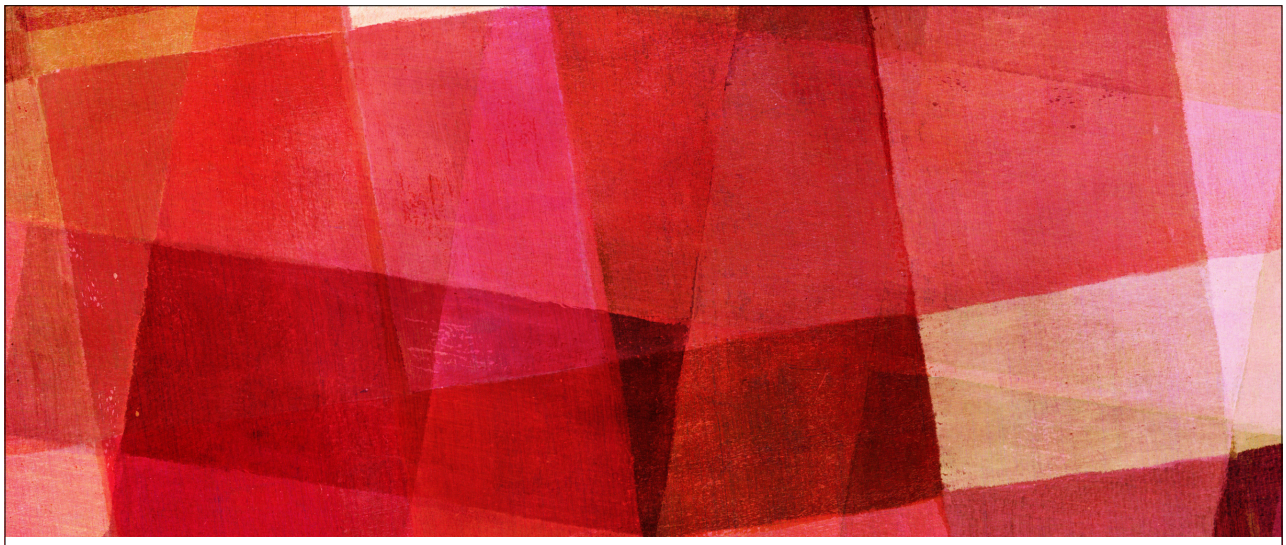


DATA PREPROCESSING

Warasinee Chaisangmongkon, PhD

FEATURE ENGINEERING

Data rarely comes in the appropriate form so one of the most important tasks in machine learning is to transform those data.

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

FEATURE ENGINEERING

- Preprocessing (filtering, cleaning, linking)
- Feature extraction
  - Create a new feature from existing feature
  - Create a new feature from 2 or more features
- Feature selection
- Dimensionality reduction
- Deep learning (a.k.a feature learning)

GETTING USEFUL FEATURE OUT OF NOT-SO-USEFUL FEATURE

Predict house price from photos

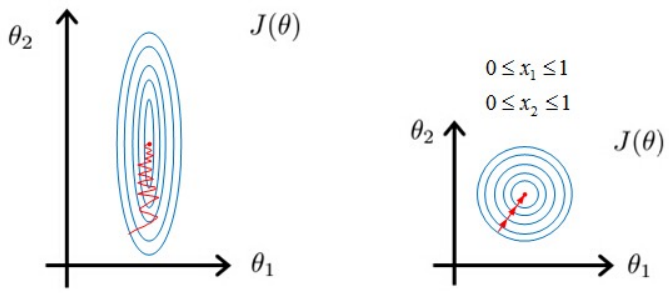
PREPROCESSING AND FEATURE ENGINEERING BASICS

(some common operations)

1. FEATURE SCALING

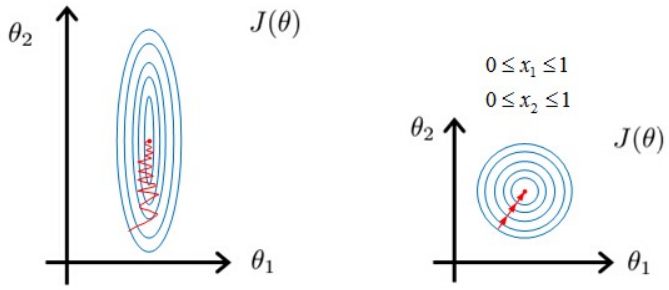
- Imagine if x1 and x2 are not similar in scale
- For example
  - x1 = number of bedrooms (0-20)
  - x2 = area of the house (24-3000 sqm)
- This means the scale of your theta will also be different.

