NUS DataScience

Introduction to Data Analytics with R

TOH Wei Zhong

31/10/2015

A little bit about me

- Graduated from NUS, Computational Biology
  - Statistics and computing onto biology and healthcare
  - E.g. –omics
- Data Scientist in NCS
  - Smart Nation projects (defense and public safety)

Agenda for this afternoon

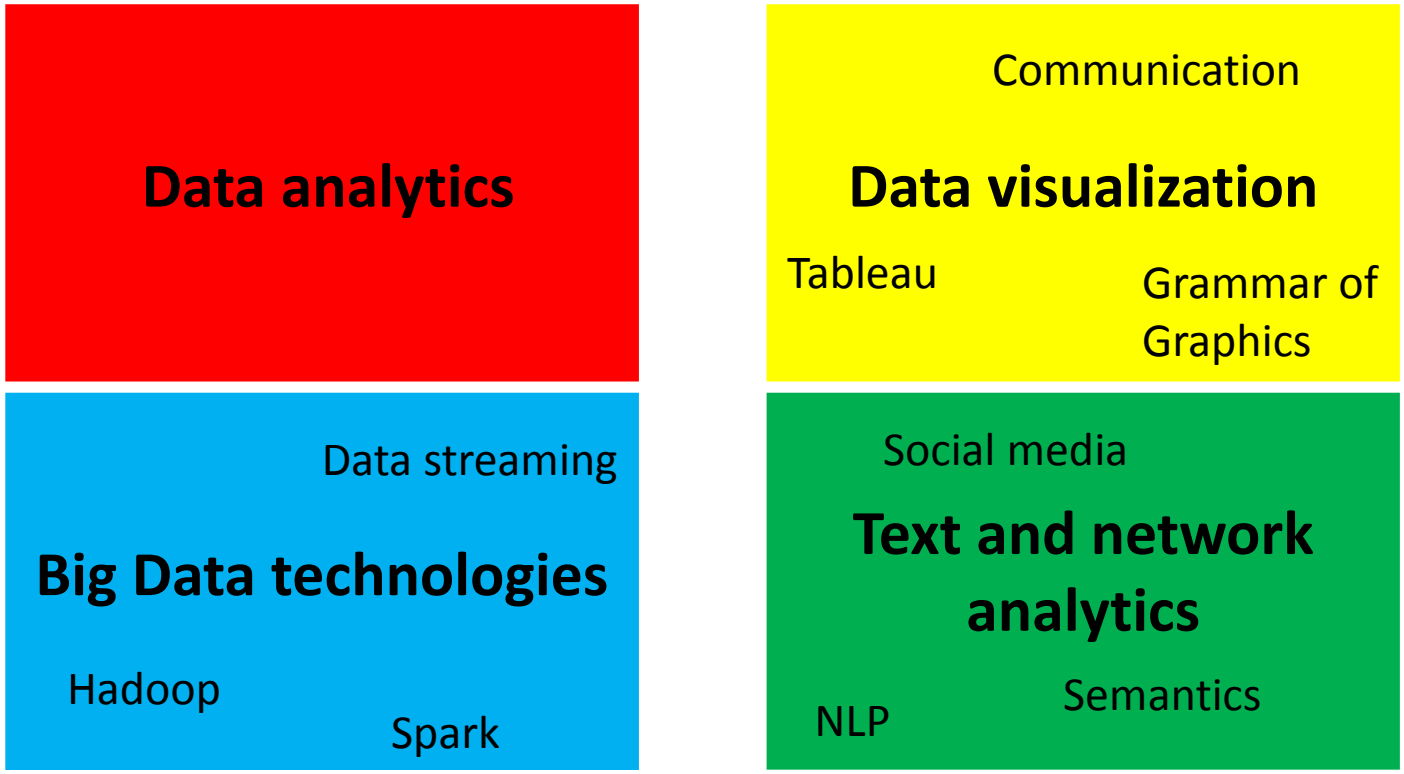
- Overview of data analytics
- Introduce key concepts for hands-on session
  - Logistic regression
  - Decision tree
  - Random forest
  - Evaluation metrics
  - Cross-validation
- Short break
- Hands-on

Overview of data analytics

What is data analytics?

