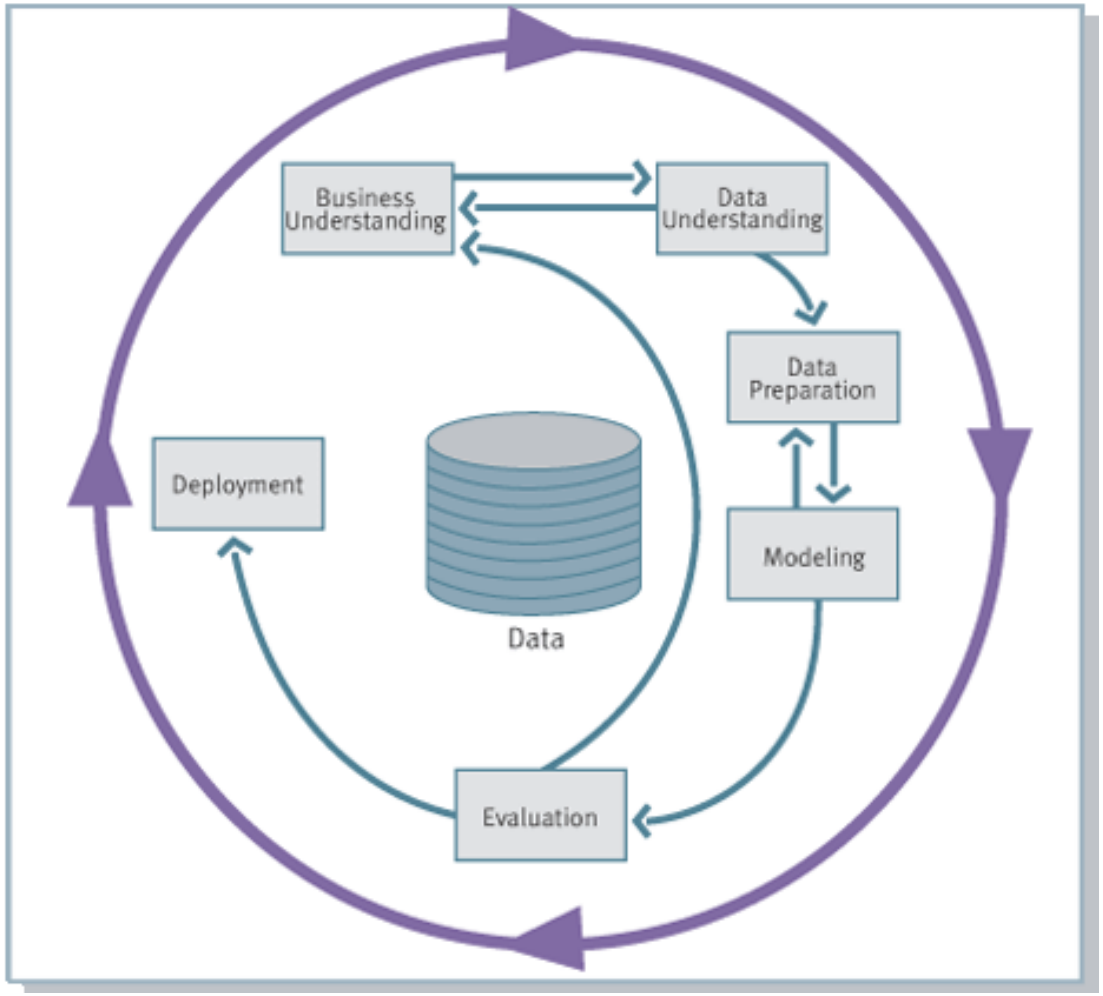
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Business Analytics (MM 5425)

Course Review

Dr. Yue (Katherine) FENG

L2 – Conduct BA from a Process View



Question example:

How to apply the BA technique to solve a problem in one business context? (Individual Assignment)

