**Leaflets three, let it be?**

**1. Introduction and Overview**

Mushroom edibility is determined by lots of different attributes. Conducting a poison test every time before eating is not realistic. Therefore, a method, which is able to judge the edibility by looking at its physical properties like color, shape, habitat etc., is really essential. In this study, we used classification to develop a rule for differentiating edible and poisonous mushrooms. The dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. It contains 8124 instances, with 2,480 missing values in the “Stalk Root” attribute. After further investigation into these missing values, we found that they were not missing completely at random. Therefore, we decided to treat the missing values as another category of this attribute.

**2. Machine Learning Method**

Due to the categorical nature of dataset, the number of methods available is limited. Eventually, we chose the naive Bayes classifier as the one to work with. The reason we chose Naïve Bayes, which is also its advantage, is that it only requires a small amount of training data to come up with an optimized classifier function. The method imposes a strong (naive) assumption that the value of any particular attribute is unrelated to the value of all other attributes, given class label. This assumption allows for the simplification of Bayes’ Theorem:

The Bayes’ Theorem is then modified to handle input that is not continuous but multinomial by changing the equation and expressing it in log-space to give a linear-classifier:

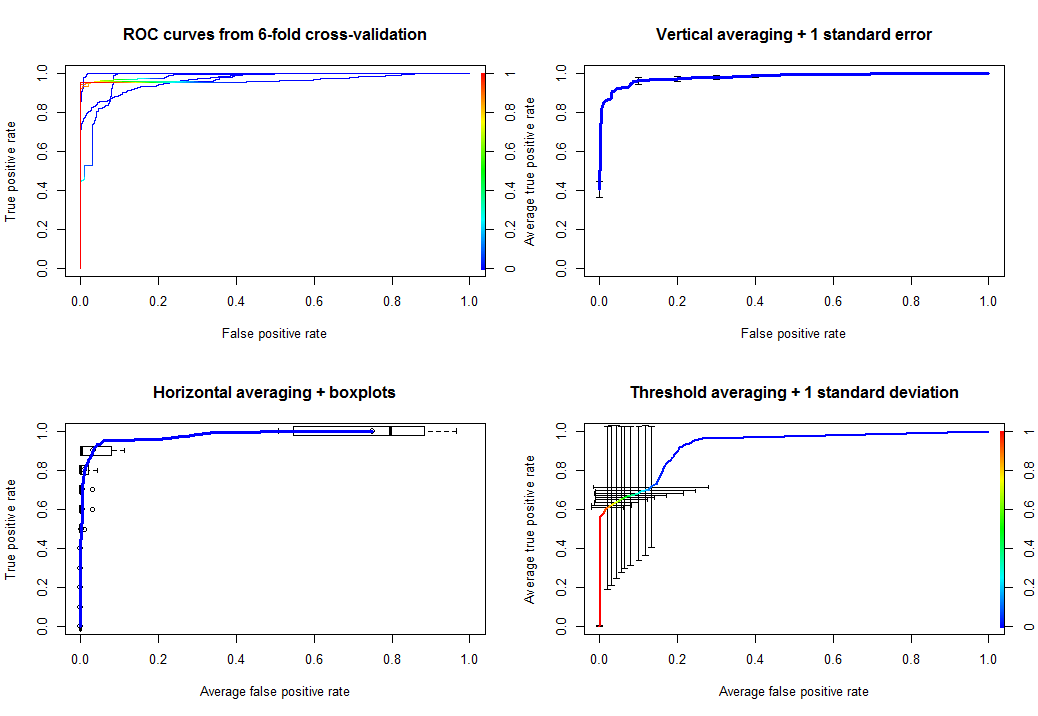
This would lead to probabilities of 0’s for a given class and attribute vector that don’t occur together in the training data set. This is problematic because **it will wipe out all information in the other probabilities when they are multiplied*.*** Therefore, it is desirable to incorporate a small-sample correction, called pseudo counts, in all probability estimation such that no probability is ever set to be exactly zero.

The algorithm uses the training data set to choose the most probable outcome, known as the “Maximum a Posteriori” decision rule. Then on the test data set, the model assigns a probability to each case being poisonous and a threshold is used to determine if the answer should be edible or poisonous. To assess the performance of this classifier in practice, we used the ***K-*fold cross validation**: we split the data set into six equally-sized subsets and applied Naïve Bayes method six times, each time leaving out a different subset to be tested while using the other five as the training set. We used the Receiver Operating Characteristics (ROC) curve to visualize the classifier's performance. And since we are dealing with a binary classification problem, misclassification rates, obtained after we fix a threshold, are also provided and averaged for overall performance.

**3. Summary of Findings**

**3.1 ROC Curve**

ROC (Receiver Operating Characteristic) curve is a graphical plot, which illustrates the performance of a binary classifier system as its discrimination threshold is varied.



3.2 Misclassification Rate

Threshold: 0.5

|  |  |
| --- | --- |
| Run | MCR |
| 1 | 0.080 |
| 2 | 0.080 |
| 3 | 0.121 |
| 4 | 0.042 |
| 5 | 0.051 |
| 6 | 0.097 |
| average | 0.079 |

**4. Discussion**

**The main assumption of naive Bayes classifier is independence between features. Despite the facts that this assumption is often inaccurate, its properties still make it efficient and competitive among other classification methods. For example, it helps alleviate problems stemming from the curse of dimensionality. Although Naïve Bayes often fails to produce a good estimate for the correct class probabilities, this may not be a requirement for many cases. For instance, Naïve Bayes classifier will make the correct decision rule classification so long as the correct class is more probably than any other class; this is true regardless of whether the probability is slightly inaccurate. Therefore, it is robust enough to ignore some deficiencies in its underlying probability model.**