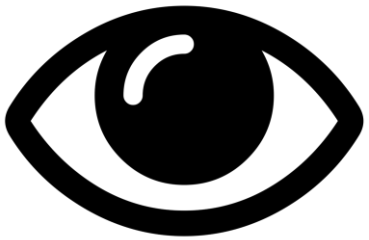
- A collection of established methods/techniques that
  - Seeks to make sense of and generate insights and knowledge from collected data (Big Data or otherwise)
  - Is statistically sound and rigorous
  - Preferably scalable
  - Is used to support decision making

Data Science



A common way to think about data analytics

Descriptive



Given existing data, generate some form of summary / aggregated view so that data can be consumed

Predictive

Given existing data, construct models so that predictions on future, yet-to-be collected data can be made

Prescriptive

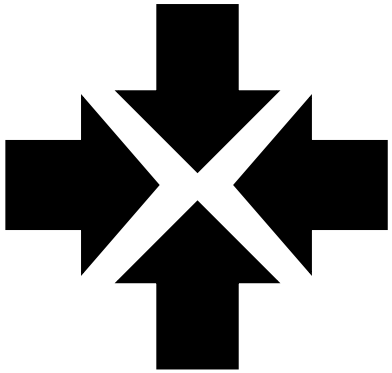
Given constructed models, recommend future decisions

Key aspects that businesses are concerned about

Accuracy

Value-adding

Interpretability



“Factors associated”

Key techniques in data analytics

- Feature selection
- Clustering
- Linear models
- Tree-based models
- Evaluation metrics
- Resampling methods
- Hypothesis testing
- Association rule mining
- Time series analysis
- Feature engineering

Statistical learning

Sometimes neglected, but nonetheless powerful

Statistical learning

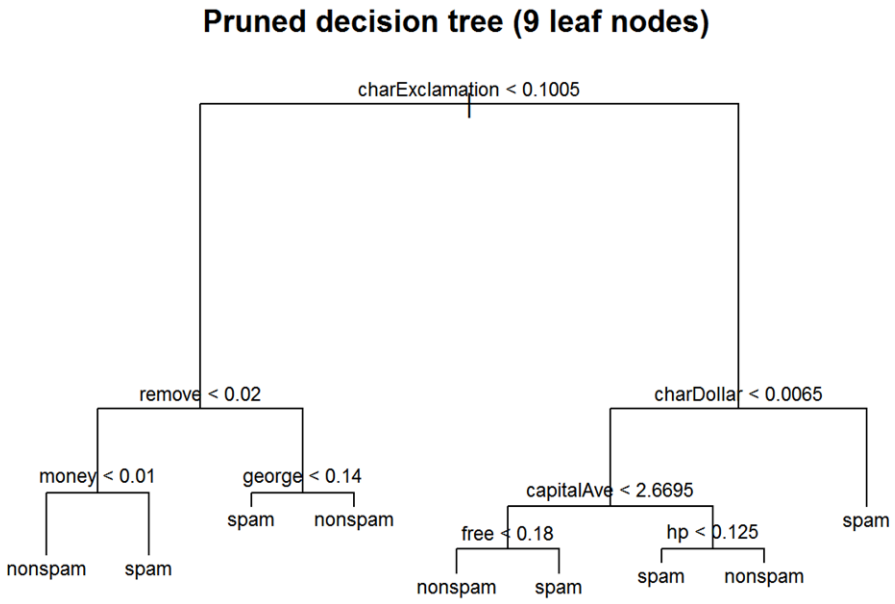
- Supervised and unsupervised learning (also, semi-supervised learning)
- Supervised learning: learning with ground truth/answers available (response variable)
  - Classification: response variable is categorical
  - Regression: response variable is continuous or numerical
- Unsupervised learning: finding intrinsic relationships between samples in the dataset
  - Clustering algorithms

