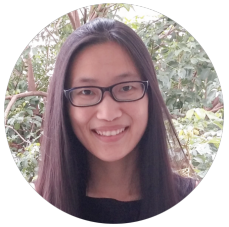


Warasinee Chaisangmongkon, PhD

Lecturer at Institute of Field Robotics, KMUTT. Data Scientist at Big Data Experience Center

- PhD from Yale University
- 10 years experiences in machine learning & data mining
- Lead scientist of corporate data science projects (banking, business digitizations)
- Found IDEA LAB - R&D in machine learning for businesses

- What is Machine Learning
- Feature, Target, Sample
- Model & Cost Function
- Overfitting

INTRODUCTION TO MACHINE LEARNING

Introduction to Machine Learning

Instructor: Warasinee Chaisangmongkon, PhD

Day 2 : Predictive Modeling

Introduction to Machine Learning
What it means to make predictive models for your data.

Feature Engineering
How to get your data ready for modeling and how to select the best features.

Linear Regression and Logistic Regression
Interactive lessons in linear and logistic regression in R

K-Nearest Neighbor
Interactive lessons on KNN methods.

Decision Trees and Random Forest Models
Interactive lessons in tree-based modeling.

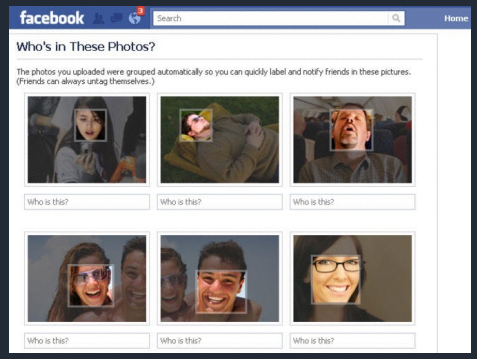
Lab
Use your knowledge to fit a model to predict house prices!

Conclusions
Other regression and classification models.

THE MOST BASIC UNDERSTANDING

- It's all about letting computer learns what 'input' is associated to what 'output'.
  - Example: given behavioral profiles of a customer A, algorithms predict the chance this customer is going to leave.
  - Example: given inputs from sensors and cameras, the robotic algorithm pushes out the appropriate movement.

INTELLIGENT SYSTEM WITH MACHINE LEARNING

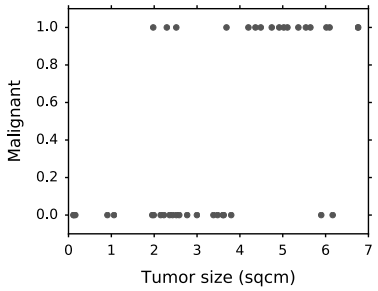


- People provide Facebook the images and tags of names in the photos.
- Over time Facebook learned to associate names with faces and can automatically recognize these people.

MACHINE LEARNING

THE SCIENCE OF GETTING COMPUTERS TO LEARN
FROM DATA WITHOUT HAVING
TO BE EXPLICITLY PROGRAMMED BY HUMANS.

A SIMPLE EXAMPLE



- A doctor has seen hundreds of patients with tumors. They measured the size of tumors and test whether they are malignant.
- The historical data let the doctor make prediction that bigger tumor is more likely to be malignant.

THE MOST BASIC EXAMPLE

In general,

x: feature (input)

y: target (output)

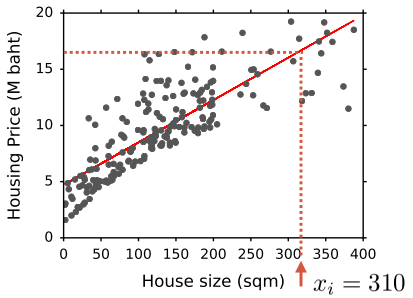
i: sample index

	x	y
i	Size (m^2)	Price (Mbaht)
1	50 $= x_1$	1.4 $= y_1$
2	128 $= x_2$	2.6 $= y_2$
3	24 $= x_3$	0.8 $= y_3$
4	78 $= x_4$	1.2 $= y_4$
i	... $= x_i$	... $= y_i$

