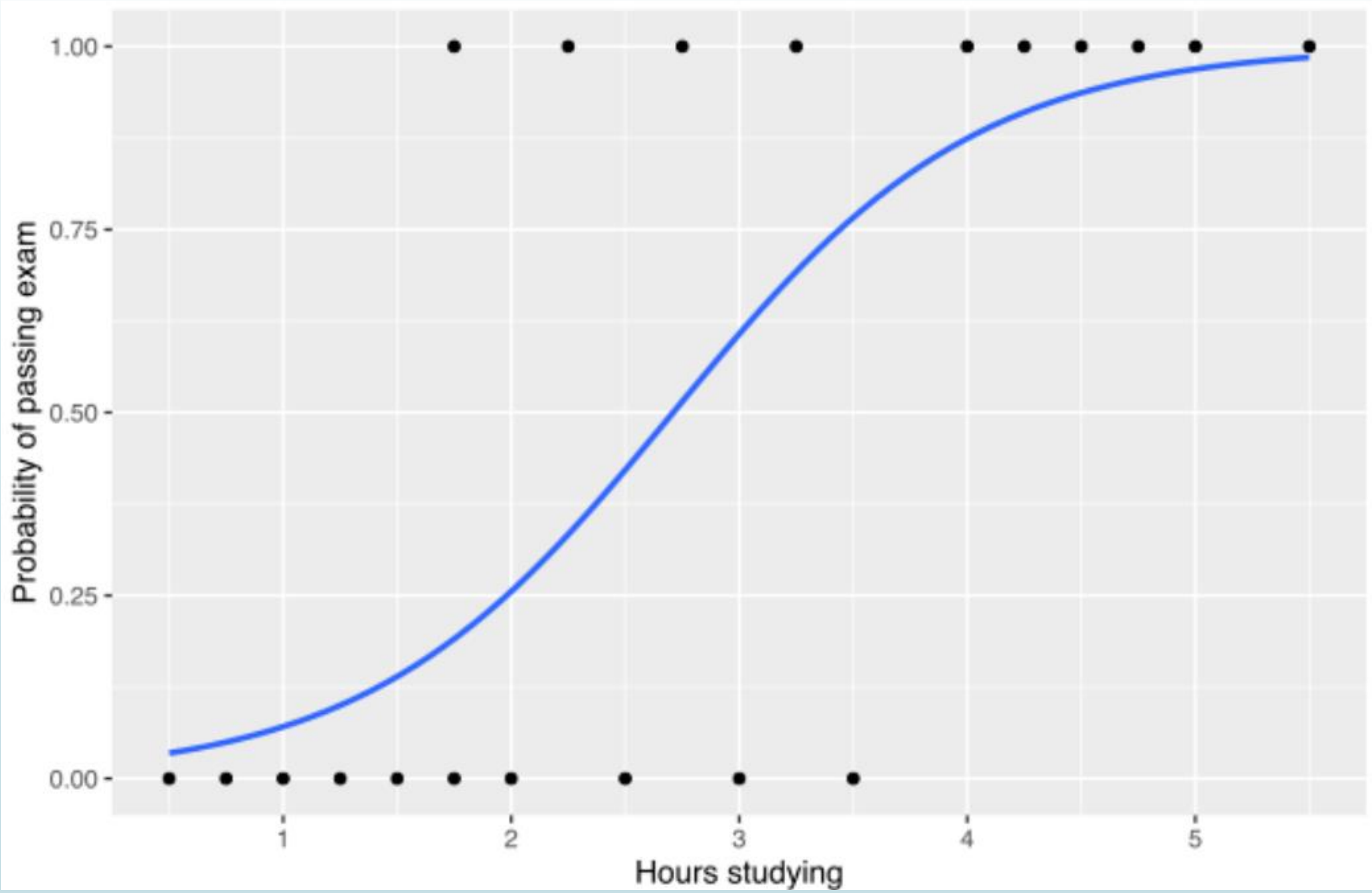
1-p = P(Tyrion doesn't die) = 1/3

odds ratio: p/(1-p) = 2.0
"He's gonna die. 2-to-1 odds"

