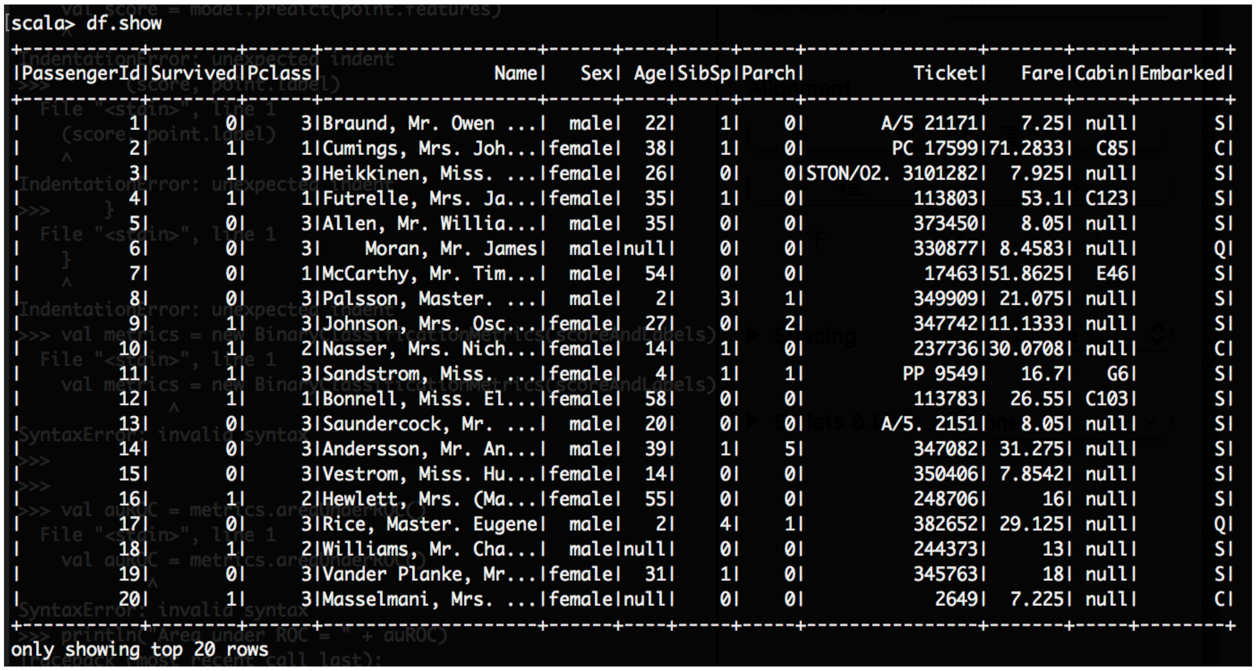
Project Report
Runchen Hu rh2619

Code: (Run in Spark Shell)

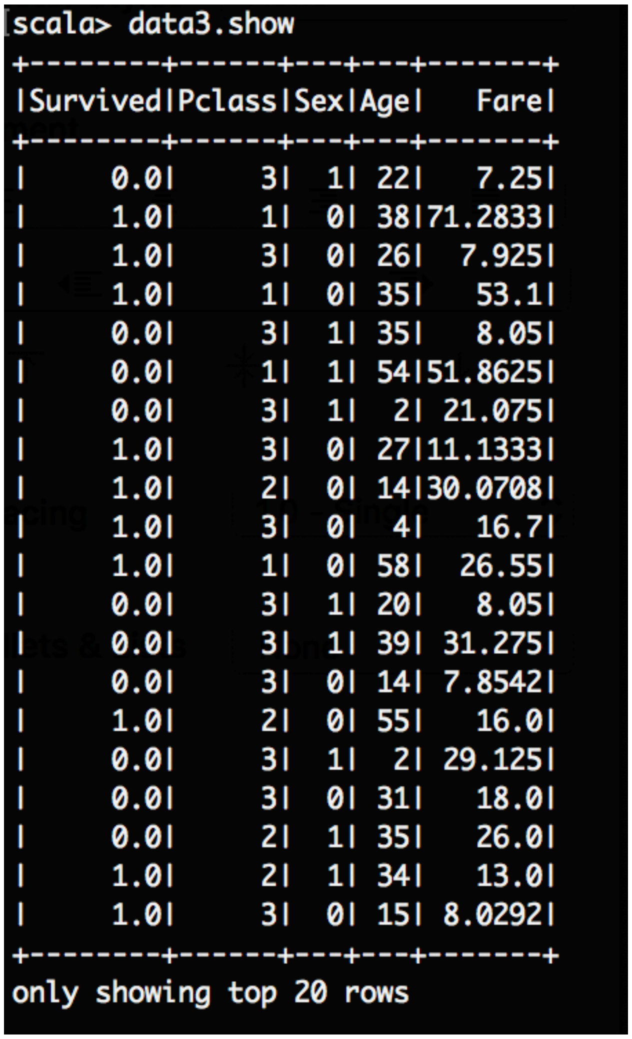
```
// SVM
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.classification.SVMWithSGD
import org.apache.spark.mllib.evaluation.BinaryClassificationMetrics

// First I standardized the dataframe and drop the rows with null values.
