Our First Model: Linear Regression

TM Quest

Overview

Overview

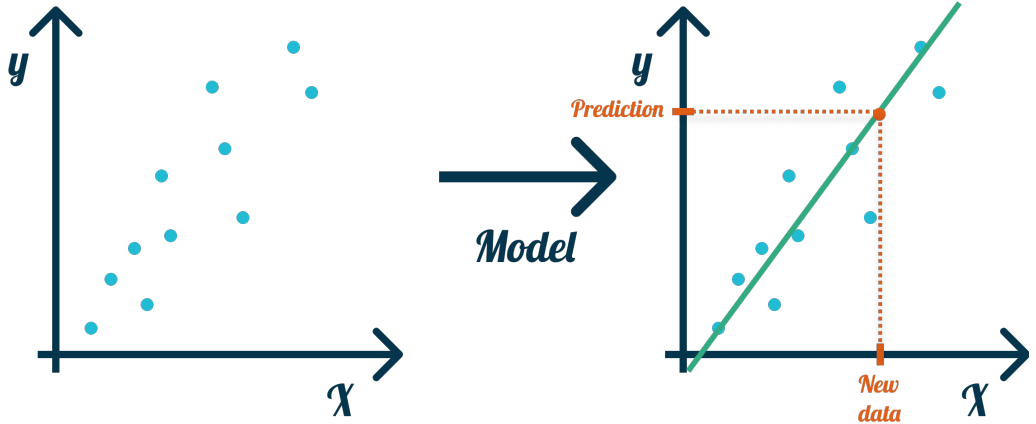
What Will we Learn in This Module?

- Linear Regression
  - What is linear regression?
  - How to train a linear regression model?
  - How to predict on new data using our model?
- Evaluating the Model
  - How can we tell if our model is good?
  - What is the mean square error?
  - How to evaluate our model?
- Test Set and Training Set
  - What is training and test sets?
  - What is the purpose of training and test sets?
  - How to split into training and test sets?

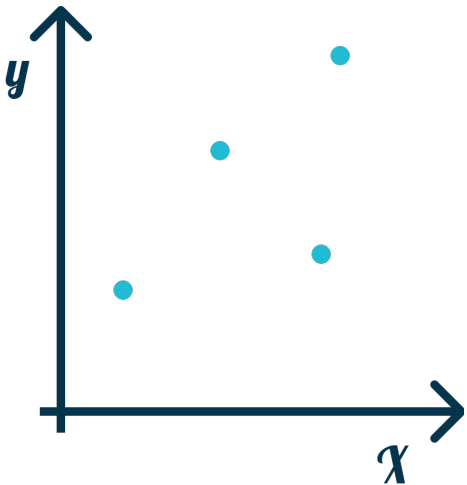
Linear Regression

What is Linear Regression?

- Is a regression model (gives our real numbers)
- Is linear

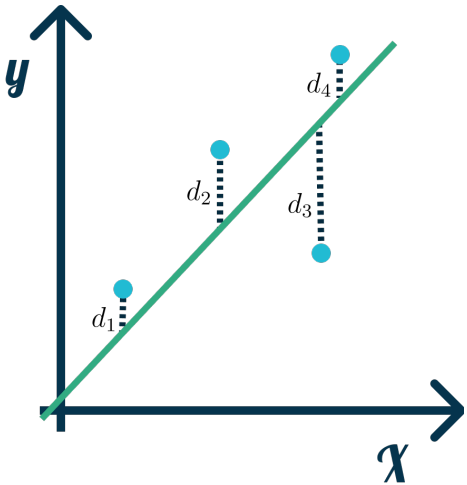


The Idea



In this case we have 4 data points and 1 feature.

The Idea



Want to find the unique line such that

$$d_1^2 + d_2^2 + d_3^2 + d_4^2$$

is as small as possible!

Finding the best line is called training.

Linear Regression Model

Some Notation

- $n + 1$ observations
- $p + 1$ features

Target Observations

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Feature Observations

$$\mathbf{X} = \begin{pmatrix} x_{00} & x_{01} & \cdots & x_{0p} \\ x_{10} & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Features

Observations

Example

Example

- Assume that we measure BMI and age of 4 people and want to predict the average blood sugar level.
- 4 Observations
- 2 Features
- Regression Problem

$$\mathbf{y} = \begin{pmatrix} 4.2 \\ 8.3 \\ 5.2 \\ 7.3 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 23 & 39 \\ 29 & 43 \\ 21 & 22 \\ 18 & 65 \end{pmatrix}$$

