Report

1. ## Load the data into Spark

   DO: Explain what the provided code does.
   The code reads the raw data line by line, separates the numbers by ",", gives every line a label according to its last number, and changes every number into float.

   DO: What does data looks like?
   In each line, we have a vector with 10 elements as features and an integer as label.

   DO: What is the schema of the data?
   |-- features: vector (nullable = true)
   |-- label: integer (nullable = true)

   DO: In our dataset, how many tumors are benign? malign?
   458 are benign, 241 are malignant.

2. ## Splitting into training and testing

   ```
   splits = data.randomSplit([9.0,1.0])
   train = splits[0]
   test = splits[1]
   ```

3. ## Building the model

   ```
   from pyspark.ml.classification import DecisionTreeClassifier
   dtc = DecisionTreeClassifier()
   bc_model = dtc.fit(train)
   ```

4. ## Testing your model

   DO: What does the DataFrame prediction contain?
   It contains rawPrediction (the number of benign tumors and malign tumors), probability (the probability of being benign), and prediction.

   DO: What is the area under ROC of our classifier?
   0.8950742240215924

   DO: What is the accuracy of our classifier?

0.927710843373494

5. Improving the model

DO: What is the area under ROC of your model now?
train-validation: 0.9736842105263158
cross-validation: 0.9439946018893388

6. Improve the classification

I chose the Logistic Regression.
Area under ROC: 0.9423076923076923
Accuracy: 0.963855421686747