Machine Learning Algorithms

0508 – Linear Regression

ML Algorithms
- Supervised ML
  - → Regression
  - → Classification
- Unsupervised ML → Clustering

Regression	**Classification**

Regression

House Price Prediction

# of rooms	Total size	Location	Price (O/P)
Independent			Dependent

Classification

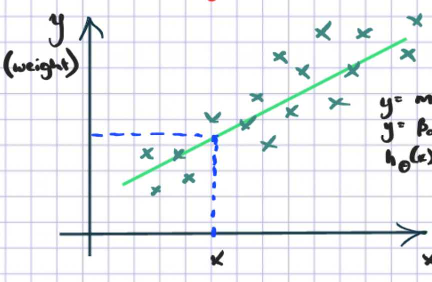
Binary or Multiclass Classification

# of play hours	# of sleep hours	hours studied	Pass / Fail (O/P)
5 hours	7 hours	1 hour	Fail

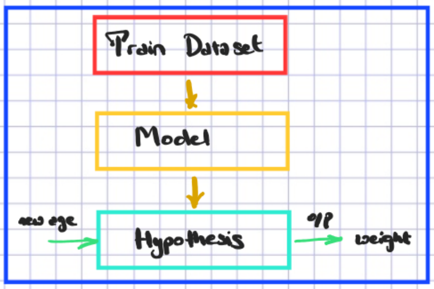
① Linear Regression
② Ridge Regression
③ Lasso Regression
④ Elastic Net
⑤ Logistic Regression [classification]

Decision Tree
Random Forest
Ada Boost
xgboost → Classification, → Regression

① Linear Regression



$y = mx + c$
$y = \beta_0 + \beta_1 \cdot x$
$h_\theta(x) = \theta_0 + \theta_1 \cdot x_1$



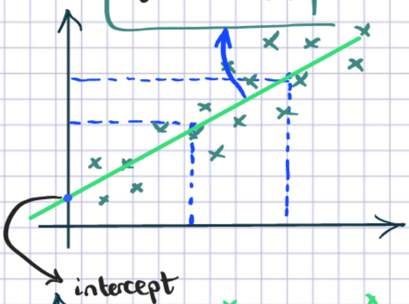
y is a linear function of x

Equation of a straight line

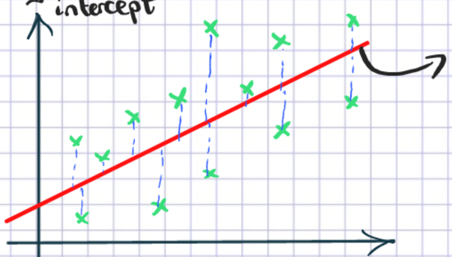
$y = mx + c$

$y = \beta_0 + \beta_1 \cdot x^{(i)}$

$\boxed{h_\theta(x) = \theta_0 + \theta_1 x_1}$



when $x = 0$
θ_0 = intercept
θ_1 = slope or coefficient
x_i = data points



best fit line

The main task of linear regression is to minimize the distance of (point - line)

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

$$(x^n)' = n \cdot x^{n-1}$$
$$\frac{\partial x^2}{\partial x} = 2x$$

Cost function

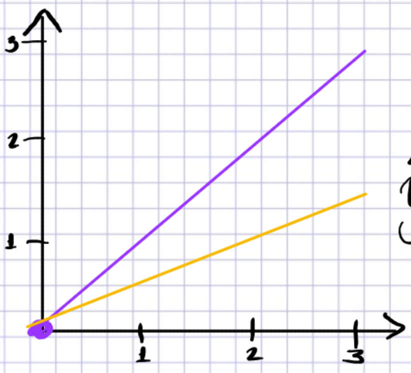
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$ ← Squared Error Function

What we need to solve?

$$\text{minimize } \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$
$$\Downarrow$$
$$\text{minimize } J(\theta_0, \theta_1)$$

$h_\theta(x) = \theta_0 + \theta_1 \cdot x_1$, if $\theta_0 = 0$

$\Rightarrow h_\theta(x) = \theta_1 x_1$



if $\theta_1 = 1 \Rightarrow J(\theta) = \frac{1}{2m} \sum_{i=1}^{3} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

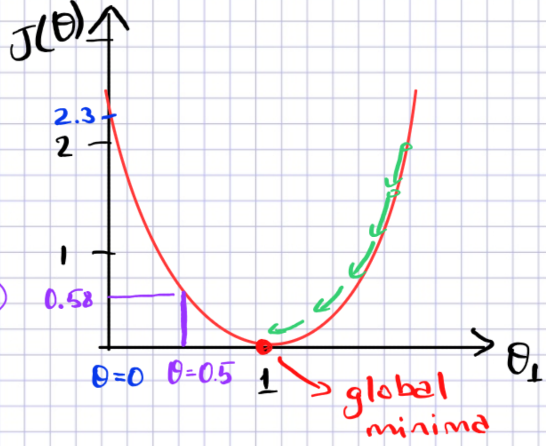
$= \frac{1}{2 \times 3} \left[(1-1)^2 + (2-2)^2 + (3-3)^2 \right] = 0$

if $\theta_1 = 0.5 \Rightarrow J(\theta) = \frac{1}{2m} \sum_{i=1}^{3} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