Supervised learning: linear models

- Generalized linear models (GLM): mainstay tool in data analytics
- Generalized in the sense of the type of response variable:
  - Continuous response variable: ordinary least squares (OLS) regression
  - Binary / multinomial response: logistic regression
  - Discrete response: Poisson regression
- Gives an equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$
- Regularization (ridge regression, LASSO regression)

Supervised learning: tree-based models

- Models that uses decision trees as fundamental building blocks
  - Random forest
  - Gradient boosting machines
  - Rotation forest
- More on decision trees and random forest later



Unsupervised learning: clustering

- Clustering: empirically grouping observations / samples / rows in a dataset together in different groups (cluster), such that the more similar observations are grouped together
- Unsupervised because there is no ground truth to guide the process, unlike e.g. regression

Feature selection

- Feature: a variable / attribute in the dataset
- Feature selection: the process of selecting relevant features that aids in the modelling process, used especially when there are too many features in the dataset to work with
- Curse of dimensionality: the more irrelevant features are used in a model, the weaker the model

Evaluation metrics

- Measures using which constructed models are assessed
- Examples include accuracy and ROC-AUC
- Later

Key concepts

Hands-on session

- For the hands-on session, we will be look at a dataset of emails, consisting of both spam and non-spam
- The objective is to construct models that can predict whether a given email is spam or non-spam

Filter

nal	project	re	edu	table	conference	charSemicolon	charRoundbracket	charSquarebracket	charExclamation	charDollar	charHash	capitalAve	capitalLong	capitalTotal	type
0.00	0.00	0.00	0.00	0	0.0	0.000	0.000	0.000	0.778	0.000	0.000	3.756	61	278	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.132	0.000	0.372	0.180	0.048	5.114	101	1028	spam
0.12	0.00	0.06	0.06	0	0.0	0.010	0.143	0.000	0.276	0.184	0.010	9.821	485	2259	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.137	0.000	0.137	0.000	0.000	3.537	40	191	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.135	0.000	0.135	0.000	0.000	3.537	40	191	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.223	0.000	0.000	0.000	0.000	3.000	15	54	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.054	0.000	0.164	0.054	0.000	1.671	4	112	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.206	0.000	0.000	0.000	0.000	2.450	11	49	spam
0.30	0.00	0.00	0.00	0	0.0	0.000	0.271	0.000	0.181	0.203	0.022	9.744	445	1257	spam
0.00	0.06	0.00	0.00	0	0.0	0.040	0.030	0.000	0.244	0.081	0.000	1.729	43	749	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.000	0.000	0.462	0.000	0.000	1.312	6	21	spam
0.00	0.00	0.00	0.00	0	0.0	0.022	0.044	0.000	0.663	0.000	0.000	1.243	11	184	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.056	0.000	0.786	0.000	0.000	3.728	61	261	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.000	0.000	0.000	0.000	0.000	2.083	7	25	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.102	0.000	0.357	0.000	0.000	1.971	24	205	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.063	0.000	0.572	0.063	0.000	5.659	55	249	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.000	0.000	0.428	0.000	0.000	4.652	31	107	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.000	0.000	1.975	0.370	0.000	35.461	95	461	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.182	0.000	0.455	0.000	0.000	1.320	4	70	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.275	0.000	0.055	0.496	0.000	3.509	91	186	spam
0.00	0.00	0.00	0.00	0	0.0	0.000	0.729	0.000	0.729	0.000	0.000	3.833	9	23	spam
0.00	0.00	0.00	0.00	0	0.0	0.042	0.101	0.016	0.250	0.046	0.059	2.569	66	2259	spam
0.00	0.00	0.00	0.00	0	0.0	0.404	0.404	0.000	0.809	0.000	0.000	4.857	12	34	spam

Showing 1 to 24 of 4,601 entries

A bit on R

- R is a statistical computing language that was developed with statistical analysis in mind
- One of the most popular tools in the data science community
- R scripts: sequence of procedures that enables step-by-step customized data crunching
- R packages: collations of R scripts (functions) that we can leverage on to do various, more complex tasks easily, e.g. manipulate data and construct models
- R and Rstudio

