

**Department of Computer Science and Engineering(UG Studies)**

**PES University, Bangalore-560085**

|  |  |
| --- | --- |
| **Session :** Aug - Dec 2017  **Credits :** 0-0-2-0-1 | UE14CS405 : Machine Learning Lab |
| **Lab # :** 07 | Implement Naïve Bayes Spam/Sentiment classifier. |

|  |
| --- |
| DataSet:  spam\_reviews :- It is the folder containing two sub folders(classes) spam and ham with sample data. |
| **Theory:** What is Naive Bayes algorithm? It is a classification technique based on [Bayes’ Theorem](https://en.wikipedia.org/wiki/Bayes'_theorem) with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.  Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.  Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:  https://www.analyticsvidhya.com/wp-content/uploads/2015/09/Bayes_rule-300x172.png How Naive Bayes algorithm works? Let’s understand it using an example. Below I have a training data set of weather and corresponding target variable ‘Play’ (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather condition. Let’s follow the below steps to perform it.  Step 1: Convert the data set into a frequency table  Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.  Bayes_4  Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.  **Problem:** Players will play if weather is sunny. Is this statement is correct?  We can solve it using above discussed method of posterior probability.  P(Yes | Sunny) = P( Sunny | Yes) \* P(Yes) / P (Sunny)  Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P( Yes)= 9/14 = 0.64  Now, P (Yes | Sunny) = 0.33 \* 0.64 / 0.36 = 0.60, which has higher probability. |
| **Executable Code Given:**  Naïve Bayes for spam prediction. File Name: naïve-bayes-spam.py |
| **To Do:**  1. With the help of given code for spam prediction Implement Naïve Bayes for Sentiment classification. (Create or Use Data Set movie reviews with pos and neg reviews)  2. Implement the algorithm by using the datasets having more number of classes. |
| **Learning outcome:**   * It is easy and fast to predict class of test data set. It also perform well in multi class prediction * When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. * It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption). |