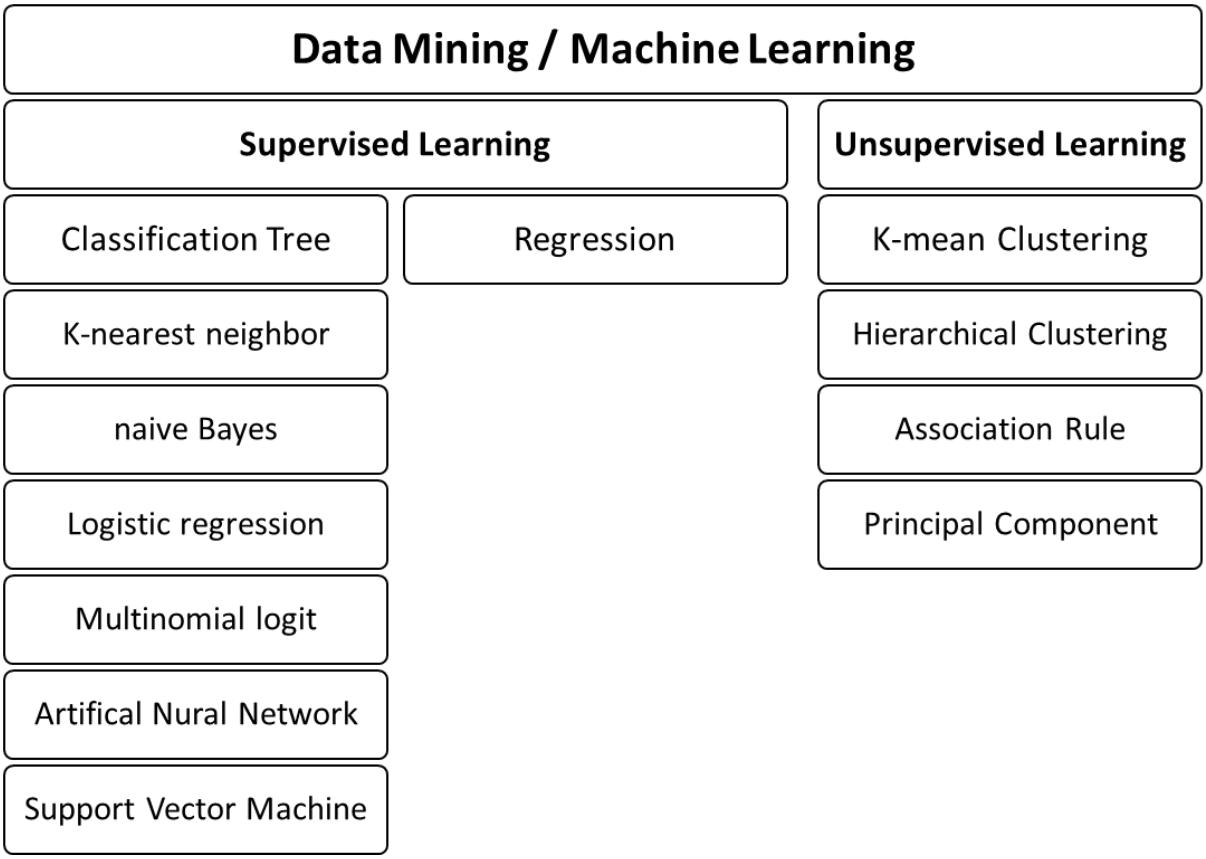
Chapter 8 Summary

Overview of methods



I Procedure for Data Mining

1. Getting to know the data and variables

The information about the dataset and variables is of great importance. Without this information, the dataset is just a collection of meaningless numbers. The success of data mining relies heavily on the understanding of the data and variables.

2. Data preparation

Data need to be process before they can be used in the analysis.

2.1 **Outlier detection**

This is important particular for the data mining methods sensitive to outliers.

2.2 **Transformation of continuous variables**

Often continuous variables are transformed using standardize transformation or using range scaling to [0,1] interval. Other transformations such as log-transformation are often used to make the normality assumption more plausible.

2.3 **Transformation of categorical variables**

The number of levels of categorical variables can be reduced by combining levels with similar meaning or fewer records.

2.4 **Variable selection**

Usually we do not want to include all variables in the analysis. We only include those useful and importance. Some methods, such as CTREE may helps to identify the importance variables.