WHAT IS A MODEL

After we have a model, we can predict y value from any x value

Model: $h(x_i) = 4.717 + 0.038 * (x_i)$



Notice that:

h(x) and y are not the same

h(x): value we predict

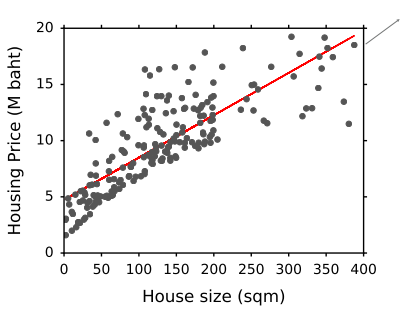
y: the actual value

THE MOST BASIC EXAMPLE

An agent has been selling 300 houses in the last years and want to be able to predict the price of a house by just knowing the size of the house.

i	Size (m^2)	Price (Mbaht)
1	50	1.4
2	128	2.6
3	24	0.8
4	78	1.2
i	...	...

WHAT IS A MODEL



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

parameters of the model

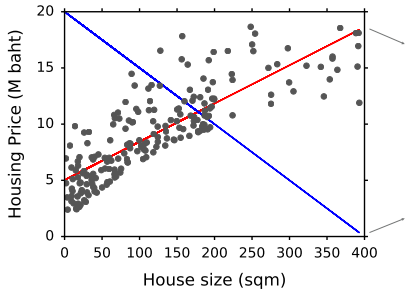
$\theta_0 = 4.717$

$\theta_1 = 0.038$

A model is a function that takes an input and yields the values we want to predict.

GOOD AND BAD MODELS

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



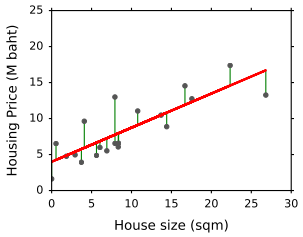
$\theta_0 = 4.717$
$\theta_1 = 0.038$

$\theta_0 = 20$
$\theta_1 = -0.05$

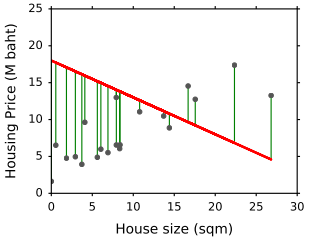
LOWERING COST FUNCTION

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

$\theta_0 = 4.717$
$\theta_1 = 0.038$

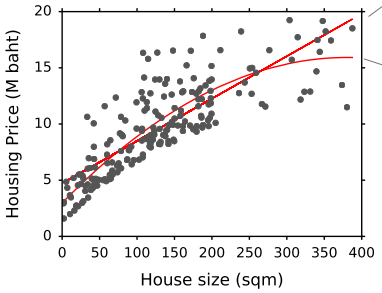


$\theta_0 = 20$
$\theta_1 = -0.05$



WHAT IS A MODEL

For different models you have different functions, with perhaps different number of parameters. Different models can be fitted to the same data set.



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

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

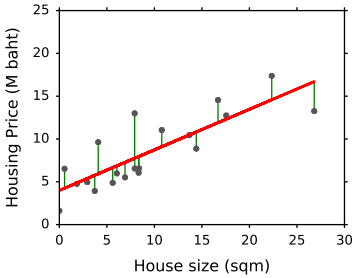
Model is a function of both features (x) and parameter (theta).

