**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. A classification model is conducted in this report to develop a way of differentiating edible and poisonous mushroom. 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 use the missing values in “Stalk Root” as another category. Since the data is categorical, a context-based similarity measure that learns distances between the values of categorical attributes was employed.

**2. Machine Learning Method**

Due to the categorical structure of dataset, the methods for machine learning were very limited. The method that we identified with the ability to cluster our data was Naïve Bayes Classifier. 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. “Naïve” indicates a very strong assumption, the value of any particular attribute is unrelated to the value of all other attributes. 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:

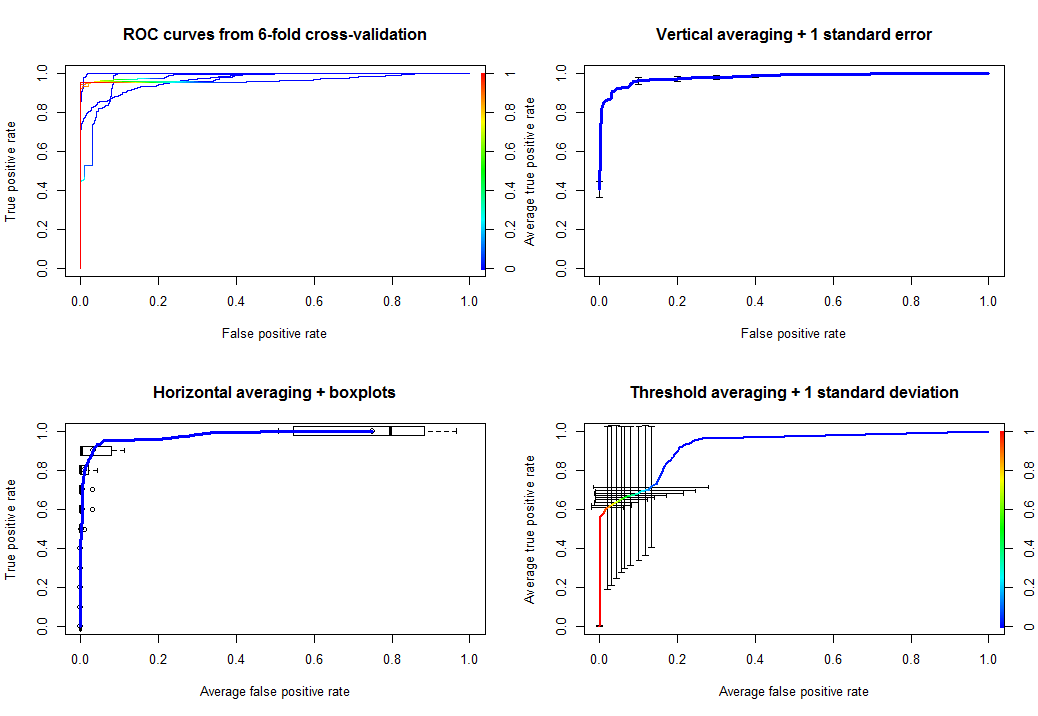
This would lead to probabilities of 0’s for a given class and attribute vector that don’t occur together in the training data set. This is problematic because **it will wipe out all information in the other probabilities when they are multiplied*.*** Therefore, it is desirable to incorporate a small-sample 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 the outcome being edible and a threshold is used to determine if the answer should be edible or poisonous. The threshold is “the magnitude or intensity that must be exceeded for a certain reaction, phenomenon, result, or condition to occur or be manifested” (Oxford Dictionary).

The cross validation technique that was used is the ***K* folds leave-one out cross validation**. In this technique, we split the data set into six subsets. We then performed the Naïve Bayes method six times each time leaving out a different one of those six subsets to be tested and the other five to be the training data set. An analysis of these six models is provided with their resulting ROC curve. Misclassification rate are also provided and averaged for overall performance.

**3. Summary of Findings**

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



**4. Discussion**

**The main assumption of this model is the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Despite the facts that the independence assumptions are often inaccurate, but 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.**