val spark =
org.apache.spark.sql.SparkSession.builder.master("local").appName("Reader").getOrCreate;
val df = spark.read.format("csv").option("header","true").load("titanic-train.csv")
val data = df.select($"Survived",$"Pclass",$"Sex",$"Age",$"Fare")
val newdata = data.withColumn("Sex", when(col("Sex") === "male",
"1").otherwise(col("Sex")))
val finaldata = newdata.withColumn("Sex", when(col("Sex") === "female",
"0").otherwise(col("Sex")))
val data2 = finaldata.selectExpr("cast(Survived as double) Survived", "cast(Pclass as int)
Pclass", "cast(Sex as int) Sex", "cast(Age as int) Age", "cast(Fare as float) Fare")
val data3 = data2.na.drop

// Then I split the dataset into training set and testing set.
val splits = data3.randomSplit(Array(0.8, 0.2), seed = 11L)
val training = splits(0).cache()
val test = splits(1)

// I converted dataframe into RDD and set feature columns and label column.
val training_rows: org.apache.spark.rdd.RDD[org.apache.spark.sql.Row] = training.rdd
val labeled_training = training_rows.map( a =>
LabeledPoint(a.getDouble(0),org.apache.spark.mllib.linalg.Vectors.dense(a.getInt(1),a.getInt(2),
a.getInt(3),a.getFloat(4))))
```

```
val test_rows: org.apache.spark.rdd.RDD[org.apache.spark.sql.Row] = test.rdd
val labeled_test = test_rows.map( a =>
LabeledPoint(a.getDouble(0),org.apache.spark.mllib.linalg.Vectors.dense(a.getInt(1),a.getInt(2),
a.getInt(3),a.getFloat(4))))

// I input labeled training set to build the SVM model.
val numIterations = 100
val model = SVMWithSGD.train(labeled_training, numIterations)

// Then I set BinaryClassificationMetrics as evaluation metrics.
model.clearThreshold()
val scoreAndLabels = labeled_test.map { point =>
      val score = model.predict(point.features)
      (score, point.label)
    }
val metrics = new BinaryClassificationMetrics(scoreAndLabels)

//Printed the ROC results.
val auROC = metrics.areaUnderROC()
println("Area under ROC = " + auROC)
```

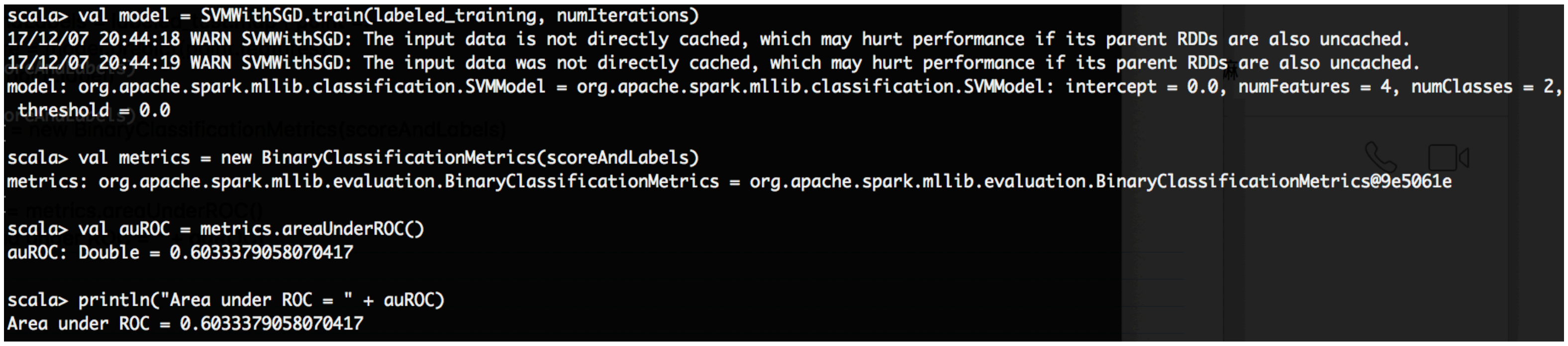
ScreenShots:



```
scala> df.show
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|          1|       0|     3|Braund, Mr. Owen ...|  male|  22|    1|    0|       A/5 21171|   7.25| null|       S|
|          2|       1|     1|Cumings, Mrs. Joh...|female|  38|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|Heikkinen, Miss. ...|female|  26|    0|    0|STON/O2. 3101282|  7.925| null|       S|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|  35|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|Allen, Mr. Willia...|  male|  35|    0|    0|          373450|   8.05| null|       S|
|          6|       0|     3|    Moran, Mr. James|  male|null|    0|    0|          330877| 8.4583| null|       Q|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|  54|    0|    0|           17463|51.8625|  E46|       S|
|          8|       0|     3|Palsson, Master. ...|  male|   2|    3|    1|          349909| 21.075| null|       S|
|          9|       1|     3|Johnson, Mrs. Osc...|female|  27|    0|    2|          347742|11.1333| null|       S|