L2 – Terminologies Clarification

> Data: record/instance/example

> Target Variable vs. Feature/Attribute

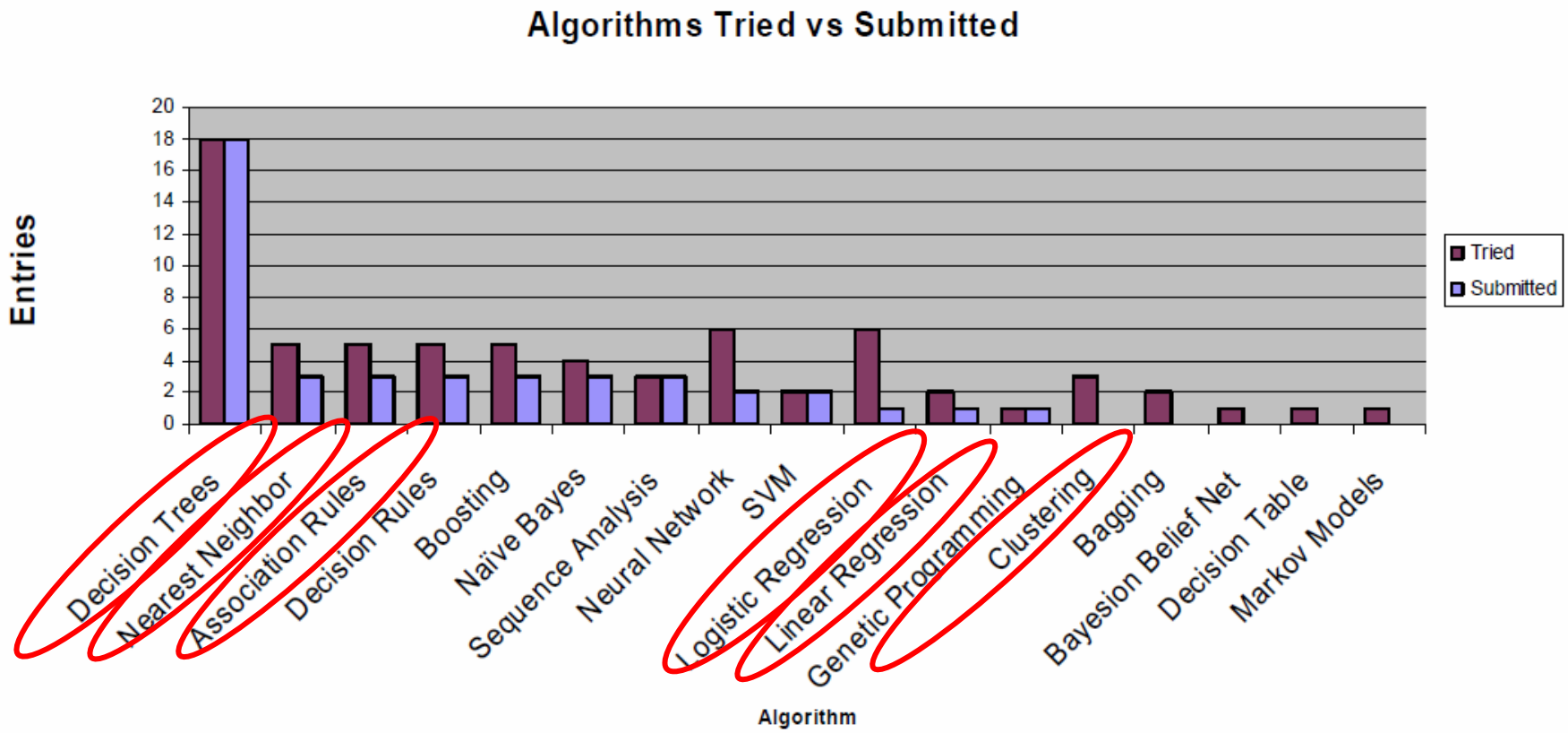
> Supervised vs. Unsupervised Learning

> Classification vs. Regression (Both are supervised learning)

> Philosophy of Predictive Modeling

> Training vs. Testing

Algorithms Taught in MM5425



Summary for Key Learning Algorithms

Supervised Learning

- **Decision Tree – *Classification & Regression***
  - Interpretable, build-in variable selection, non-linear, fast, may overfit
- **Linear Regression – *Regression***
- **Logistic Regression – *Classification***
  - Probability estimation, numeric interpretation, naturally avoid overfitting
- **K-Nearest Neighbor (KNN) – *Classification & Regression***
  - Memory-based, slow, sensitive to the similarity measure

Unsupervised Learning

- **Association rule**
  - Rule generation, interpretable, unfocused
- **Clustering**
  - Similarity-based pattern, not stable, natural groups

Decision Tree

> Tree representation and terminology

> Classification tree induction

- Entropy and information gain

> Class probability and the use of class probability

> Geometric interpretation

> Pros and cons of decision tree

> Regression tree

Linear/Logistic Regression

> A brief idea on Linear Regression

> Logistic Regression - Model Setup

- Parameter estimation (optional)
- Model Interpretation

> Decision Tree vs. Logistic Regression

KNN

> One application based on similarity for classification/regression

> Flow of KNN

> Selection of *k*: overfitting control (if *k* is too small or too large)

> Strength and weakness of KNN

- KNN: Memory-based Learning
- One application of KNN: Collaborative Filtering (optional)

Overfitting

> Symptom of overfitting problem

> Overfitting control for decision tree

- Pruning a tree: the role of hyperparameters (e.g., max tree depth, min samples in a node required for splitting, min impurity decrease required, etc.)
- How to find the best hyperparameters: training, validation, and testing

Model Evaluation

> Cross-Validation

> Accuracy

> Confusion Matrix

- Precision, recall, F1 score
- Bad Positives and Harmless Negatives (uneven cost)
- Using expectation value for model evaluation (cost/benefit matrix)

> ROC Analysis

- The role of decision threshold
- The meaning of special points on ROC curve

> Performance evaluation for regression

Association Rule

> Basic idea of association rule and its application

> How to generate association rules?