1. FEATURE SCALING



- A symmetric contour
- Landscape of cost function is like **ขนมเบื้อง**
- Gradient descent algorithm has a hard time with such surface

1. FEATURE SCALING



- Simple solution would be to scale the features.
  - x1: number of bedroom -> x1 = #bedroom / max #bedroom
  - x2: area of the house -> x2 = area / max area
- Try to get features to stay within -1 to 1

MEAN NORMALIZATION AND Z-SCORE

- These two techniques are suitable for most problems

mean normalization

$$x_1 := \frac{x_1 - mean(x_1)}{max(x_1)}$$

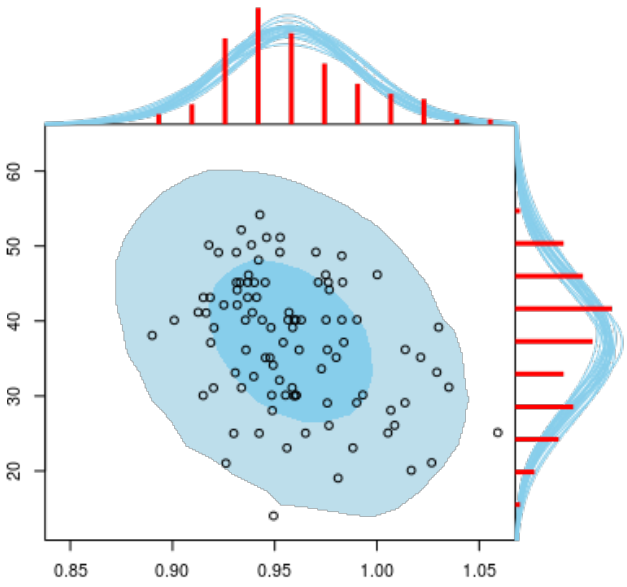
$$x_1 := \frac{x_1 - mean(x_1)}{max(x_1) - min(x_1)}$$

z-score

$$x_1 := \frac{x_1 - mean(x_1)}{std(x_1)}$$

What's the range of these rescaled variables?

2. VARIABLE WITH SKEWED DISTRIBUTION

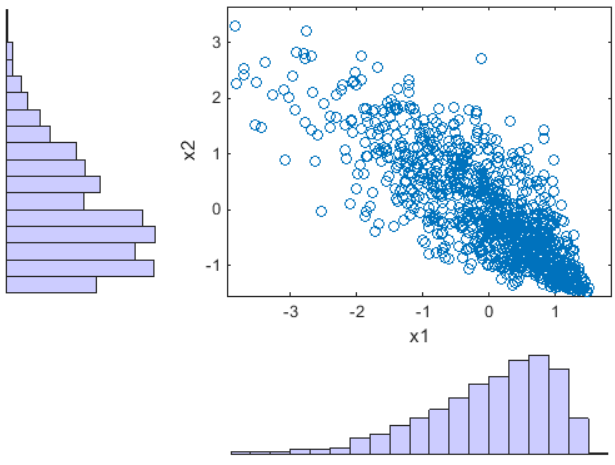


Normally, we operate regression with the assumption that variables are normally distributed.

