TP 1 Data Mining: Naive Bayes (OBLIGATORY)

Tuesday 26th February, 2019
To be send to Frantzeska.Lavda@etu.unige.ch
deadline: Monday 11th March, 2019, 23:59

This TP is obligatory. In this TP you are going to implement the Naive Bayes (NB) algorithm for categorical (titanic) and continuous (iris) data using Python 3. First you will implement it for categorical data and then for continuous data. In the continuous case you are going to implement NB algorithm using three different strategies.

Theory Reminder on Naive Bayes

Remember that Naive Bayes works on computing the *Maximum A Posteriori Probability*, based on the assumption that the attributes of the training examples are independent.

$$\begin{aligned}
C_{MAP} &= argmax_{c_i \in C} P(c_i|X) && (1)\\
&= argmax_{c_i \in C} \frac{P(X|c_i)P(c_i)}{P(X)} && (2)\\
&= argmax_{c_i \in C} \frac{P(X|c_i)P(c_i)}{\sum_{c_j \in C} P(X|c_j)P(c_j)} && (3)\\
&\propto argmax_{c_i \in C} P(X|c_i)P(c_i) && (4)\\
&= argmax_{c_i \in C} P(X_1, X_2, X_3, ..., X_d|c_i)P(c_i) && (5)
\end{aligned}$$

Under the assumption of independent attributes we have :

$$\begin{aligned}
C_{MAP} &= argmax_{c_i \in C} P(X_1, X_2, X_3, ..., X_d|c_i)P(c_i) && (6)\\
&= argmax_{c_i \in C} P(X_1|c_i)P(X_2|c_i)...P(X_d|c_i)P(c_i) && (7)\\
&= argmax_{c_i \in C} P(c_i)\Pi_{j=1}^{d} P(X_j|c_i) && (8)
\end{aligned}$$

How to classify a new instance

In order to classify a new instance we will be using the formula 7. For example consider the following instance :

$$< a_1, a_2, a_3, ..., a_d, ? >$$

Formula 7 will be then written as

$$? = C_{MAP} = argmax_{c_i \in C} P(X_1 = a_1|c_i)P(X_2 = a_2|c_i)...P(X_d = a_d|c_i)P(c_i)$$

In practice, to find the class label that maximizes the above quantity we will have to compute:

$$P(X_1 = a_1|c_i)P(X_2 = a_2|c_i)...P(X_d = a_d|c_i)P(c_i)$$

for every class i, and select the one that gives the highest probability.

Computing the necessary probabilities to perform classification
Before we are able to classify a new instance we have to compute the following:

1. The probability $P(c_i)$ for every class

2. For each attribute $X_j, 1 \leq j \leq n$, with k distinct values : $a_{j1}, a_{j2}, ..., a_{jn}$, compute for every distinct value a_{jz}, $1 \leq z \leq k$ the conditional probability:
$$P(X_j = a_{jz}|c_i)$$

Exercises

Split your dataset in training and testing data sets (2/3 for training 1/3 for testing).

1. Implement Naive Bayes for categorical data, apply in titanic dataset and report your misclassification rate.

2. Implement the Naive Bayes for continuous data (apply it in iris dataset) using the following strategies:

   (a) Assume that each attribute of iris data follows normal distribution.

   (b) Discretize the data and use the same techniques as for categorical data (ex1).

   (c) (To be done in the next tp) Estimate the probability distribution from the data.

For all the different cases draw the decision surfaces induced by naive Bayes. For the visualization purposes choose two attributes as predictive attributes and color the plane defined by these two attributes on the basis of class labels that Naive Bayes predicts, compute the percentage of misclassified instances on the test set for the different strategies.

Hint: The visualization can be trained on the complete data set however, misclassification cannot.

General instruction

You have to send me your **code** and a **report**, both saved using the format: TP1_LASTNAME_Firstname.

The code should be well written with detailed comments to explain what you do at each step. Try to avoid the for loops using the nymPy library. Your code should be generic and use functions. For example the code for the *titanic data* should be applicable, to the *iris data* after the discretization (ex2b).

The report should be sent in **.pdf** format. Explain the problem and discuss the results, it should be a summary of the main findings, factual, clear, and concise. Include in your report the visualization on iris and your comments on the decision surfaces (what form do they have? linear? quadratic? other? what is the effect of the different strategies?). It should also contain the misclassification percentage for the different strategies and a discussion of performance.