Regression Overview

Tozammel Hossain

Data Science & Analytics
University of Missouri

What is a regression Problem?

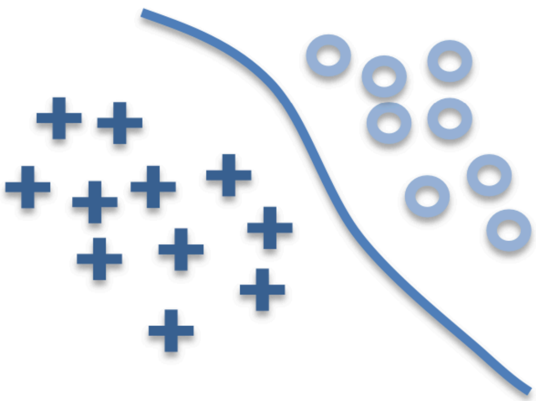
- **A subcategory supervised learning**
- **A way of measuring the relationship between two or more variables**
- **Task: predicting a continuous variable**
  - Fit a curve given a set of data points
- **One feature**
  - Linear Regression
- **More than one features**
  - Multiple linear regression
  - or Multivariate linear regression

Regression Setting

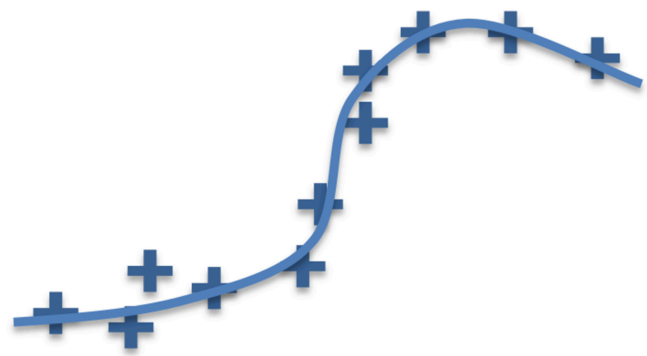
- **Given**
  - A dataset D
    - D = $\{X_1,X_2,..\}$ is a set of rows/instances/examples
    - Each instance in D is described by values for a set of features/attributes X = $\{x_1,x_2,...\}$
    - Each instance in D is associated with a real number, y = $\{y_1,y_2,...\}$; So, y is a continuous variable
  - Learning
    - y = f(input instance)
  - Predict
    - target value for new instances
- **Classification vs Regression**
  - Predicting discrete vs Continuous

x_1	x_2	...	x_n	y
				y_1
				y2
				y2
				Y3
				y1

Difference between Classification & Regression



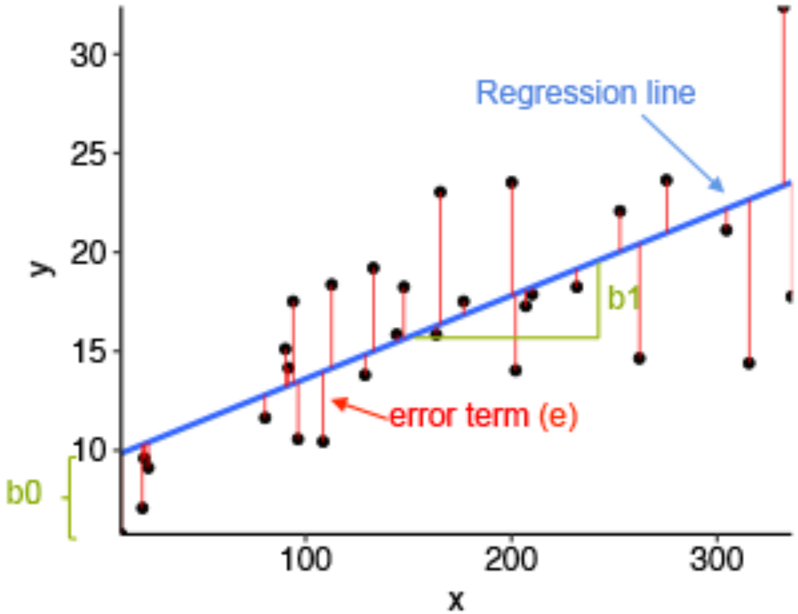
Classification



Regression

source: https://zenodo.org/record/4429576#.YFUz3GRufJ8

Linear Regression



Assumptions:

- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X
- **Independence**: Observations are independent of each other
- **Normality**: For any fixed value of X, Y is normally distributed

How to fit the line?

$$min_b \|Xb - y\|^2$$

LR: $y_i = b_0 + b_1 x_1^i + e$

MLR: $y_i = b_0 + b_1 x_1^i + b_2 x_2^i + \ldots + b_n x_n^i + e$

$$e \sim N(0, \sigma^2)$$

Linear Regression

sklearn.linear_model.LinearRegression

class `sklearn.linear_model.`**`LinearRegression`**(**, fit_intercept=True, normalize=False, copy_X=True, n_jobs=None, positive=False*) [source]

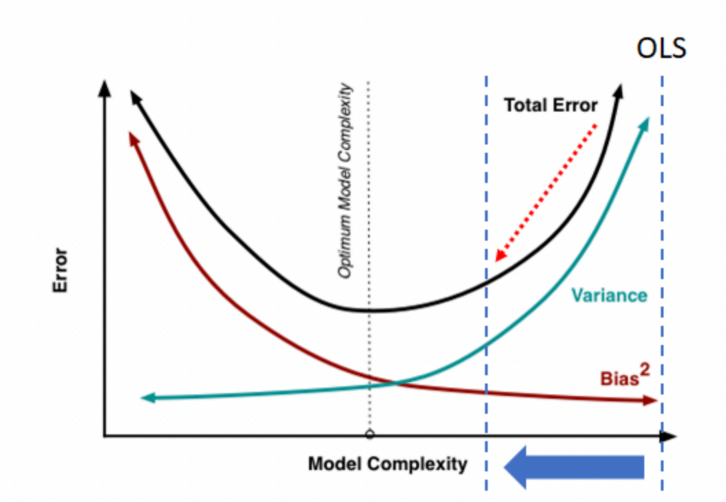
- **Parameters to look at**
  - fit_intercept
  - normalize

Ridge Regression

- **A variation of linear regression**
- **Addresses some problems in linear regression**
  - OLS doesn't consider which independent variable is more important than others
    - OLS is an unbiased estimator/model
  - We need to consider bias for better performance
    - Bias means how equally a model cares about its predictors/features

Ridge Regression

- **The OLS estimation usually gives low bias, high variance**
  - Why?
    - OLS treats all the variables equally; becomes more complex as new variables are added
  - bias is related with a model failing to fit the training set
  - variance is related with a model failing to fit the testing set



Ridge Regression

- **A constraint is added to linear regression**
  - Aka Ridge constraint

$$y_i = b_0 + b_1 x_1^i + b_2 x_2^i + \ \dots + b_n x_n^i + +e$$

$$e \sim N(0, \sigma^2)$$

subject to

$$b_0^2 + b_1^2 + b_2^2 + \ \dots + b_n^2 \ \leq C^2$$

  - Optimize

$$min_b \, \|Xb \ - y\|_2^2 + \alpha \, \|b\|_2^2$$

Ridge Regression

sklearn.linear_model.Ridge ¶

class `sklearn.linear_model.`**`Ridge`**(*alpha=1.0*, ***, *fit_intercept=True*, *normalize=False*, *copy_X=True*, *max_iter=None*, *tol=0.001*, *solver='auto'*, *random_state=None*) [source]

