Mushroom Classification

DS 6003 Spark Homework

Runhao Zhao

rz6dg

Motivation

As it is recorded, there are more than 10000 described mushroom species exist worldwide. Many types of mushrooms can be identified as poisonous mushrooms by their shapes, sizes or colors. I am interested in applying various machine learning methods to predict if a mushroom is safe to eat based on multiple observed features. Learning which characteristics spell death and which are most palatable is also important and helpful in identifying edible/poisonous mushrooms. Ultimately, I hope this could be used to help mushroom lovers avoid poisonous mushrooms when spending time outdoors.

Code snippet

```
#create a logistic regression model
from pyspark.ml.classification import LogisticRegression
lr = LogisticRegression(featuresCol = 'features', labelCol = 'label', maxIter=10)
lrModel = lr.fit(train)
trainingSummary = lrModel.summary
roc = trainingSummary.roc.toPandas()
plt.plot(roc['FPR'],roc['TPR'])
plt.ylabel('False Positive Rate')
plt.xlabel('True Positive Rate')
plt.title('ROC Curve')
plt.show()
```

The codes above are building a logistic regression model on the data set to predict the edibility of a mushroom. The features used in the model include all 28 features in the data set. An ROC curve is created to evaluate the model performance

```
for col_name in df1.columns:
    indexer = StringIndexer(inputCol=col_name, outputCol="new_"+col_name)
    model = indexer.fit(df1)
    df1 = model.transform(df1)
```

The codes above are looping through each column and applying label encoding to transform non-numerical labels to numerical labels.

Visualizations