Key concepts to be used

- Logistic regression
- Decision tree
- Random forest
- Cross-validation
- Evaluation metrics: accuracy and ROC-AUC

Logistic regression

- A type of generalized linear model (GLM)
- Assigns each variable used in the model with a coefficient that can be used in summation to predict log-odds
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$
- Log-odds = log(odds) $\propto$ probability
- In our case, probability of an email being a spam email

Pros and cons of logistic regression

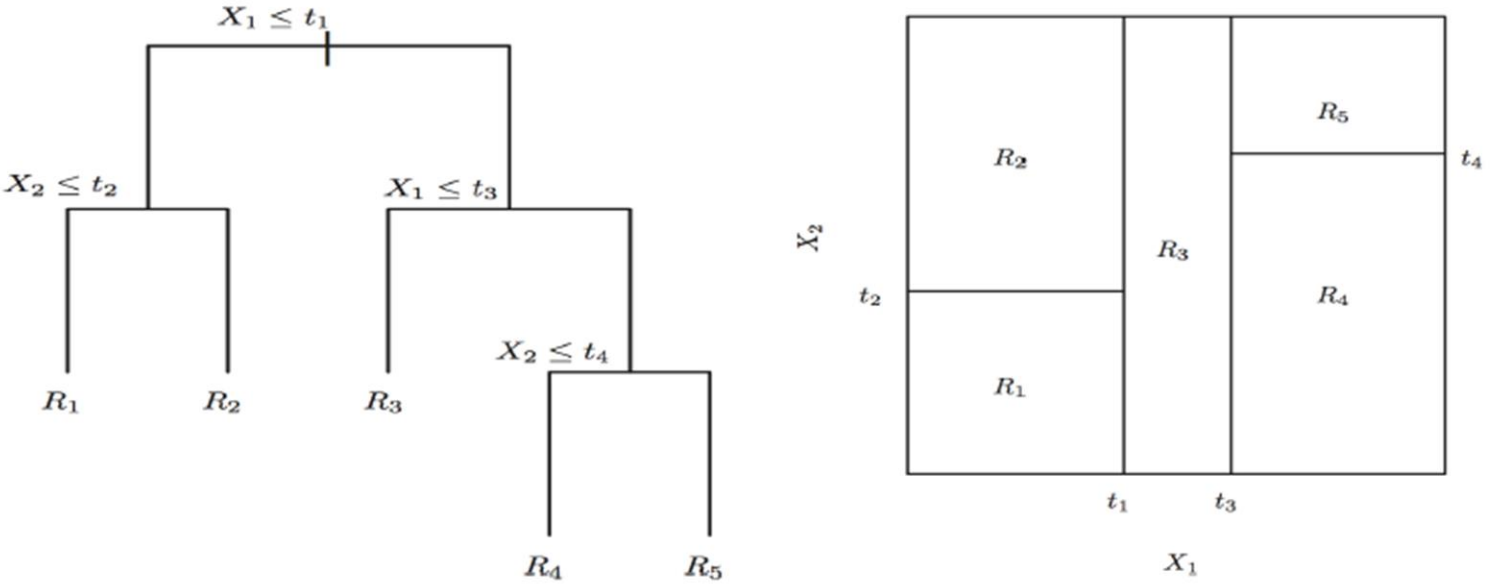
- Pros:
  - Easy to interpret – the idea of regression is familiar and intuitive
- Cons:
  - Requires certain statistical assumptions to hold true in the data
  - Generally low predictive accuracy

Decision trees

- A simple model used in supervised learning
- CART, C4.5 – amongst top 10 most popular data mining algorithms
- Can handle both classification and regression
- The **tree** package that we are using uses the recursive partitioning algorithm