- Support, confidence, lift and leverage
- 3 steps to find association rules
- APRIORI algorithm: find frequent itemsets (optional)

> Pros and Cons of Association Rule

Clustering

> Basic idea: difference from association rule and application

> Similarity/distance measure

> Flow of K-means clustering

> Several problems of clustering

- Center initialization
- Selection of *k*: elbow method + domain knowledge + practical use

> How to evaluate clustering results

- Supervised vs. Unsupervised Learning

Discussions (True or False)

- **Question 1**: After a model is generated and tested by performing a supervised learning algorithm, the model will be eventually applied to new data in which the value of the target variable is known.

- **Question 2**: The choice of learning algorithm is the only factor determining the accuracy of predictive model generated.

Discussions

Q: You would like to build a model to **classify online posts** (e.g., online review, tweets) by whether the users express **positive or negative sentiment** about some product. Your goal is to tell if a new post is positive or negative sentiment using the model you built. Please describe the steps of data analysis you will take to achieve your goal.

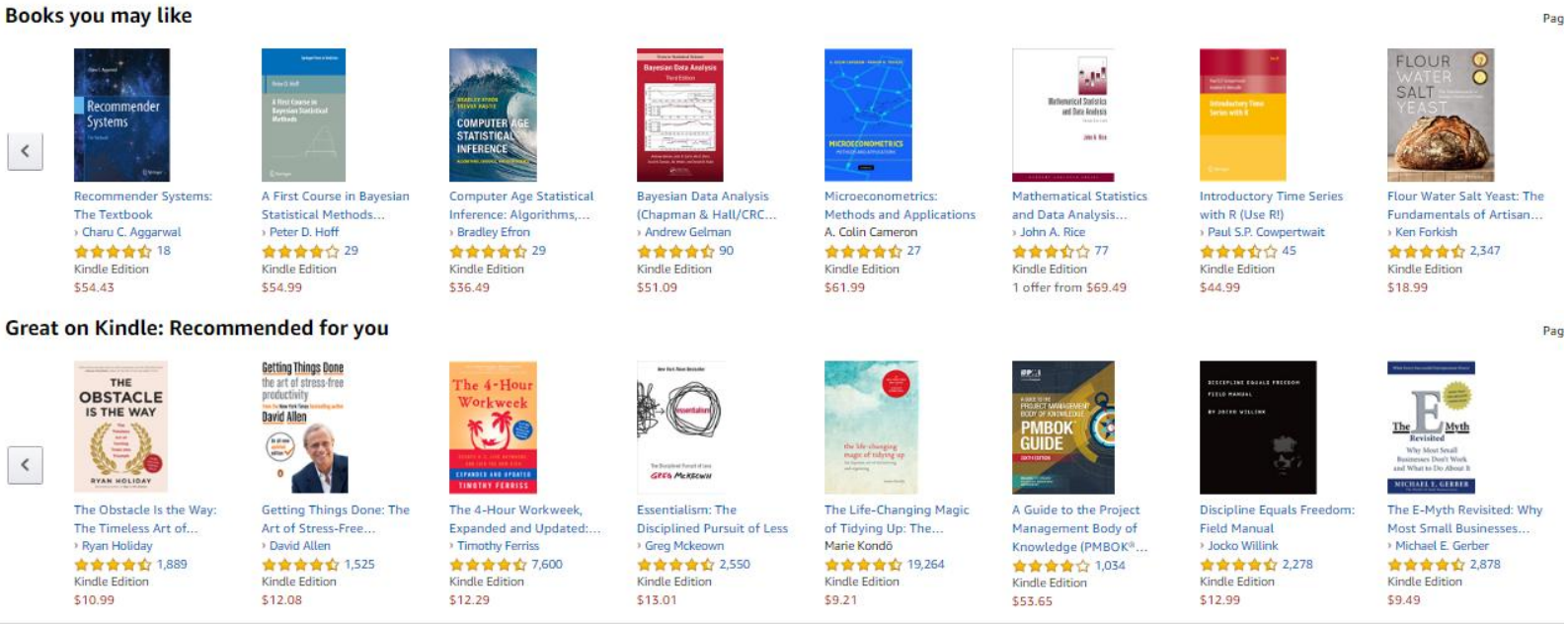
- Formulate problem
- Prepare data and preprocess: labeling, convert vector of words to numeric measures
- Modeling: supervised learning
- Evaluation: simple hold-out or cross-validation
- Apply and deployment

Discussion - Recommender Systems

> How to use association rule learning in recommender systems?

Discussion - Recommender Systems

> Another perspective to generate recommendations



Discussion:
Will you make decisions based on BA or human insights? Why?

Thank You !

IDEAS Innovation-driven Education and Scholarship | Faculty of Business 工商管理學院

Discover · Design · Deliver

Department of MANAGEMENT & MARKETING 管理及市場學系

Opening Minds • Shaping the Future
啟迪思維 • 成就未來