Notes 03-1: Introduction to Classification

Classification techniques attempt to derive a model of the data that assigns labels to objects that describe and distinguish classes of objects with similar properties.

A *training set* (i.e., objects whose class label is already known) is used to derive specific parameters of the model (known as the *training phase*).

The derived model is used to predict the class label of a previously unseen or unknown object (known as the *classification phase*).

More formally, the classification problem is defined as follows: Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of tuples and a set of classes $C = \{C_1, C_2, \ldots, C_m\}$, the classification problem is to define a mapping $f : D \rightarrow C$, where each t_i is assigned to one class. A class, C_j, contains precisely those tuples mapped to it; that is, $C_j = \{t_i \mid f(t_i) = C_j\ (1 \leq i \leq n) \wedge t_i \in D\}$.

Three basic methods used to solve the classification problem include: specifying boundaries, using known probability distributions, and using posterior probabilities.

Specifying boundaries: Divide the input space of potential database tuples into regions, where each region is associated with one class.

Example – Specifying boundaries

EXAMPLE = Classification.A.3.b

Using known probability distributions: For any given class, C_j, $P(t_i \mid C_j)$ is the probability density function for the class evaluated at t_i. If the probability of occurrence for each class, $P(C_j)$, is known, then $P(C_j)P(t_i \mid C_j)$ is used to estimate the probability that t_i is in class C_j. Assign t_i to the class with the highest probability.

Example – Using known probability distributions

Tuple	Gender	Height	Class
t_1	*f*	1.6	*short*
t_2	*m*	2.0	*tall*
t_3	*f*	1.9	*medium*
t_4	*f*	1.8	*medium*
t_5	*f*	1.7	*short*
t_6	*m*	1.85	*medium*

t_7	*f*	1.6	*short*
t_8	*m*	1.7	*short*
t_9	*m*	2.2	*tall*
t_{10}	*m*	2.1	*tall*
t_{11}	*f*	1.8	*medium*
t_{12}	*m*	1.95	*medium*
t_{13}	*f*	1.9	*medium*
t_{14}	*f*	1.8	*medium*
t_{15}	*f*	1.75	*medium*

Note: The values in the Tuple column are not the t_i's referred to in the probabilities in the explanation of using known probability distributions above. The values in the Tuple column are just the unique identifiers assigned to each tuple. The t_i's in the probabilities correspond to particular attribute values in the Gender and Height columns.

Consider those tuples where t_i = 1.9 and C_j = *medium*. What is the probability that 1.9 is in the *medium* class? Now,

$$P(1.9|medium) = \frac{\#(medium \wedge 1.9)}{\#(medium)} = \frac{2}{8} = 0.25$$

and

$$P(medium) = \frac{\#(medium)}{N} = \frac{8}{15} = 0.53,$$

where N = the number of tuples. So,

$$P(medium)P(1.9|medium) = (0.53)(0.25) = 0.1325.$$

Using posterior probabilities: Given a data value t_i, determine the probability that t_i is in C_j, denoted as $P(C_j \mid t_i)$ and known as the posterior probability. Determine the posterior probability for each class containing t_i and then assign t_i to the class with the highest probability.

Example – Using posterior probabilities

Tuple	Gender	Height	Class
t_1	*f*	1.6	*short*
t_2	*m*	2.0	*tall*
t_3	*f*	1.9	*medium*

t_4	*f*	1.8	*medium*
t_5	*f*	1.7	*short*
t_6	*m*	1.85	*medium*
t_7	*f*	1.6	*short*
t_8	*f*	1.8	*tall*
t_9	*m*	1.7	*short*
t_{10}	*m*	2.2	*tall*
t_{11}	*m*	2.1	*tall*
t_{12}	*f*	1.8	*medium*
t_{13}	*m*	1.95	*medium*
t_{14}	*f*	1.9	*medium*
t_{15}	*f*	1.8	*medium*
t_{16}	*f*	1.75	*medium*
t_{17}	*m*	1.8	*tall*

Consider those tuples where t_i = 1.8. What class should 1.8 be assigned to? Now,

$$P(medium|1.8) = \frac{\#(1.8 \wedge medium)}{\#(1.8)} = \frac{3}{5} = 0.6,$$

$$P(tall|1.8) = \frac{\#(1.8 \wedge tall)}{\#(1.8)} = \frac{2}{5} = 0.4,$$

and

$$P(short|1.8) = \frac{\#(1.8 \wedge short)}{\#(1.8)} = \frac{0}{5} = 0.0.$$

Therefore, 1.8 should be assigned to the *medium* class.

Machine Learning Approach to Classification

The machine learning approach to classification is based on using posterior probabilities.

The probabilities are inferred from data and represented as a model. The model can be represented as:

- Decision lists (a decision list is an ordered list of `if-then` rules)

- Decision trees

- Mathematical formulas
- Neural networks
- Etc.

Example – Soybean disease classification

The data consists of 680 descriptions of examples of diseased soybean plants (i.e., each example represents one plant).

Each example is represented by 35 attributes, each describing a different characteristic of the diseased plant.

Sample Attributes	# Possible Values	Sample Value
plant height	2	*normal*
seed treatment	3	*fungicide*
leaf condition	2	*abnormal*
stem condition	2	*normal*

Each example is labeled with one of 17 diseases as determined by the diagnosis of an expert in plant biology.

An `if-then` rule learned from the data

```
if leaf condition = normal
   and stem condition = abnormal
   and stem canker = below soil line
   and canker lesion color = brown
then
   diagnosis = root rot
```

A decision tree could be constructed where one of the paths from the root to a leaf corresponds to the `if-then` rule given in the previous example.