Data Mining

TP1 – Quentin Rivollat

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

By the Bayes formula :

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)}$$

Using the Naïve hypothesis, we eventually want to calculate :

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

To do so using Python, we implement the NB algorithm in 2 steps :

1. Learning-NB (with a training set)
2. Classify-NB (with a testing set)

In this TP, we have two different dataset : Titanic data, which is a discrete dataset, and Iris data, which is continuous.

For each, we will divide in similar steps :

- Obtain the data (-> *obtainData* function)
- Divide the dataset, between a training set (2/3) and a testing set (1/3) (-> *splitDataset* function)
- Distinguish, separate the class of the datasets (-> *classSeparation* function)
- Get the number of each element for each attribute, knowing the class (= result) (-> *calculateAttribut* function)
- Calculate every probability (-> *getProba)*
- Use the testing set to test our model (-> *testNB* function)

In iris data, for one situation, we discretize the data (after obtaining them), and then we follow the previous steps. For the other situation, to calculate the probability, we assume that each attribute of iris data follows normal distribution.

The algorithms of irisDiscretize and Titanic are same except for the path of the data (titanic.csv or iris.csv), and in irisDiscretize, we just use one more function that discretize the data.

Moreover, to have better test, we divide the dataset with randomness, which means it's never (or almost) the same testing set.

Results :

After numerous test, there are the ratio of correct answers, out of the number of tests :

- Titanic : around 78% success. It is always between 76% and 80%, which is a low 'variance' of result
- Iris Discretized : around 92% success. It depends on the parameter for discretize the data. It can go from 50 % if success, with bad discretization, until 99%, with better discretization.
- Iris with Normal law : around 96% success. It is usually between 92% and sometimes 100%

This means that the misclassification percentage is : 22% for Titanic, 8% for iris discretized, and 4% for iris with normal law.

With these results, we can say that in general, the Naïve Bayes is a good way to classify data, with a good accuracy.

And if the data is continuous, the discretization method can give us very good results, and even better if we consider that data follow a normal law, almost until perfection.

With all of this, we can say that Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods, and are easy to implement.