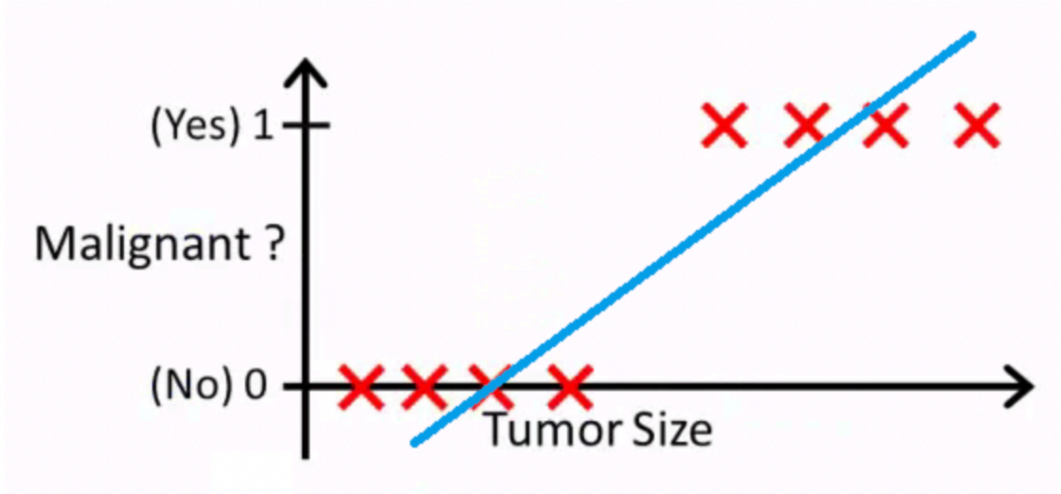
- **Parameters to look at**
  - alpha
  - auto
    - Optimization solver

Logistic Regression

- **A special type of linear regression**
  - Target variable is categorical
  - Features could be any types
- **Used for classification**
  - Then, why is it called regression?
    - It's not a classifier on its own
    - Logistic regression + a decision rule = a classifier
- **It a probabilistic classifier**
  - Discriminative model

$$\boxed{P(C \mid \mathbf{X}) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)}$$

Why not use linear regression?



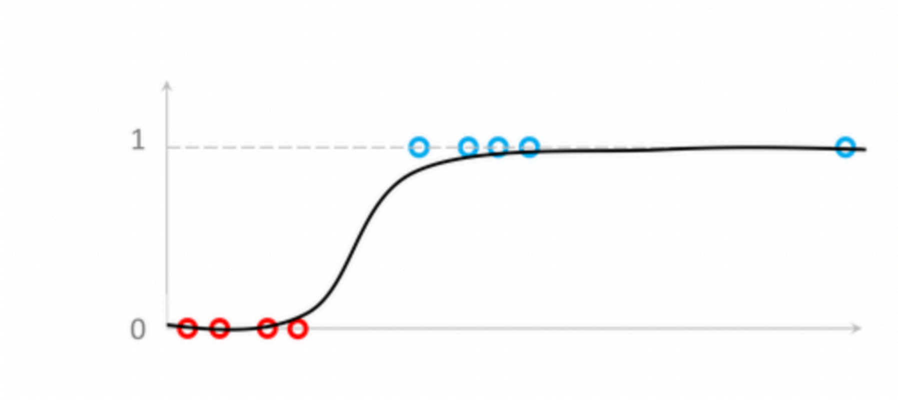
Enter Logistic Regression

- **Remember linear regression**

$$y_i = b_0 + b_1 x_1^i + b_2 x_2^i + \ldots + b_n x_n^i + +e$$

- **In logistic regression:**

$$P(C = c|X) = y_i = \frac{1}{1 + e^{-(b_0 + b_1 x_1^i + b_2 x_2^i + \ldots + b_n x_n^i)}_0}$$



Logistic Regression

- **A special type of regression**
  - Target variable is categorical
  - Features could be any types
- **Used for classification**
  - Then, why is it called regression?
    - It’s not a classifier on its own
    - Logistic regression + a decision rule = a classifier

Conclusion

- **Three linear models**
  - Linear regression
  - Ridge regression
    - Linear regression with a constraint
  - Logistic regression
    - Used in classification