3. Analyzing techniques

Methods to be used depend on the dataset and situation. Keep in mind that an all-time winner does not exist. Some methods may work well on this dataset but may work poorly on other data. We may even combine several methods to give us a better result.

3.1 **CTEE:** rule discovery, simple and easy to interpret. Robust to outliers and identify important variables. Random forest may help but computationally intensive.

3.2 **Knn**: Lazy learner, distance-based method. Quite robust to outliers.

3.3 **Naïve Bayes classifier**: Lazy learner based on Bayes reasoning. Having a good probability interpretation and give the posterior probability of each record. Sensitive to outlier.

3.4 **Logistic regression / multinomial logit**: extension of the familiar regression method to binary and categorical response. Sensitive to outliers.

3.5 **ANN**: Flexible and work well in many complicated domains. Work like a black-box and hard to interpret. Can be used to predict continuous response as well. Recent development of **Deep learning** based on ANN has great achievements in many applications. This becomes the most active research in Artificial Intelligence.

3.6 **K-means clustering**: Unsupervised learning method to discover homogeneous and heterogeneous group of data.

3.7 **Association analysis**: Unsupervised learning and rule discovery method. Based on simple frequency counting of records.

4. Assessment and refinement

4.1 Usually the dataset is large enough to partition into training dataset and testing dataset. The testing dataset often consist of 1%-30% of the total records. If the data are too noisy, we may need more for the training dataset.

4.2 Training dataset is used to build or to train the model while the testing dataset is used to assess the accuracy of the model.

4.3 Many Data Mining methods have parameters need to be tuned and refined. Classification error rate as well as the lift chart is used to assess and compare different methods.

II Comparison of Data Mining methods

Here is the comparison of classification table and error rate of all the methods and examples used in this course.

IRIS data

CTREE: (using whole dataset)	**Multinomial logit:**
cl 1 2 3 (error=6/150=4%) 1 50 0 0 2 0 49 5 3 0 1 45	pred 1 2 3 (error=2/150=1.33%) 1 50 0 0 2 0 49 1 3 0 1 49
knn cl 1 2 3 (error=3/50=6%) 1 14 0 0 2 0 17 1 3 0 2 16	**Naïve Bayes** cl 1 2 3 (error=3/50=6%) 1 14 0 0 2 0 18 2 3 0 1 15
ANN: (linear output) 1 2 3 (error=1/150=0.07%) 1 50 0 0 2 0 49 1 3 0 0 50	**ANN: (logistic output)** 1 2 3 (error=1/150=0.07%) 1 50 0 0 2 0 49 0 3 0 1 50

HMEQ data

CTREE: Training data cl 0 1 (error=119/2045=5.8%) 1 1840 117 2 2 86	**CTREE: Testing data** cl 0 1 (error=52/1022=5.1%) FALSE 943 48 TRUE 4 27
	Random Forest: Testing data cl 0 1 (error=48/1022=4.7%) FALSE 946 47 TRUE 1 28
knn cl 0 1 (error=102/1022=9.98%) 0 910 91 1 11 10	knn (scale) cl 0 1 (error=54/1022=5.28%) 0 915 48 1 6 53
knn (stand) cl 0 1 (error=44/1022=4.31%) 0 919 42 1 2 59	**Naïve Bayes** cl 0 1 (error=70/1022=6.84%) 1 912 61 2 9 40
Logistic regr: Training data pr 0 1 (error=147/2045=7.2%) 0 1856 135 1 12 42	**Logistic regr: Testing data** cl 0 1 (error=77/1022=7.5%) 0 912 68 1 9 33
ANN: Training data pr 0 1 (error=110/2045=5.4%) 0 1867 109 1 1 68	**ANN: Testing data** pred 0 1 (error=58/1022=5.7%) 0 920 57 1 1 44

Titanic data

CTREE: Training data cl no yes(error=413/1980=20.8%) 1 1322 394 2 19 245	**CTREE: Testing data** cl no yes (error=48/221=21.7%) 1 148 47 2 1 25
knn cl 0 1 (error=48/221=21.7%) 0 148 47 1 1 25	**Naïve Bayes** cl no yes (error=54/221=24.4%) 1 136 41 2 13 31
ANN: Training data cl no yes(error=413/1980=20.8%) 1 1322 394 2 19 245	**ANN: Testing data** cl no yes (error=48/221=21.7%) 1 148 47 2 1 25