Naive Bayes

1 Bayes Theorem

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)},$$

where:

- $P(A|B)$ is the conditional probability of event A given B,
- $P(B|A)$ is the conditional probability of event B given A,
- $P(A)$ and $P(B)$ are *prior* probabilities of events A and B.

2 Naive Bayes Classifer

The Bayes theorem can be used for classification by calculating the probability of each class for a given input value:

$$P(c_y|x) = \frac{P(c_y)P(x|c_y)}{\sum_{c_i \in C} P(c_i)P(x|c_i)}.$$

Classification of multi-dimensional data requires modelling the joint probability of multiple input variables.

$$P(c_y|x_1, \ldots, x_d) = \frac{P(c_y)P(x_1, \ldots, x_d|c_y)}{\sum_{c_i \in C} P(c_i)P(x_1, \ldots, x_d|c_i)}.$$

Assuming the **mutual independence of attributes** (this is a naive assumption, hence the name of the classifier), the joint conditional probability is:

$$P(x_1, \ldots, x_d|c_y) = \prod_{i=1}^{d} P(x_i|c_y).$$

To select the class with the highest probability it is enough to compare the numerator in the equation (the denominator will be equal for all classes):

$$P(c_y|x_1, \ldots, x_d) \sim P(c_y) \prod_{i=1}^{d} P(x_i|c_y).$$

2.1 Smoothing

It is possible that a given value of an attribute may not appear in any example for one of the classes:

$$P(x_i|c_y) = \frac{x_i}{N} = 0.$$

This would set the probability of that class to 0. We can remedy this by using smoothing:

$$P(x_i|c_y) = \frac{x_i + 1}{N + d},$$

where d is the number of possible values of the attribute.

Questions

Question 1.

Using the training set in `Playgolf.xlsx`, classify the following examples using the Naive Bayes classifier:

outlook	**temp**	**humidity**	**windy**
sunny	cool	high	true
overcast	mild	normal	false
overcast	cool	high	false

Mini-project: Naive Bayes

The aim is to classify mushrooms from the dataset in `agaricus-lepiota.data` (source) as either poisonous (class `p`) or edible (class `e`) using the Naive Bayes classifier.

Implement the classifier and test it on the test set in `agaricus-lepiota.test.data`. The decision attribute is in the **first** column.

Use smoothing where necessary.

The program should output the accuracy, precision, recall and F-measure.