$= \frac{1}{2 \times 3} \left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2 \right] \approx 0.58$

if $\theta_1 = 0 \Rightarrow J(\theta) = \frac{1}{2m} \sum_{i=1}^{3} \left(h_\theta(x^{i}) - y^{i} \right)^2$

$= \frac{1}{2 \times 3} \left[(0-1)^2 + (0-2)^2 + (0-3)^2 \right] \approx 2.3$



Gradient Descent Convergence Algorithm

repeat until convergence {

$$\theta_j := \theta_j - \underset{\text{LR}}{\alpha} \boxed{\frac{\partial J(\theta_j)}{\partial \theta_j}}$$ → calculate slope

}

$\theta_{new} = \theta_{old} - \eta(+ve \text{ value})$

$= \theta_{old} - (+ve \text{ value})$

$\Rightarrow \theta_{new} << \theta_{old}$

$\theta_{new} = \theta_{old} - \eta(-ve \text{ value})$

$= \theta_{old} + ve \text{ value}$

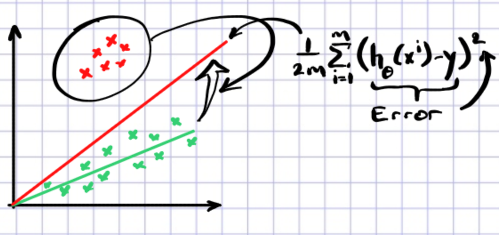
$\Rightarrow \theta_{new} >> \theta_{old}$

Multi Linear Regression: $h_\theta(x) = \theta_0 + \theta_1 \cdot x_1 + \dots + \theta_n x_n$

Advantages: ① There is one global minima

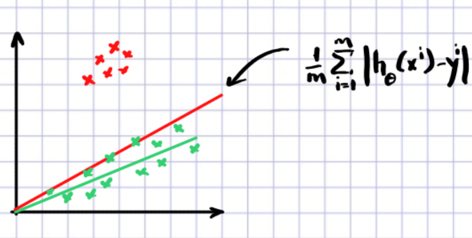
② It is differentiable

Disadvantage: ① Not robust to outliers [MSE penalizes Error]

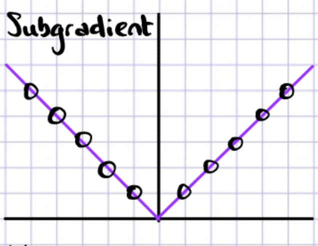


Mean Absolute Error

$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} |\hat{y} - y|$, where $\hat{y} = h_\theta(x)$



Subgradient



Using 5 number summary, remove outliers and use MSE.

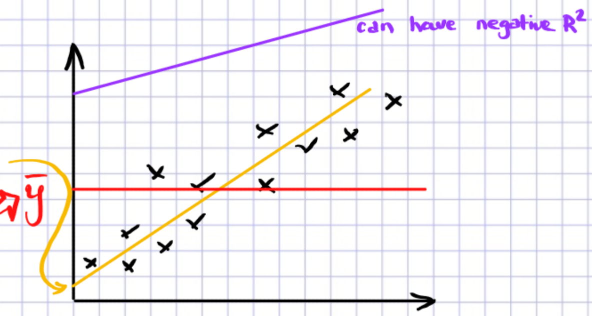
Performance Metrics

R^2 and Adjusted R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{m} (y^{(i)} - \bar{y}^{(i)})^2}$$

SS_{res} = Squared sum error of regression line (sum of squares of residuals)

SS_{total} = Squared sum error of mean line (total sum of squares)



Adjusted R^2

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$, where N = # of data points

p = # of predictors

House Pricing

	# of rooms	Total size	Location	Gender	PRICE
R^2		= 0.89	↑ = 0.95	↑ = 0.96	— " —
Adj. R^2		= 0.87	↑ = 0.89	↓ = 0.85	— " —

RMSE

Root mean squared error = $\sqrt{\frac{\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2}{N}}$

Linear Regression:

* There is a linear relationship between x any y
* It is always good if feature is normally distributed
* Always take care of multicollinearity
* Homoscedacity: all the features have the same variance (use one of them)
* Feature Scaling is required, since algorithm use Gradient Descent
* Heteroscedeacity: condition in which the variance of the residual term, or error term, in a regression model varies widely