2. VARIABLE WITH SKEWED DISTRIBUTION

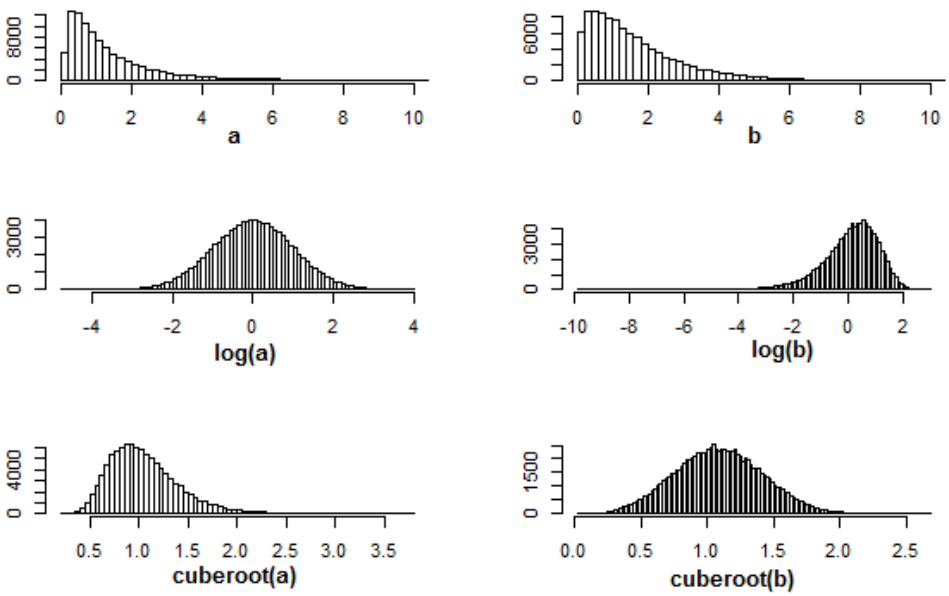
- Regression has the following assumptions:
  - Linear relationship
  - Multivariate normality
  - No or little multicollinearity
  - No auto-correlation
  - Homoscedasticity (all variables have same variance)

SKEWED DISTRIBUTION



Regression does not guarantee solution for skewed distribution.

FIXING SKEWED DISTRIBUTION



3. DEALING WITH MISSING DATA

- If data are assumed to be missing at random, we may simply ignore the data
  - You went to all houses and randomly for some houses, you took time to do detailed house measurement
  - In this case, you simply remove the missing data from the analysis (cut the whole row)
  - Be aware that missing data might not be as random as you think

3. DEALING WITH MISSING DATA

i	Size (m^2)	Price	# Bed	Price
1	50	1.4	2	1.4
2	128	2.6	?	2.6
3	24	0.8	1	0.8
4	?	1.2	2	1.2
i	…	…		

- Listwise means cutting the whole row (reduced sample)
- Pair wise means cutting only the missing value (can't do multivariate analysis)

3. DEALING WITH MISSING DATA

- Mean or mode substitution
  - Missing income? Fill it with average income?
  - Don't know if the patient is left-handed or right-handed, assume right-handed, because it's more common.
  - Weaken correlation and covariance.

3. DEALING WITH MISSING DATA

- Dummy variable
  - Customers are divided to high (3), medium (2), low income (1), assume that customers with missing income is of category 0.
  - In some cases, this is perfect because people that avoid filling in income might have interesting characteristics.
  - In some cases, this method is not so good because you introduced and extra value that is not driven by fact.
  - For example, if customers forget to fill in their ages, you would not assume all the missing data have something in common.

3. DEALING WITH MISSING DATA

- Regression imputation.

i	Size (m^2)	Price	# Bed	Price
1	50	1.4	2	1.4
2	128	2.6	?	2.6
3	24	0.8	1	0.8
4	?	1.2	2	1.2
i	…	…		

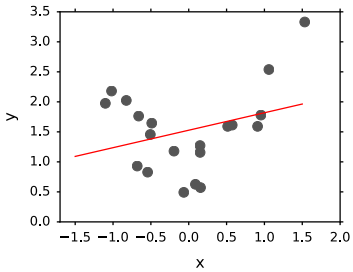
- Use regression model to estimate the relationship between variables to fill in the missing value.
- Overestimate model fit (features are too smooth).