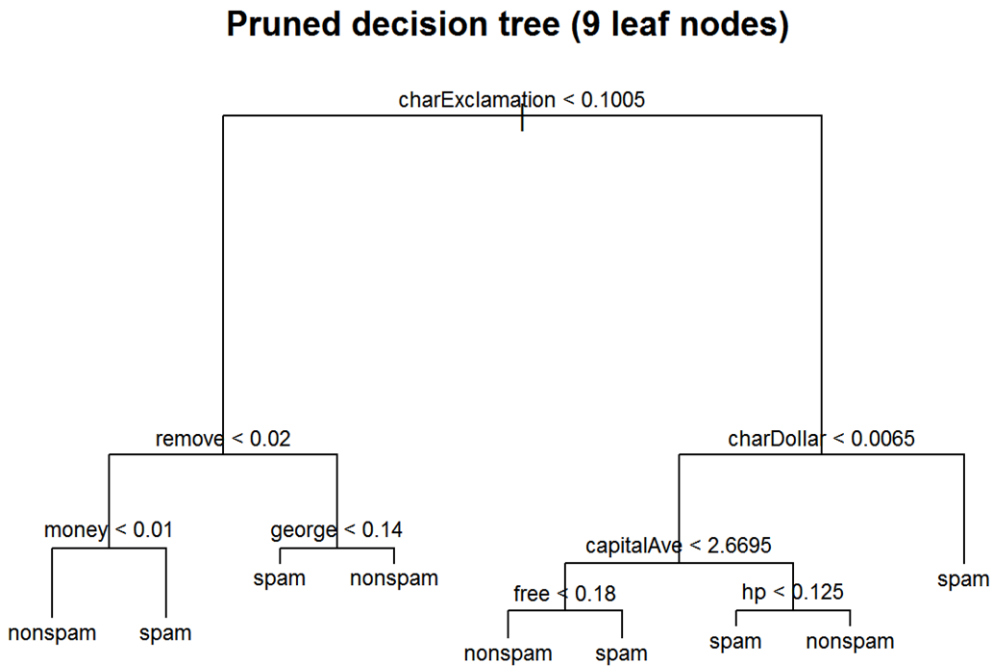
Equivalents

- Tree == Binary partitioning of dataset
- Each partition is represented by the mode (classification) or mean (regression)



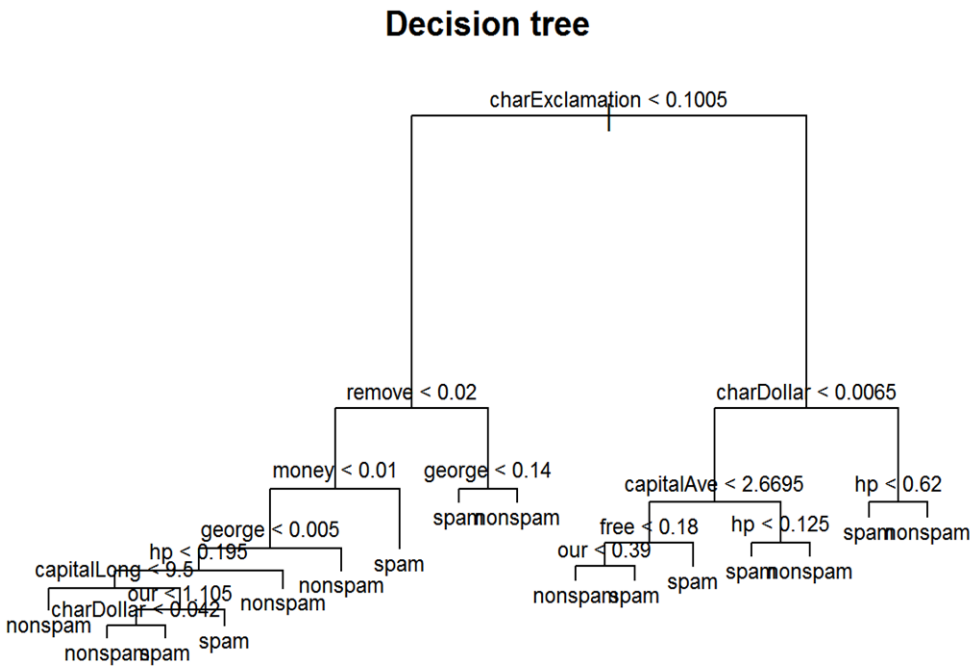
Terminologies

- Depth
- Node
  - Leaf nodes
  - Non-leaf nodes
- The size of a tree sometimes refers to the number of leaf nodes
- Parents and children
- Branching factor



