

Fundamentals of Machine Learning

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST

icmoon@kaist.ac.kr

Weekly Objectives

- Learn the most classical methods of machine learning
 - Rule based approach
 - Classical statistics approach
 - Information theory approach
- Rule based machine learning
 - How to find the specialized and the generalized rules
 - Why the rules are easily broken
- Decision Tree
 - How to create a decision tree given a training dataset
 - Why the tree becomes a weak learner with a new dataset
- Linear Regression
 - How to infer a parameter set from a training dataset
 - Why the feature engineering has its limit

2/10
10/10

LINEAR REGRESSION

How about statistical approach?

(rule-based vs. stat.)

- <http://archive.ics.uci.edu/ml/datasets/Housing>
- Housing dataset
 - 13 numerical independent values
 - 1 numerical dependent value : True / False label
- How to create an approximated function? : (linear vs. non-linear approx. = LR)
 - Do you remember that the machine learning is the function approximation process?
- Here,
 - Our hypothesis is (가정하는 모델, 피팅을 잘. - 최적의 함수를 찾는 것)
 - The house value will be the linearly weighted sum of the feature values.
 - $\hat{f}(x; \theta) = \theta_0 + \sum_{i=1}^n \theta_i x_i = \sum_{i=0}^n \theta_i x_i$
 - n is the number of the feature values.
 - Two aspects: the linearly weight sum (the model), the parameter θ
 - The first effort is finding the better θ , just like the thumbtack

Finding θ in Linear Regression

- To make the hypothesis better, we need to find the better θ

- $$h: \hat{f}(x; \theta) = \sum_{i=0}^n \theta_i x_i \rightarrow \hat{f} = X\theta$$

- $$X = \begin{pmatrix} 1 & \cdots & x_n^1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_n^D \end{pmatrix}, \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

(true f \neq \hat{f} (prediction))
 (error)

- The reality would be the noisy, so....

- $$f(x; \theta) = \sum_{i=0}^n \theta_i x_i + e = y \rightarrow f = X\theta + e = Y$$

- The difference is the error from the noise, so let's make it minimum

- $$\begin{aligned} \hat{\theta} &= \operatorname{argmin}_{\theta} (f - \hat{f})^2 = \operatorname{argmin}_{\theta} (Y - X\theta)^2 \\ &= \operatorname{argmin}_{\theta} (Y - X\theta)^T (Y - X\theta) = \operatorname{argmin}_{\theta} (Y - X\theta)^T (Y - X\theta) \\ &= \operatorname{argmin}_{\theta} (\theta^T X^T X \theta - 2\theta^T X^T Y + \cancel{Y^T Y}) = \operatorname{argmin}_{\theta} (\theta^T X^T X \theta - 2\theta^T X^T Y) \end{aligned}$$

- Now, we need to optimize θ

($\because \perp \theta$)

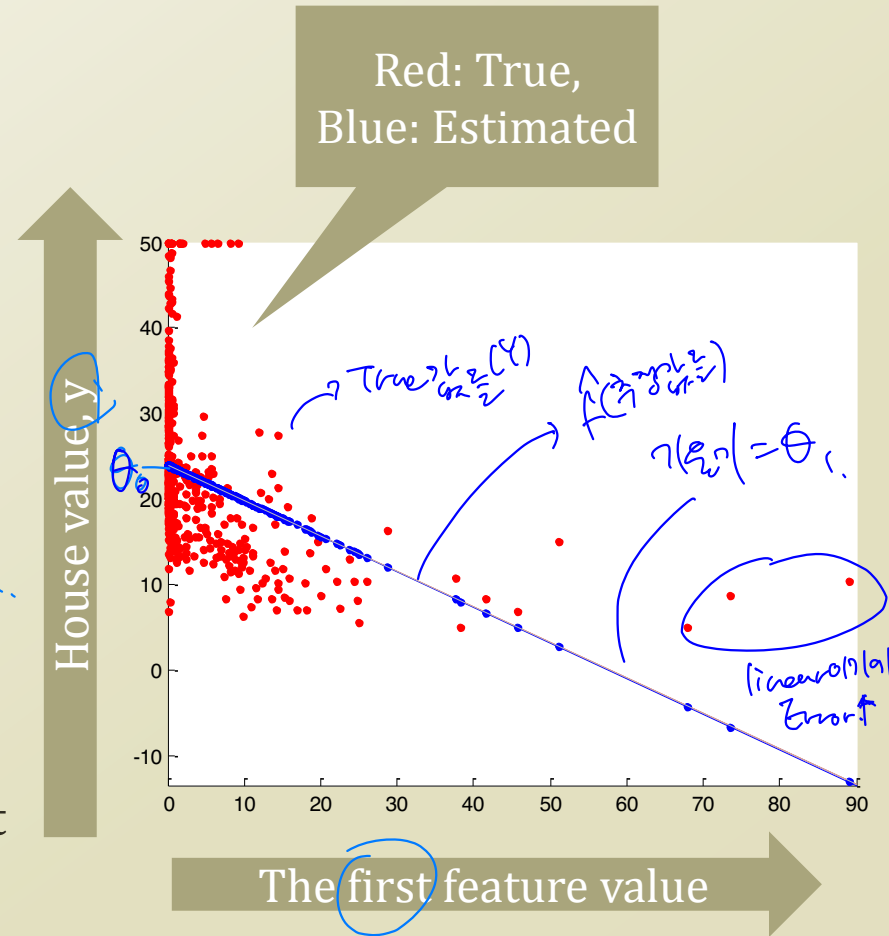
Optimized θ

Handwritten notes at the top right of the slide:

$$1 \quad 1 \quad \frac{2}{3} \quad 1 \quad \frac{2}{3} \quad \frac{1}{3}$$

Below these, there are some scribbles and the letter 'T'.