OVERFITTING

OVERFITTING

underfit

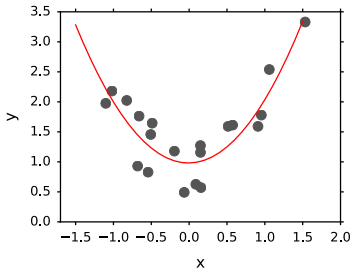
cost = 0.21



$h(x) = \theta_0 + \theta_1 x$

just right

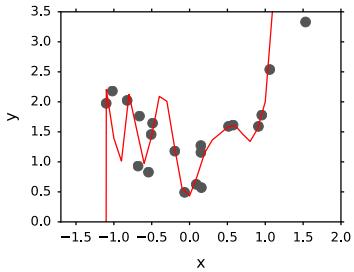
cost = 0.05



$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

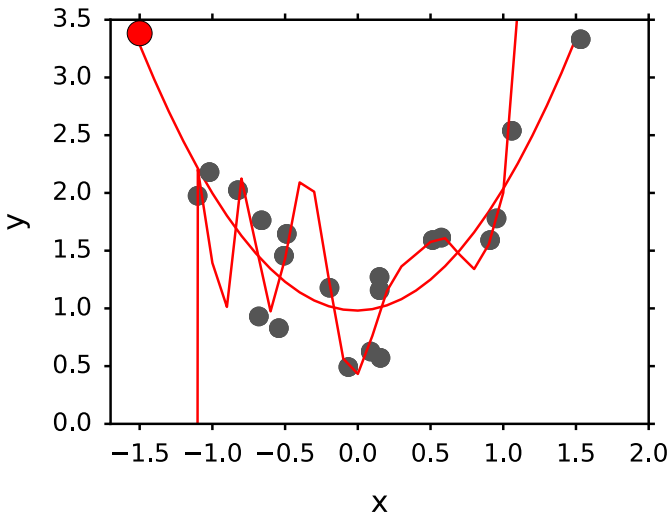
overfit

cost = 0.02



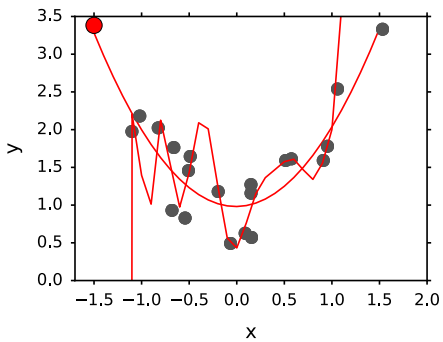
$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + ... + \theta_{15} x^{15}$

OVERFITTED MODELS COULD NOT GENERALIZED



FACTS ABOUT OVERFITTING

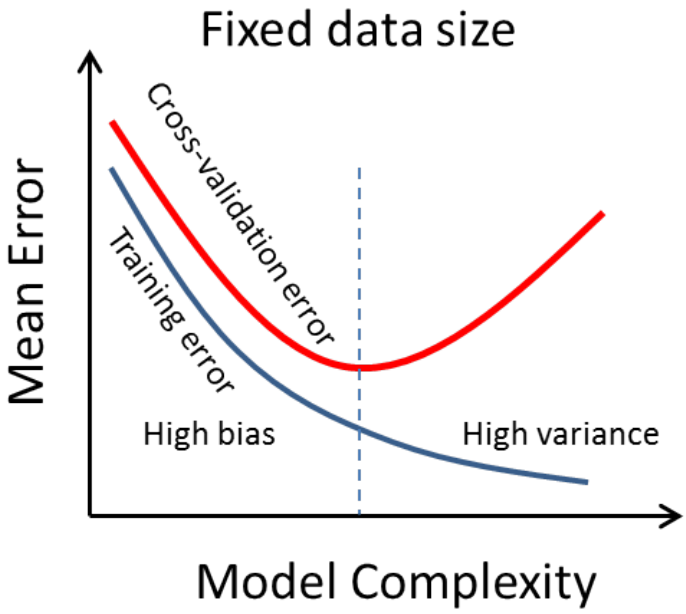
- If the number of features (n) is really large, you can fit y with really high precision.
- The more data you have (compared to parameters) the less likely your model will overfit, because noise will be more likely to average out.