$h(x; \theta)$

COST FUNCTION

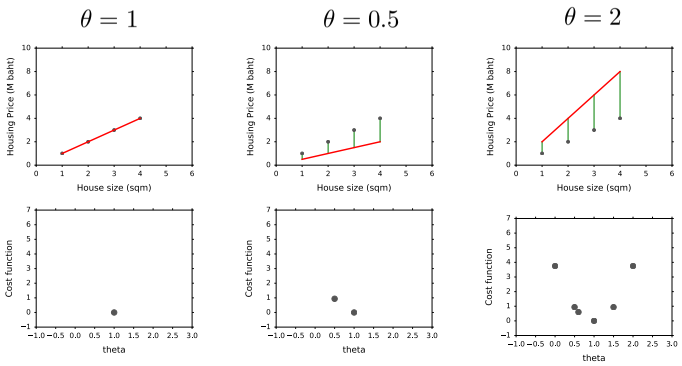
Cost function (or loss function) is a measure of whether a model is a good fit to the data.

For example, a famous cost function is called 'squared error' function that takes the difference between what you predict and the actual data value and square it.

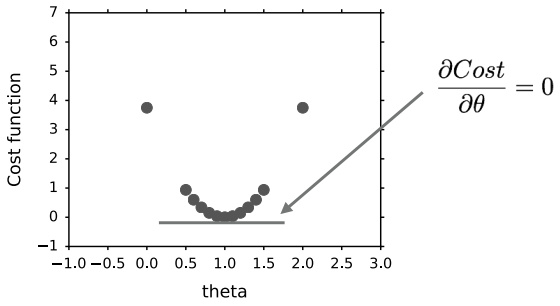


$$\sum_i (h(x_i) - y_i)^2$$

MINIMIZING COST FUNCTION

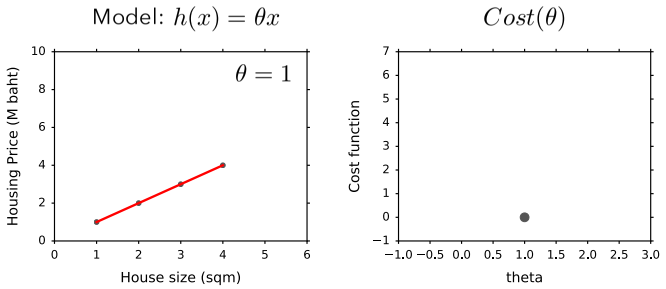


MINIMIZING COST FUNCTION

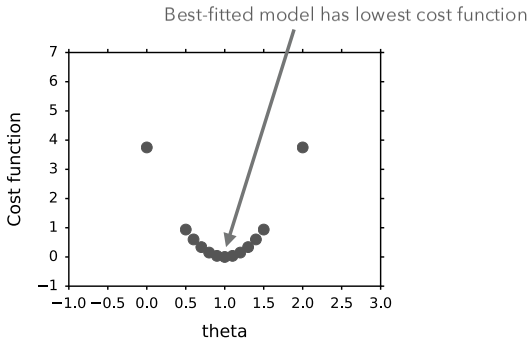


Some optimization algorithm is so simple,
you just solve an equation and done.

MINIMIZING COST FUNCTION



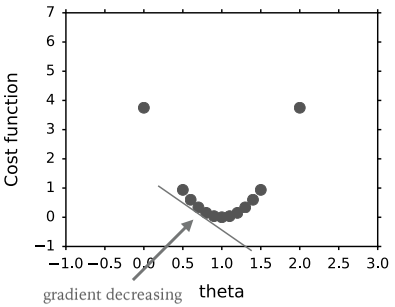
MINIMIZING COST FUNCTION



Training machine learning models = minimizing cost function

Often accomplished by using 'optimization algorithms'

MINIMIZING COST FUNCTION



Gradient Descent

- pick a random theta
- find gradient
- update theta following the descending gradient

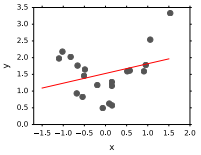
$\theta \rightarrow \theta - \alpha * gradient$