The Model

- $\mathbf{x}_i$ be the i'th observation of $\mathbf{X}$ with corresponding target y_i.
- Then the model is

$$\hat{y}_i = \mathbf{a} \cdot \mathbf{x}_i + b = a_0 x_{i\,0} + a_1 x_{i\,1} + \cdots + a_p x_{i\,p} + b,$$

  where $\mathbf{a} = (a_0, a_1, \ldots, a_p)$ and b are constant.
- When we have 1 feature, then

$$\hat{y}_i = a_0 \cdot x_{i\,0} + b.$$

- The training step finds $\mathbf{a}$ and b such that

$$\sum_{i=0}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=0}^{n} d_i^2$$

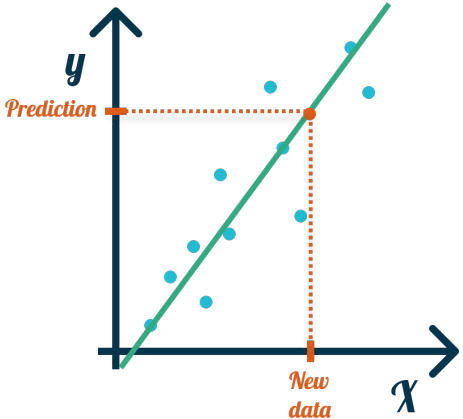
  is as small as possible.

New Predictions

Let $\mathbf{x} = (x_0, x_1, \ldots, x_p)$ be a new observation. Then the predicted value $\hat{y}$ is given by

$$\hat{y} = \mathbf{a} \cdot \mathbf{x} + b,$$

for the values of $\mathbf{a}$ and b found in the training step.



Example

Example

Using the data

$$\mathbf{y} = \begin{pmatrix} 4.2 \\ 8.3 \\ 5.2 \\ 7.3 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 23 & 39 \\ 29 & 43 \\ 21 & 22 \\ 18 & 65 \end{pmatrix}$$

we have

$$\mathbf{a} = (0.219, 0.073) \quad \text{and} \quad b = -1.802.$$

Hence

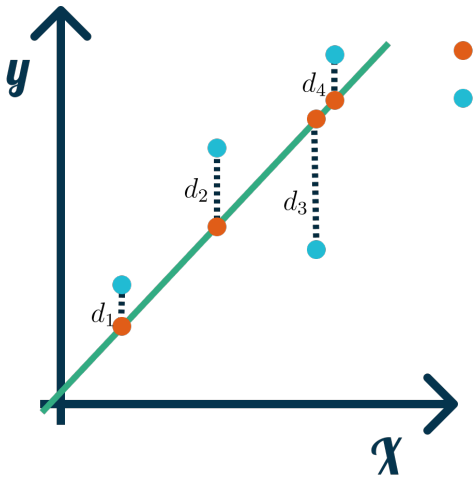
$$\hat{y} = 0.219x_0 + 0.073x_1 - 1.802.$$

If we have a new patient with $\mathbf{x} = (25, 30)$ we have that

$$\hat{y} = 0.219 \times 25 + 0.073 \times 30 - 1.802 = 5.863.$$

Measuring the Error

Mean Square Error



Predicted Value

Actual Value

Definition

The Mean Square Error (MSE)

$$MSE = \frac{1}{m}\sum_{j=0}^{m-1}(\hat{y}_j - y_j)^2.$$

Mean Square Error

Example

Model:

$$\hat{y} = 0.219x_0 + 0.073x_1 - 1.802.$$

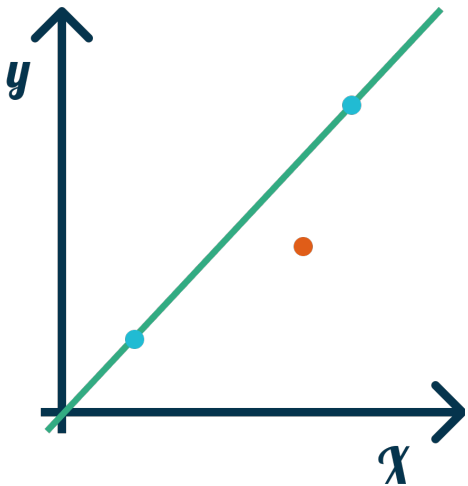
Let us say that we have the two new data points:

- $\mathbf{x}_0 = (25, 30)$ with $y_0 = 5.0$,
- $\mathbf{x}_1 = (30, 39)$ with $y_1 = 8.0$.

Using the model gives $\hat{y}_0 = 5.849$ and $\hat{y}_1 = 7.599$.

$$\text{MSE} = \frac{1}{2}\sum_{j=0}^{1}(\hat{y}_j - y_j)^2 = \frac{1}{2}((5.849 - 5)^2 + (7.599 - 8)^2) = 0.442$$

Data Leakage



Data Leakage

- Measuring the error on the training set will result in lower error than on new data.
- The training set and the test set must be distinct. Also, the training set should not influence the test set at all.
- If not, we say that data leakage has occurred.

Test Set

Definition (Test Set)

The test set is a subset of the dataset used to test the model using a metric (e.g. MSE).

Train-Test Split

Given a data set X and y, train-test split divides it randomly into two parts:

- Training set, used for training the model.
- Testing set, used to testing the model.

One usually set a higher percentage of data in the training set than in the testing set, e.g. 70% training, 30% testing.

Train-Test Split