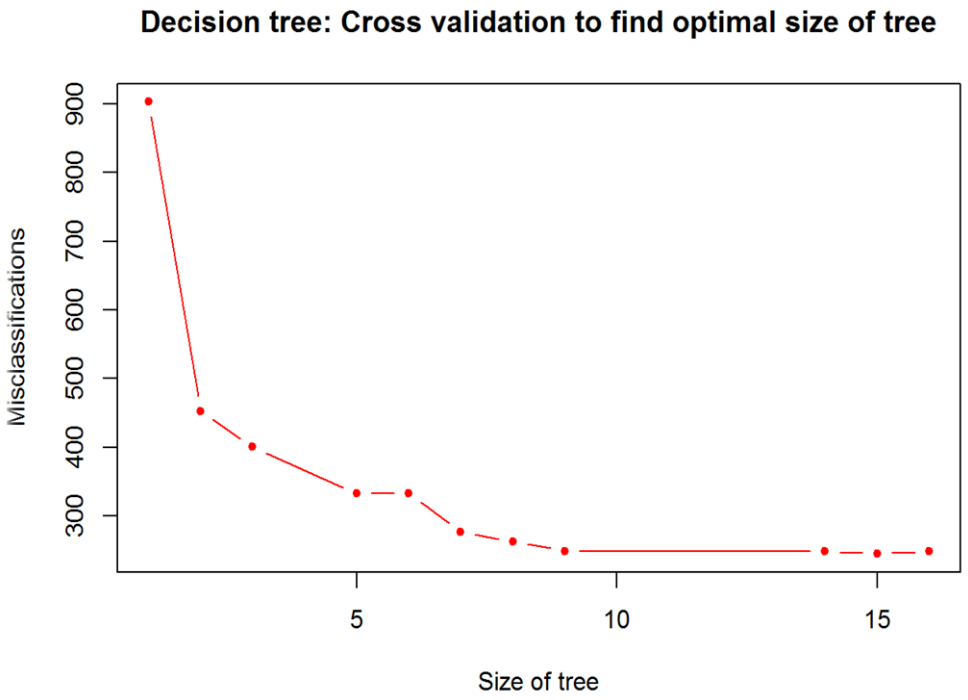
Pruning

- Typically after the construction of a decision tree, we would want to prune the tree, because the tree may be overly complicated



Pruning (2)

- Pruning refers to the process of trimming the tree to a more compact and concise one, without sacrificing much performance
- The **tree** package uses cost-complexity pruning
  - Comparing the relationship between number of leaf nodes and performance of model



Pros and cons of decision trees

- Pros:
  - Very easy to interpret and communicate to others, because it is similar to how humans think and make decisions
  - Easy to construct

- Cons:
  - Generally unstable
  - Generally low predictive accuracy

Random forest

- In the RF model, instead of using one decision tree to do predictions, we use multiple of them
- The idea is to build decision trees on different subsets of the training data
  - Each subset is known as a “bag”
  - Each bag yields one decision tree
- To make a prediction, we ask each tree to make a predictions
  - To get the overall prediction of the RF model, we take a majority vote

Pros and cons of random forest

- Pros:
  - One of the top-performing models in supervised learning
  - With some basic understanding of sampling and bootstrapping, RF can be easy to communicate. The intuition of voting as a mechanism to make decisions is simple
  - Able to derive variable importance measures
- Cons:
  - Computationally intensive

Evaluation metrics: assessing the performance of a supervised learning model

- In order to know whether the models constructed can perform well in reality, we need to assess some metrics to assess their performance
- Classification: accuracy / error rate
  - Sensitivity, specificity etc.
- Regression: mean squared error
  - $MSE = \frac{1}{n}\sum(prediction - actual)^2$