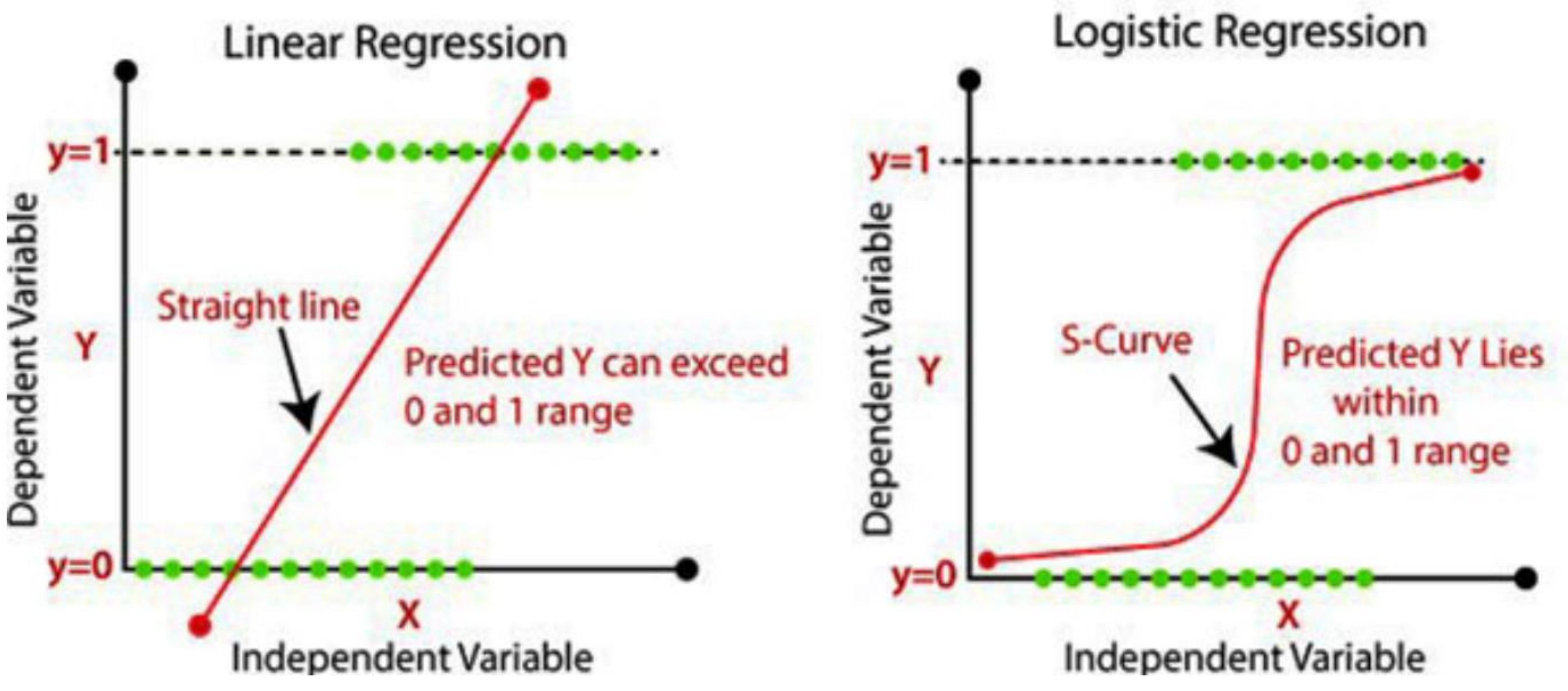
log-odds ratio: ln[p/(1-p)] = 0.693
"He's gonna die. .693 log-odds"

Example

- Probability of passing exam vs hours of studying
  - Output looks like an S-curve showing P(Y=1) based on the value of X



Comparison


Linear Regression
y=1
Dependent Variable
Y
Straight line
Predicted Y can exceed
0 and 1 range
y=0
X
Independent Variable
Logistic Regression
y=1
Dependent Variable
Y
S-Curve
Predicted Y Lies
within
0 and 1 range
y=0
X
Independent Variable

Threshold for logistic regression

- To predict the Y label
  - spam/not spam
  - cancer/not cancer
  - fraud/not fraud
  - …

- We have to set a probability cutoff, or threshold,
  - for a positive result.
  - For **example**:
    - "If our model thinks the probability of this email being spam is higher than 70%, label it spam.
    - Otherwise, don't."

Threshold for logistic regression

- Depends on your tolerance for **false positives** vs. **false negatives**.
  - Example
    - In diagnosing cancer,
    - We'd have a very low tolerance for false negatives,
    - because even if there's a very small chance the patient has cancer, you'd want to **run further tests to make sure**.
    - So you'd set a very low threshold for a negatives result

    - In the case of fraudulent loan applications,
    - the tolerance for false positives might be higher, particularly for smaller loans, since further vetting is costly
    - and a small loan may not be worth the additional operational costs processing.

New Homework

- برای مقاله انتخابی، در صورتی که از Classification استفاده کرده است، شما نیز همان روش و الگوریتم‌های ایشان را بر دیتای مقاله اجرا کنید و همان نتایج را ایجاد نمایید و گزارش دهید که آیا تفاوت قابل توجه در نتایج حاصل شده وجود دارد یا خیر؟ اگر تفاوت قابل توجه وجود دارد، دلایل احتمالی را شرح دهید.
- چه آموخته‌های اضافی یادگرفته‌اید یا می‌دانید (مانند پیش‌پردازش یا بالانس کردن داده‌ها یا به کارگیری سایر روش‌ها از قبیل روش‌های Ensemble یا ...) که می‌تواند در بهبود نتایج مفید باشد؟
- اگر مقاله انتخابی شما از Regression استفاده کرده است، روند بالا را برای آن انجام دهید و علاوه بر آن، بر دیتاست دیابت که در کلاس حل تمرین استفاده کردید، روش‌های Classification را اجرا کنید و عملکرد آن‌ها را از نظر Accuracy و F1-Measure مقایسه کنید و روش برتر را گزارش نمایید.

Thanks

Social & Biological Network Analysis Laboratory