FACTS ABOUT OVERFITTING

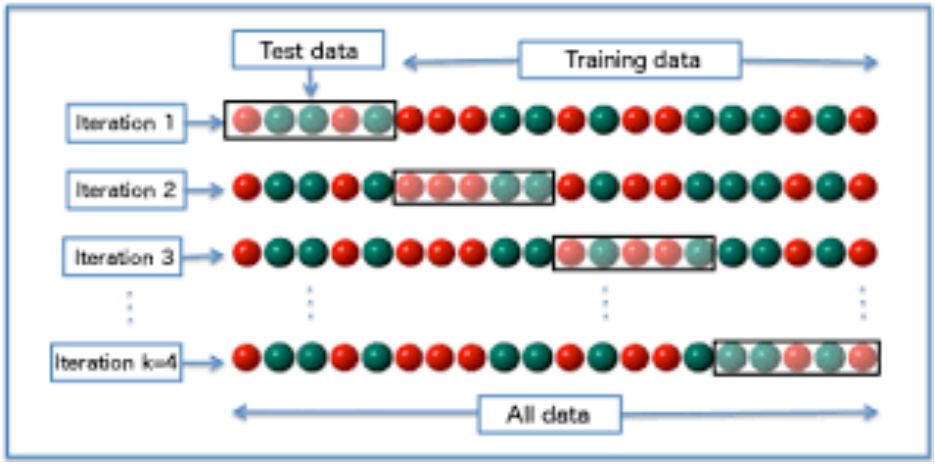
- The more parameters, the larger solution space you fit to the data, and this is not a good thing.
- The less parameters, the more likely your model will underfit (not having enough feature to make predictions about the target).
- This is call bias-variance tradeoff (bias: model has bias, i.e. too strong assumption) (variance: too much errors/noise is fitted).

FACTS ABOUT OVERFITTING



AVOIDING OVERFITTING: CROSS VALIDATION

- Cross-validate your models: train models with one set of data (training set) and test them with another set of data (test set)



AVOIDING OVERFITTING: CROSS VALIDATION

- Reduce the number of features and parameters
  - Manual selection
  - Algorithmic selection (information gain, pruning, stepwise algorithm)
  - Regularization

REGULARIZATION

- Regularization is a mathematical way to reduce overfitting automatically, by limiting the influence of each feature.

$$\sum_i (h(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$$

The normal cost function *norm of parameter vector*

- Usually lambda is really huge like 1,000
- This means if we increase theta a little bit, cost function will go up a lot

REGULARIZATION

- Regularization is a mathematical way to reduce overfitting automatically, by limiting the influence of each feature.

$$\sum_i (h(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$$

The normal cost function *norm of parameter vector*

- To minimize this cost function you have to minimize both first and second terms
- Minimize the second term means less overfitting.

REGULARIZATION

$$\sum_i (h(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$$

The normal cost function *norm of parameter vector*

- Lambda is called 'regularization parameter'
- Because we adjust lambda to adjust the weight of the two cost function terms
- High lambda: severely limit parameter size
- Low lambda: allow parameters to scale up more freely, give more importance to lowering error

REGULARIZATION

$$\sum_i (h(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$$

The normal cost function *norm of parameter vector*

- Lambda is called 'regularization parameter'
- Because we adjust lambda to adjust the weight of the two cost function terms
- High lambda: severely limit parameter size
- Low lambda: allow parameters to scale up more freely, give more importance to lowering error

REGULARIZATION

- If lambda is too high, you will be underfitting.
- There are two common types of regularization:
  - L2-regularization is the equation above
  - L1-regularization use $|\theta|$ instead of $|\theta|^2$

PREPROCESSING LAB