CS256 – Midterm Exam Study Guide

By: Zayd Hammoudeh

Classification	**Training Set**	**Model**	**Test Set**	**Example Classification Techniques**
Task of assigning objects to one of several predefined categories.	A collection of records. Each **record** contains a set of attributes one of which is the **class**.	A function from the value of record attributes to the class attribute.	A collection of records used to determine the accuracy of the classification model.	1. **Neural Networks** 2. **Decision Tree** 3. **Rule Based Classifier** 4. **Memory Based Reasoning** 5. **Support Vector Machines** 6. **Naïve Bayes and Bayesian Belief Networks**

Induction

Using a training set to generate a model.

Deduction

Process of applying a model to a training set.

Decision Tree Induction

- **Greedy Strategy**
- **Key Decision #1:** Attribute to expand next
- **Key Decision #2:** When to stop expanding

Hunt's Decision Tree Induction Algorithm:

- Let D_t be the set of training records that reach a node t.

1. If D_t contains records that **all belong to the same class y_t**, then t is a leaf node with class value y_t.
2. If D_t is an **empty set**, then t is a leaf node with default value y_d.
3. If D_t contains **records that belong to more than one class and there are no attributes left**, then t is a leaf node with default value is a leaf node with default value y_d.
4. If D_t contains **records that belong to more than one class**, then use an attribute test to split the data into smaller subsets. Recursively apply the same procedure above.

Attribute Types

- **Binary** – Attribute with exactly two possible values.
- **Nominal** – Two or more class values with no intrinsic Order
- **Ordinal** – Two or more class values that can be ordered or ranked
- **Continuous** – Quantitative attribute that can be measured along a continuum.

Splitting Nominal and Ordinal Attributes

- **Binary** – Divides attribute values into two subsets. **This requires the additional step of finding optimal partitioning.**
- **Multi-way** – Use as many partitions as distinct values.

Splitting Based on Continuous Attributes

- **Discretization** – Form an ordinal categorical attribute.
  - **Static** – Discretize once at the beginning
  - **Dynamic** – Ranges can be found by equal interval bucketing, equal frequency bucketing, or clustering.
- **Binary Decision** (A < v or A >v) – Consider all possible splits and find the best cut.

Homogeneity/Low Impurity – Extent to which nodes in the decision tree have the same class value/distribution.

Nodes with high levels of homogeneity (i.e. low levels of impurity) are preferred.