SVM Classification of Mushroom Edibility

Ronald Estevez and Reilly Fitzgerald

Define Problem:

Mushrooms can be a tasty treat, but if not picked responsibly some are poisonous to consume. According to UCI "there is no *simple* rule for determining ediblity of [mushrooms]." We believe we can leverage machine learning to combine many different attributes of mushrooms to decide whether or not a mushroom is safe to eat. The difficulty to determine the safety of mushrooms means it is important to find classifiers which work to determine the safety of wild mushrooms. Our goal is to implement a support-vector machine that can be trained to correctly predict whether a certain mushroom is edible or not. Code for this project can be found at the following source: https://bit.ly/3aq228i.

Show connection to class:

In class, support vector machines (SVM) were introduced as "one of the most important developments in pattern recognition in the last decades." The goal of an SVM is to produce an optimal hyperplane that delineates distinct classes of data in a population. This makes SVM very well suited for populations where labeling is binary in nature. For this reason, we decided that an SVM could be a very effective algorithm for our project where we seek to label mushroom samples as edible or poisonous. Although feature selection is not the focus of this project, we decided to also use and observe the effectiveness of a greedy forward feature selection algorithm. The dataset under examination contains 22 distinct attributes describing each mushroom. The kernel trick allows SVM to easily handle many attributes and still make accurate predictions; however, for the sake of efficiency we are using the greedy forward feature selection algorithm in hopes that this will allow us to run more trials on our data by reducing the dimensions of it.

Describe Data:

The mushrooms in this dataset consist of 23 species, all within the *Agarlicus* and *Lepiota* families of mushrooms. It belongs to the UCI Machine Learning repository but was obtained from the following source for our project: https://www.kaggle.com/uciml/mushroom-classification/. The attribute we are trying to identify will be edibility, which is either definitely edible, definitely poisonous, or unknown (which is recorded as poisonous in the dataset). The data consists of 8125 individual mushrooms. Each is defined by 22 attributes and each attribute is represented by a single character variable. See either of the links attached above for more details on the range of these attributes and a description of each.

Simple preliminary analysis of the data revealed that two of the 22 features were arbitrary or contained null values that would cause issues for our SVM classifier. The feature with null values, 'stalk-root', was ignored by our classifiers and removed from the dataset. The arbitrary feature, ‘veil-type’, had the same value for every single sample in the dataset so it too was ignored by our classifiers and removed from the dataset.

For each classifier we ran, we chose to randomly split the data such that 80% of the samples were used for training and 20% of the samples were used for testing. This proportion was chosen due to its widespread use in machine learning, the decently large size of our dataset which allows for a smaller proportion of samples to be used for testing, and because it performed similarly or better than other data splits. As is the case with most machine learning algorithms, SVM requires that the data under examination be of numerical form. We chose to encode the features in our project (which consists entirely of single, lowercase characters) by converting each value to its unicode value and subtracting the unicode value of the ‘a’ from it. The labels of the data were encoded as well for the sake of uniformity and the poisonous label was designated the number 0 and the edible label the number 1.

Show Results:

Using All Features

Using all 22 features, we tuned the SVM classifier to use a variety of kernels and ran it once on each. We found a linear kernel provided an accuracy of 95.82%. A quadratic polynomial kernel resulted in 99.75% accuracy for the tests. A cubic polynomial kernel and a radial basis kernel (RBF) both yielded 100% accuracy in classification. With a Sigmoid kernel, we only achieved an accuracy of about 55%.

Greedy Feature Selection

We wanted to do further testing with tweaked parameters, and decided that in order to conduct more tests rapidly, it was necessary to narrow down the number of features. We chose to use a greedy forward feature selection algorithm. Although this algorithm is suboptimal for selecting the best attributes, the SVM was still able to give 88% accuracy with a linear kernel, with the use of only 2 to 5 features (since the greedy algorithm returns a different amount of features each time).

We tested the attributes from a single run of the greedy algorithm (odor, spore print color, habitat, bruises, stalk-surface above ring, and gill-spacing) and ran it through multiple SVMs, 10 times each so we could determine a Std. Deviation and Mean Accuracy.

Kernel	Mean Accuracy	Std. Deviation
Linear SVM	0.883	0.00732
Quadratic Polynomial	0.963	0.00665
Cubic Polynomial	0.992	0.00165
RBF (Run 100 times)	0.996	0.00351
Sigmoid	0.534	0.136

It is interesting to note that the RBF kernel seems to perform best for the given dataset. In addition, it is also the fastest kernel and this allowed us to run 100 trials with it as well as use it for the kernel for our greedy forward selection algorithm.

Sigmoid Kernel

As seen above the Sigmoid kernel's performance was underwhelming, with an almost 50/50 accuracy/error rating. In an attempt to improve performance, the features were normalized. Yet even then the performance does not improve, and in fact stays around 50 even when all 22 attributes are used. Normalization likely did very little as the data was mostly qualitative with no range in between (red or gray cap color for instance). This suggests that the sigmoid kernel is unsuitable for our data, at least without further data modification.

Conclusions

Although it is impossible to know for sure what kernel best performs for a certain dataset, simply running several trials on a dataset can yield potent results. For our dataset, the RBF kernel performed the best out of the several common kernels we tuned our classifier to use (which were linear, quadratic, cubic, RBF, and sigmoid kernels). It also ran the quickest of the kernels examined. For this specific dataset, normalization was not required. Results on the dataset without normalization were very accurate and the worse performing kernel, the sigmoid kernel, did not benefit much from normalization. However, dimensional reduction does benefit our data. As mentioned above, the greedy forward feature selection algorithm narrowed the scope of dimensions required from 2 to 5. If our research were to be extended to focus more on feature selection than learning, it would be interesting to take note of which features exactly perform best for this dataset as clearly all 22 are not needed. Overall, our learner performed exceedingly well and proved that SVM is a useful technique for predicting edibility of mushrooms of the *Agarlicus* and *Lepiota* families.