MANCHESTER 1824

The University of Manchester

# Revision (Part II)

**Ke Chen**

Revision slides are going to summarise all you have learnt from Part II, which should be helpful for you to prepare your exam in January along with those non-assessed exercises.

# Naïve Bayes Classifier

- Probabilistic Classifiers
  - discriminative vs. generative classifiers
  - Bayesian rule used to convert generative to discriminative
- Naïve Bayesian Assumption
  - Conditionally independent assumption on input attributes
- Naïve Bayes classification algorithm
  - Estimate conditional probabilities for each attribute given a class label and prior probabilities for each class label
  - MAP decision rule
- Relevant issues
  - Zero conditional probability due to short of training examples
  - Applicability to problems violating the naïve Bayes assumption

# Clustering Analysis Basics

- Clustering Analysis Task
  - discover the “natural” clustering number
  - properly grouping objects into sensible clusters
- Data type and representation
  - Data type: continuous vs. discrete (binary, ranking, ...)
  - Data matrix and distance matrix
- Distance Measure
  - Minkowski distance (Manhattan, Euclidean ...) for continuous
  - Cosine measure for nonmetric
  - Distance for binary: contingency table, symmetric vs. asymmetric
- Major Clustering Approach
  - Partitioning, hierarchical, density-based, graph-based, ensemble

# K-Means Clustering

- Principle
  - A typical partitioning clustering approach with an iterative process to minimise the square distance in each cluster
- K-means algorithm
  1) Initialisation: choose K centroids (seed points)
  2) Assign each data object to the cluster whose centroid is nearest
  3) Re-calculate the mean for each cluster to get a updated centroid
  4) Repeat 2) and 3) until no new assignment
- Relevant issues
  - Efficiency: O(*tkn*) where $t, k << n$
  - Sensitive to initialisation and converge to local optimum
  - Other weakness and limitations
  - Clustering validation

# Hierarchical Clustering

- Hierarchical clustering
  - Principle: partitioning data set sequentially
  - Strategy: divisive (top-down) vs. agglomerative (bottom-up)
- Cluster Distance
  - Single-link, complete-link and averaging-link
- Agglomerative algorithm
  1) Convert object attributes to distance matrix
  2) Repeat until number of cluster is one
     - Merge two closest clusters
     - Update distance matrix with cluster distance
- Relevant concepts and techniques
  - Construct a dendrogram tree
  - Life-time of clusters achieved from a dendrogram tree
  - Determine the number of clusters with maximum *k* life-time

# Cluster Validation

- Cluster Validation
  - Evaluate the results of clustering in a *quantitative* and *objective* fashion
  - Performance evaluation, clustering comparison, find cluster num.
- Two different types of cluster validation methods
  - Internal indexes
    - No ground truth available and sometimes named “relative index”
    - Defined based on “common sense” or “a priori knowledge”
    - Variance-based validity indexes
    - Application: finding the “proper” number of clusters, ...
  - External indexes
    - ground truth known or reference given
    - Rand Index
    - Application: performance evaluation of clustering, clustering comparison...
  - There are many validity indexes, still an active research area in unsupervised learning

# Examination Information

- Three Sections (total 50 marks)
  - Section 1 (20 marks)
    - 20 multiple choice questions totally
    - Questions 11-20 relevant to Part II
  - Section 2 (15 marks)
    - One compulsory question relevant to Part I
  - Section 3 (15 marks)
    - One compulsory question relevant to Part II
- Length: two hours
- Calculator (without memory) allowed