- gradient will decrease
- repeat until gradient is zero

OVERFITTING

underfit

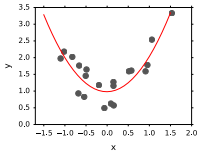
cost = 0.21



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

just right

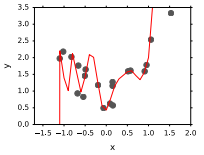
cost = 0.05



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

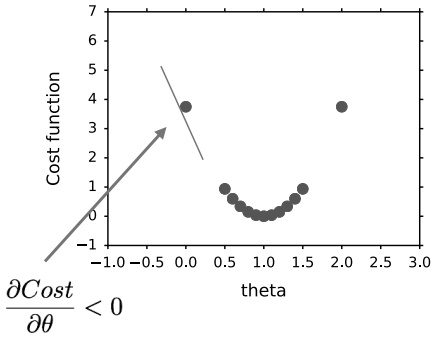
overfit

cost = 0.02



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

MINIMIZING COST FUNCTION



$\frac{\partial Cost}{\partial \theta} < 0$

Gradient Descent

- pick a random theta
- find gradient
- update theta following the descending gradient

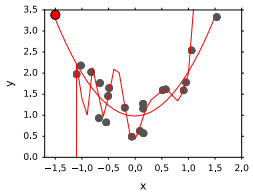
$\theta \rightarrow \theta - \alpha * gradient$

TAKE-HOME MESSAGES

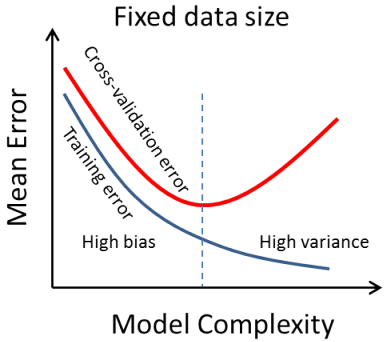
- Machine learning is a science of letting computers take input (features), pass them through a model (an equation or a system of equations) and return a predicted output.
- Models consist of parameters we can adjust to fit the data.
- The goodness of fit is judged by a cost function which measures how much our predictions are far away from actual data (target).
- Machine learning application always involves an algorithm which optimizes cost function.

FACTS ABOUT OVERFITTING

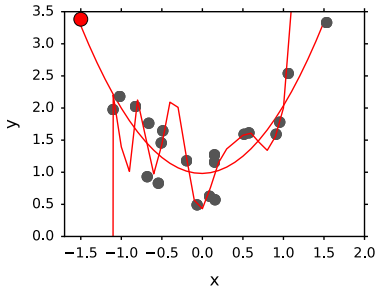
- If the number of features (n) is really large, you can fit y with really high precision.
- The more data you have (compared to parameters) the less likely your model will overfit, because noise will be more likely to average out.



FACTS ABOUT OVERFITTING



OVERFITTED MODELS CANNOT GENERALIZE

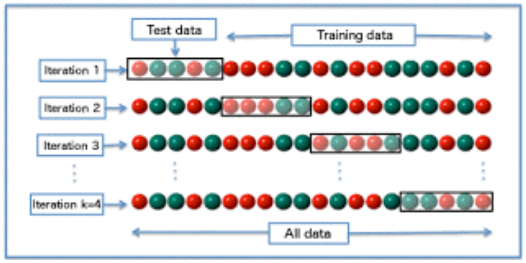


FACTS ABOUT OVERFITTING

- 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).

AVOID OVERFITTING

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



REGULARIZATION

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

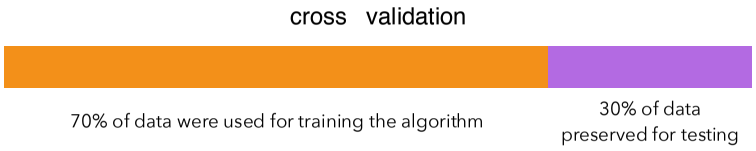
$$\sum_i (h(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$$

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.

