NOTES 1

Introduction to Data Mining

Acknowledgement: some of the contents are borrowed with or without modification from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R.Tibshirani.

What is Data Mining

- “Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.”

- “Methods for automatically learning and recognizing complex patterns from data.”

Business Analytics Procedure

- Understand your problem

Notice the difference of statistical hypothesis and data mining;
In Data mining, we usually forced to have data rather than hypothesis.

- Data Collection
  - Availability

data structure:
** Size; data type

- Data Processing
  - Exploratory Data Analysis
  - Missing Values
  - Data cleaning
  - Reduce your data

especially data visuallization

- Model Development

The more data u have, the more presentative of model;
However, if we use all data to build model, then we'll lack the data to test the model that we built

- Model Assessment and Interpret the results
- Deploy models, assess results, update models, design new models

Models Covered in Class

- Unsupervised Learning: Measurements for each observation, but no associated response
  - Given measurements $X_1; : : : X_n$, learn some underlying group structure based on similarity. We don't have $Y_1; : : : Y_n$,
  - Clustering Ch.10
  - Association Rules Tan et al. Ch. 5

eg: use historical order to predict users' products that they might be interested in.

We don't use Y to build model, we use Y to measure our model that is built.

- Supervised Learning: For each training example both the input variables and the associated response are available
  - Given measurements $(X_1; Y_1); : : : (X_n; Y_n)$, learn a model to predict Y_i from X_i

Supervised Learning: Classification and Regression

All Supervised Learning; The only difference is the Y is continuous or not.

- Regression: the output Y is quantitative or continuous or numerical;
  - Linear Regression: Ch.3
  - Performance Evaluations: AIC, BIC, MSE and etc. Ch 6.1

- Classification: the output Y is qualitative or categorical or binary or discrete;
  - Classification Tree, Random Forest, Bagging Ch.8
  - KNN Ch.4
  - Logistic regression Ch.4
  - Performance Evaluation: precision-recall and etc. Ch.4

Model Assessment

- Training/testing dataset

- Resampling methods: Ch. 5
  - Cross-Validation
  - Bootstrap

One group for testing, others used for training.

Bias-Variance Tradeoff

There always have a trade-off between bias and vaiance

- You always only have limited data…..*variance*
- There is no perfect model…..*bias*
- *Variance* refers to the amount by which your model would change if we estimated it using a different training data set

High:
too acurate

- *Bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

too simple

High:
general

Prediction and Inference

- *Prediction*: predict the output Y given inputs X using the model

- *Inference*: understand relationship between Y and predictors X
  - Avoid the model to be a black box

Interpretability-Flexibility Tradeoff

- A model with the best prediction performance may be a hard-to-interpret **black box**

- Depending on your purpose, we may select a more interpretable but with less accuracy model.