- $\hat{\theta}$
 $= \operatorname{argmin}_{\theta} (\theta^T X^T X \theta - 2\theta^T X^T Y)$
- Same technique as in Thumbtack
 - $\nabla_{\theta} (\theta^T X^T X \theta - 2\theta^T X^T Y) = 0$
 - $2X^T X \theta - 2X^T Y = 0$
 - $\theta = (X^T X)^{-1} X^T Y$
- Great! We know X and Y, so we can compute θ — given Data.
- Let's calculate and watch the performance!
 - For demonstration purpose, we limit the dimension to the constant and the first feature
- MLE if error follows the normal distribution

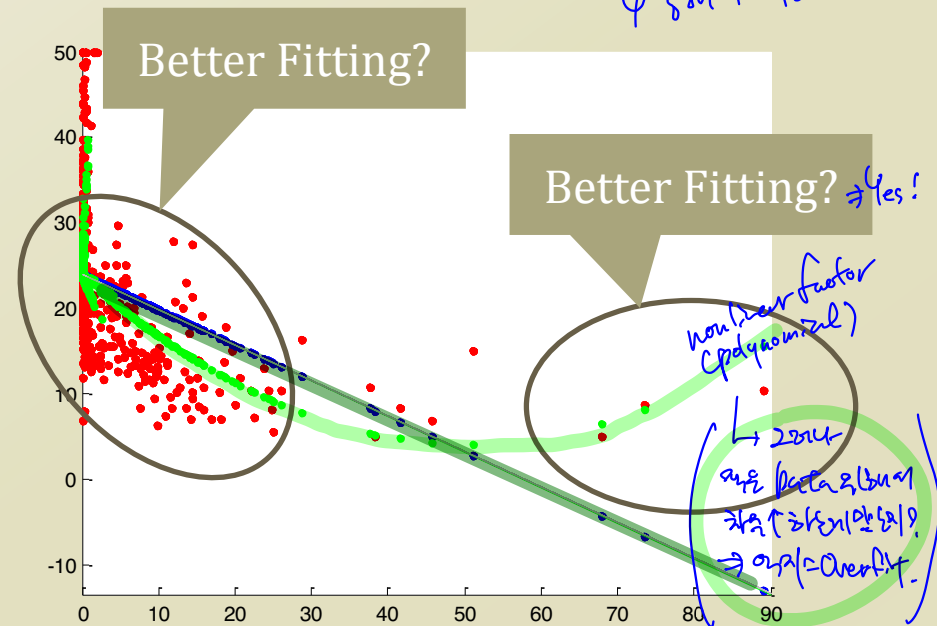
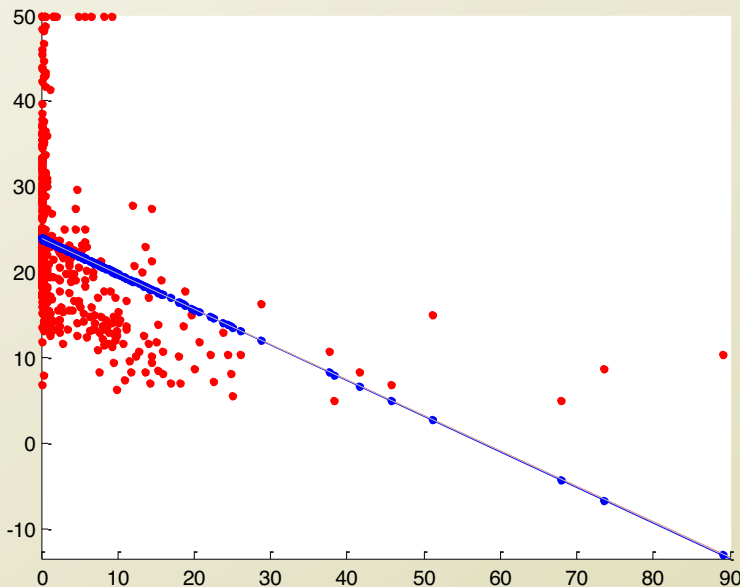


If you want more....

- Actually, you can increase the number of features, a.k.a. dimension
 - You have a value x in the previous fitted figure
 - How about adding x^2, x^3, x^4, \dots ? (polynomial)
- You can improve the result!
 - Is that right?
- We are going to come back!

$$h: \hat{f}(x; \theta) = \sum_{i=0}^n \sum_{j=1}^m \theta_{i,j} \phi_j(x_i)$$

ϕ_j is a feature vector.



Too Brittle to Be Used Naively

- What we are doing
 - Approximating a function to the dataset
 - The function can be
 - Discrete logics
 - Statistical model
 - Often, the function type is given
 - The parameters of the functions are the target of the analysis
- Alternatives in finding the parameter
 - MLE or MAP
 - Engineering the features
 - Setting the generalization and the specialization level
- Best choice among the alternatives
 - If we have the perfect data, we will know
 - But we don't