AVOID OVERFITTING

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



AVOID OVERFITTING

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

AVOID OVERFITTING

- Cross-validate your models: train models with one set of data (training set) and test them with another set of data (test set)
- Reduce the number of features and parameters
  - Manual selection
  - Algorithmic selection (information gain, pruning, stepwise algorithm)
  - Regularization: cost function is split to two parts

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

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

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.

- What is it and why
- Data preprocessing
- Preprocessing w/t R

FEATURE ENGINEERING

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)

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.

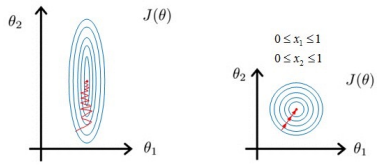
DATA SCIENTIST TASKS

- 50% - Obtaining data and cleaning data
- 30% - Feature engineering
- 20% - Modeling

COMMON PREPROCESSING OPERATIONS

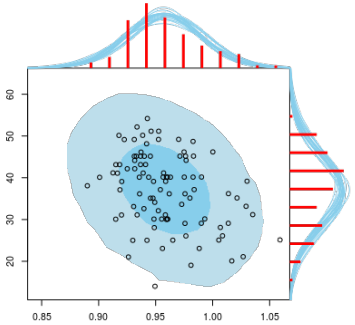
- Feature scaling
- Dealing with skewed data
- Dealing with missing data

FEATURE SCALING



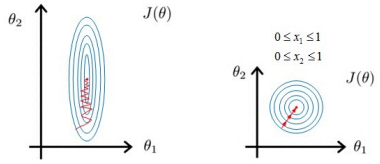
- Simple solution would be to scale the features.
  - x1: number of bedroom -> x1 = #bedroom / max #bedroom
  - x2: area of the house -> x2 = area / max area
- Try to get features to stay within -1 to 1

SKEWED DATA



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

FEATURE SCALING



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

FEATURE SCALING

- These three techniques are suitable for most problems

normalization with max, min, mean

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

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

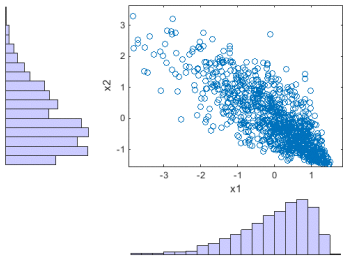
$$x_1 := \frac{x_1 - mean(x_1)}{max(x_1) - min(x_1)}$$

standardization a.k.a. z-score

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

What's the range of these rescaled variables?

SKEWED DATA



Regression does not guarantee solution for skewed distribution.

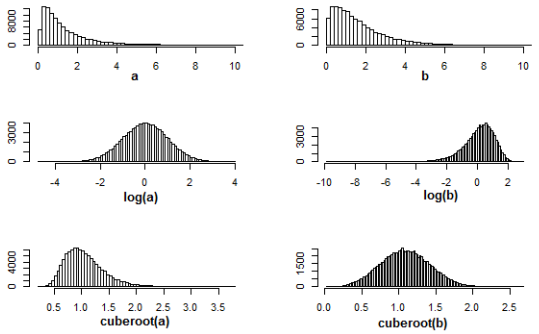
MISSING DATA

- 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

SKEWED DATA

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

SKEWED DATA



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.

MISSING DATA

i	Size (m^2)	Price (Mbaht)	# 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).

MISSING DATA

i	Size (m^2)	Price (Mbaht)	# 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)
- Pairwise means cutting only the missing value (can't do multivariate analysis)

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.

Big Data

- Most problems you will face in the real world is gigantic.
  - Millions of rows
  - Hundreds or thousands of features
  - Your algorithm will take forever to run
- What can we do about it?
  - We might be able to look through all the features and manually select them.
  - But that would waste so much time and resources
  - So maybe do automated feature selection?