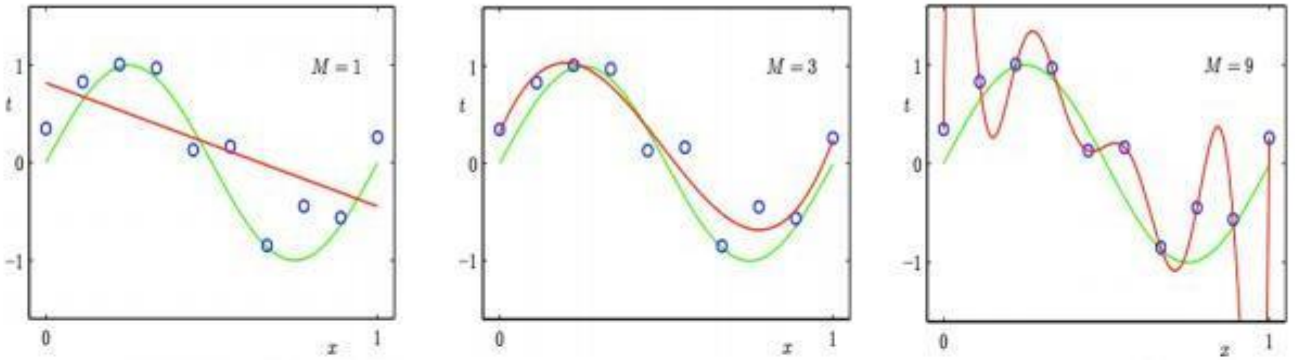
- Also, there are two types of classification models:
  (1) Those that output classes / categories as predictions
  (2) Those that output probabilities as predictions

- (2): can use ROC-AUC as a measure of performance

Cross-validation

- Gives rise to the idea of training and testing datasets
- Rationale:
  - Recall that the constructed models are ultimately meant to do predictions on future, unknown observations
  - Models are constructed/trained using input datasets. We call them training data
  - If the models constructed are too attuned to the training data => overfitting

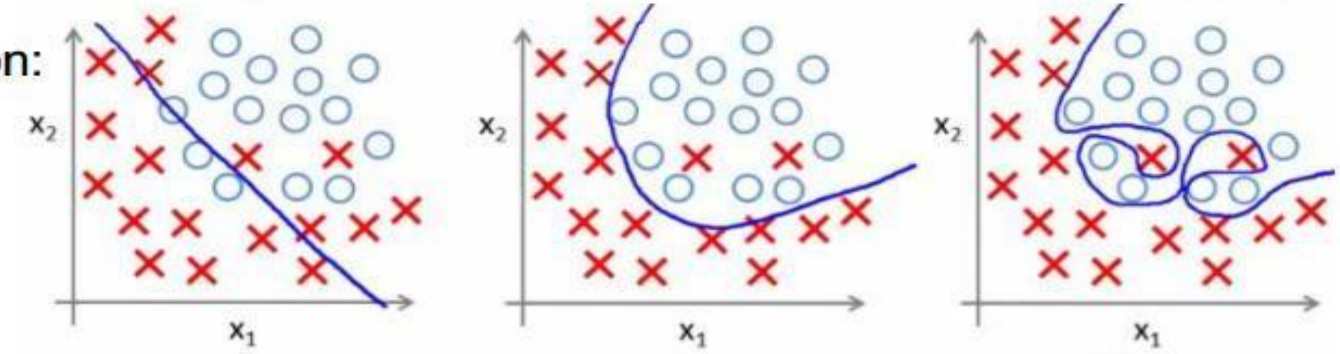
Regression:



predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:



http://www.turingfinance.com/regression-analysis-using-python-statsmodels-and-quandl/

Cross-validation

- In order to know whether our models are overfitted to the data, we use cross-validation
  - Split the dataset in two parts: training and testing
  - Use the training set to build the models
  - Use the models to make predictions on the testing set
- A way to think about this: studying for an examination

Hands-on

Thanks!
Questions?

github.com/tohweizhong/NUS-DataScience

sg.linkedin.com/in/tohweizhong

tohweizhong@u.nus.edu