Introduction/Background

In this lab I used a perceptron and logistic regression algorithm to analyze a mushroom data set. The data set contained over 8000 mushrooms, for which we knew whether or not it was poisonous, as well as about 100 of its other features. We split the data into a training set, a validation set, and a test set, and used the training set and validation set to train the algorithms, and used the test set to test how accurate the models were. The purpose of this was to see if there was a strong enough correlation between the features of a mushroom and its toxicity, that the models could correctly predict if a mushroom was poisonous just based on its features.

Results for Perceptron

The perceptron model converged instantaneously in only two iterations and was able to correctly predict the toxicity of every mushroom in the training set, validation set, as well as in the testing set. Hence, the precision and recall of the model was 100%. The perceptron found that having a creosote odor, followed by having a foul odor, followed by having a green spore print, were the three most indicative features of a poisonous mushroom. On the other hand, having an almond odor, followed by having no odor, followed by having a broad gill size, were the three most indicative features of an edible mushroom.

Results for Logistic Regression

The accuracy of the logistic regression model was a bit more complicated to measure than that of the perceptron, because the accuracy of the logistic regression model varied based on its learning rate. I found that with a 0.0001 learning rate, it was wrong about 4% of the time and took about 20 iterations (two minutes) to converge, with a 0.001 learning rate, it was wrong about 1-2% of the time and took about 7 iterations (40 seconds) to converge, but with a learning rate between 0.25 and 2, it converged in 1 to 3 iterations (15 seconds), and often predicted the toxicity of mushrooms without making any mistakes. When it did make mistakes, there were fewer than 10 of them across all data sets. Most mistakes that occurred were in the training set. This was caused by the fact that the training set was 8 times bigger than the validation and testing sets. The test set was almost always correctly sorted, so its precision and recall were always above 99%, and usually 100%. I did not see any signs of the model overfitting for the training or validation sets. The logistic regression model found that having bruises, followed by having a creosote odor, followed by having a foul odor, were the three most indicative features of a poisonous mushroom. On the other hand, having an anise odor, followed by having an almond odor, followed by having a broad gill size, were the three most indicative features of an edible mushroom.

Analysis

A good model for classifying mushrooms is first and foremost a model that accurately detects if a mushroom is poisonous or not. Given the potential outcomes of mistakenly labeling a poisonous mushroom as edible, it is more important to have a high recall than to have a high precision. Secondarily, between two equally accurate models, the efficiency should be taken into consideration. Though correct outcomes are the primary objective, machine learning models are notorious for being slow, and in some cases it can even hinder our ability to use an algorithm to analyze a data set.

Both the perceptron and logistic regression models performed really well at classifying the data, however the perceptron was slightly more accurate and significantly faster. It also had the benefit of always being accurate, whereas the accuracy of the logistic regression model was dependent on the imputed learning rate. They both had similar weights, meaning that there was significant overlap between what features they deemed important. Specifically, a mushroom's odor was most important is determining if a mushroom was poisonous or edible. Both models also found that mushrooms with a broad gill were highly likely to be edible. For different learning rates, the logistic regression model found other features to be of importance. Perhaps, the odor is objectively the most distinguishing feature, so when the linear regression model was tuned right, it picked up on it.

Overall, both models were very accurate. If either one of them were used to predict the toxicity of mushrooms, it would probably be other obstacles, such as noise in the data, mushrooms within one species having slightly different features, or faulty mushroom feature documentation, that would contribute more errors than the models.