# 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* = {*t1*, *t2*, …, *tn*} of tuples and a set of classes *C* = {*C1*, *C2*, …, *Cm*}, the classification problem is to define a mapping *f* : *D* → *C*, where each *ti* is assigned to one class. A class, *Cj*, contains precisely those tuples mapped to it; that is, *Cj* = {*ti* | *f*(*ti*) = *Cj* (1 ≤ *i* ≤ *n*) ∧ *ti* ∈ *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, *Cj*, *P*(*ti* | *Cj*) is the probability density function for the class evaluated at *ti*. If the probability of occurrence for each class, *P*(*Cj*), is known, then *P*(*Cj*)*P*(*ti* | *Cj*) is used to estimate the probability that *ti* is in class *Cj*. Assign *ti* to the class with the highest probability.

Example – Using known probability distributions

|  |  |  |  |
| --- | --- | --- | --- |
| Tuple | Gender | Height | Class |
| *t1* | *f* | 1.6 | *short* |
| *t2* | *m* | 2.0 | *tall* |
| *t3* | *f* | 1.9 | *medium* |
| *t4* | *f* | 1.8 | *medium* |
| *t5* | *f* | 1.7 | *short* |
| *t6* | *m* | 1.85 | *medium* |
| *t7* | *f* | 1.6 | *short* |
| *t8* | *m* | 1.7 | *short* |
| *t9* | *m* | 2.2 | *tall* |
| *t10* | *m* | 2.1 | *tall* |
| *t11* | *f* | 1.8 | *medium* |
| *t12* | *m* | 1.95 | *medium* |
| *t13* | *f* | 1.9 | *medium* |
| *t14* | *f* | 1.8 | *medium* |
| *t15* | *f* | 1.75 | *medium* |

Note: The values in the Tuple column are not the *ti*’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 *ti*’s in the probabilities correspond to particular attribute values in the Gender and Height columns.

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

and

,

where *N* = the number of tuples. So,

.

*Using posterior probabilities*: Given a data value *ti*, determine the probability that *ti* is in *Cj*, denoted as *P*(*Cj* | *ti*) and known as the posterior probability. Determine the posterior probability for each class containing *ti* and then assign *ti* to the class with the highest probability.

Example – Using posterior probabilities

|  |  |  |  |
| --- | --- | --- | --- |
| Tuple | Gender | Height | Class |
| *t1* | *f* | 1.6 | *short* |
| *t2* | *m* | 2.0 | *tall* |
| *t3* | *f* | 1.9 | *medium* |
| *t4* | *f* | 1.8 | *medium* |
| *t5* | *f* | 1.7 | *short* |
| *t6* | *m* | 1.85 | *medium* |
| *t7* | *f* | 1.6 | *short* |
| *t8* | *f* | 1.8 | *tall* |
| *t9* | *m* | 1.7 | *short* |
| *t10* | *m* | 2.2 | *tall* |
| *t11* | *m* | 2.1 | *tall* |
| *t12* | *f* | 1.8 | *medium* |
| *t13* | *m* | 1.95 | *medium* |
| *t14* | *f* | 1.9 | *medium* |
| *t15* | *f* | 1.8 | *medium* |
| *t16* | *f* | 1.75 | *medium* |
| *t17* | *m* | 1.8 | *tall* |

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

,

,

and

.

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.