What Feature Matters Most?

- The algorithm predicts whether the email is spam or not. Which feature is most useful for the prediction?
  - Feature 1: whether the email contains the word 'viagra'
  - Feature 2: whether the email is sent from a Nigerian Prince
  - Feature 3: whether the email is sent from one person to a massive amount of people
- For all 1000 features you have calculated, maybe only a few features are important.
- Feature selection algorithm gives you insight and interpretability of your model.

- Why Feature Selection
- How to do feature selection

53

FEATURE SELECTION

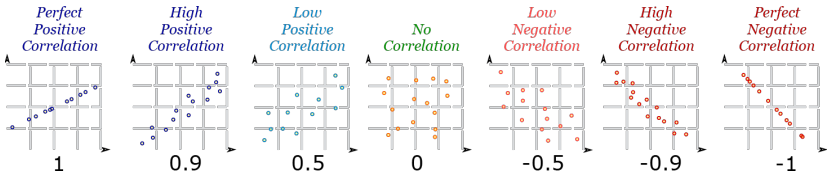
55

Feature Selection :
The process of selecting the most relevant features to be included in the machine learning model

Curse of Dimensionality

- This is one of the most important problems in machine learning.
  - Take linear regression for example
  - The more features you have, the more parameters you need to fit the model
  - If you have 2 features, your solution space has 2 dimensions (small possible values)
  - If you have 1000 features, your solution space has 1000 dimensions (huge amount of possible values.
  - You algorithm can take a lot of time to find solution.

How to find a good feature

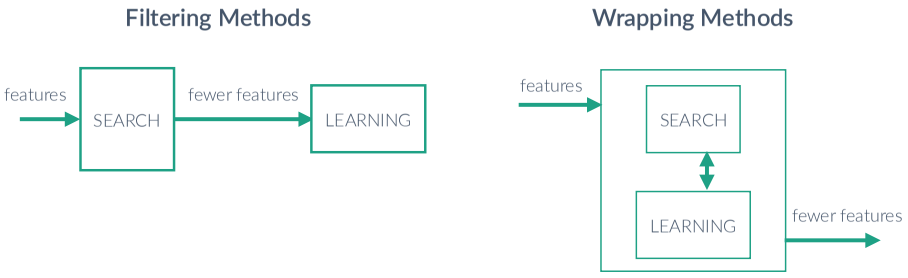


- You need to find features that have high correlation to your target.
  - Don't care whether it's a positive or negative correlation
  - The larger the number the better

Curse of Dimensionality

- This is one of the most important problems in machine learning.
  - Imagine you have a large amount of features, each can have infinite number of values.
  - You will need an enormous amount of training data is required to ensure that there are several samples with each combination of values.
  - If you have limited samples, which do not cover the whole space, your model loses predictive power.

Feature Selection



Analysis of Variance

- Classification problem: Y can only be class 0 or class 1. Find variance of X within class and between classes.

$$\text{F-Value} = \frac{\text{Variance between classes}}{\text{Variance within class}}$$

V between class is **high**
V within class is **low**
F-Value is **high**
Feature X is **important**

V between class is **low**
V within class is **high**
F-Value is **low**
Feature X is **not important**

How to find a good feature

- Correlation is not the only measure that tells you 'how much x is related to y' there are other measures we use.
- Such as:
  - ANOVA (Analysis of Variance)
  - Chi2
  - Mutual Information

Analysis of Variance Quiz

Weight	Class
50	Adult
80	Adult
12	Children
30	Children
...	...

V between class is ...
V within class is ...
F-Value is ...
Feature is ...

Weight	Class
68	Thailand
75	China
80	Thailand
82	China
...	...

V between class is ...
V within class is ...
F-Value is ...
Feature is ...

Weight	Class
65	Human
1000	Animal
0.1	Animal
60	Animal
30	Human

V between class is ...
V within class is ...
F-Value is ...
Feature is ...