|         10|       1|     2|Nasser, Mrs. Nich...|female|  14|    1|    0|          237736|30.0708| null|       C|
|         11|       1|     3|Sandstrom, Miss. ...|female|   4|    1|    1|         PP 9549|   16.7|   G6|       S|
|         12|       1|     1|Bonnell, Miss. El...|female|  58|    0|    0|          113783|  26.55| C103|       S|
|         13|       0|     3|Saundercock, Mr. ...|  male|  20|    0|    0|       A/5. 2151|   8.05| null|       S|
|         14|       0|     3|Andersson, Mr. An...|  male|  39|    1|    5|          347082| 31.275| null|       S|
|         15|       0|     3|Vestrom, Miss. Hu...|female|  14|    0|    0|          350406| 7.8542| null|       S|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|  55|    0|    0|          248706|     16| null|       S|
|         17|       0|     3|Rice, Master. Eugene|  male|   2|    4|    1|          382652| 29.125| null|       Q|
|         18|       1|     2|Williams, Mr. Cha...|  male|null|    0|    0|          244373|     13| null|       S|
|         19|       0|     3|Vander Planke, Mr...|female|  31|    1|    0|          345763|     18| null|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|null|    0|    0|            2649|  7.225| null|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 20 rows
```



```
scala> data3.show
+--------+------+---+---+-------+
|Survived|Pclass|Sex|Age|   Fare|
+--------+------+---+---+-------+
|     0.0|     3|  1| 22|   7.25|
|     1.0|     1|  0| 38|71.2833|
|     1.0|     3|  0| 26|  7.925|
|     1.0|     1|  0| 35|   53.1|
|     0.0|     3|  1| 35|   8.05|
|     0.0|     1|  1| 54|51.8625|
|     0.0|     3|  1|  2| 21.075|
|     1.0|     3|  0| 27|11.1333|
|     1.0|     2|  0| 14|30.0708|
|     1.0|     3|  0|  4|   16.7|
|     1.0|     1|  0| 58|  26.55|
|     0.0|     3|  1| 20|   8.05|
|     0.0|     3|  1| 39| 31.275|
|     0.0|     3|  0| 14| 7.8542|
|     1.0|     2|  0| 55|   16.0|
|     0.0|     3|  1|  2| 29.125|
|     0.0|     3|  0| 31|   18.0|
|     0.0|     2|  1| 35|   26.0|
|     1.0|     2|  1| 34|   13.0|
|     1.0|     3|  0| 15| 8.0292|
+--------+------+---+---+-------+
only showing top 20 rows
```

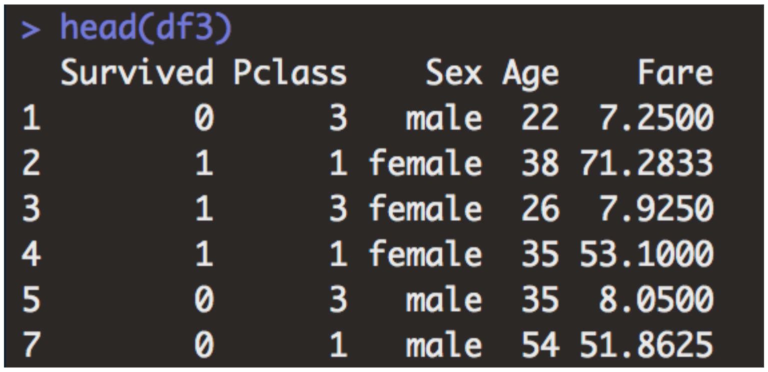


```
scala> val model = SVMWithSGD.train(labeled_training, numIterations)
17/12/07 20:44:18 WARN SVMWithSGD: The input data is not directly cached, which may hurt performance if its parent RDDs are also uncached.
17/12/07 20:44:19 WARN SVMWithSGD: The input data was not directly cached, which may hurt performance if its parent RDDs are also uncached.
model: org.apache.spark.mllib.classification.SVMModel = org.apache.spark.mllib.classification.SVMModel: intercept = 0.0, numFeatures = 4, numClasses = 2,
 threshold = 0.0

scala> val metrics = new BinaryClassificationMetrics(scoreAndLabels)
metrics: org.apache.spark.mllib.evaluation.BinaryClassificationMetrics = org.apache.spark.mllib.evaluation.BinaryClassificationMetrics@9e5061e

scala> val auROC = metrics.areaUnderROC()
auROC: Double = 0.6033379058070417

scala> println("Area under ROC = " + auROC)
Area under ROC = 0.6033379058070417
```

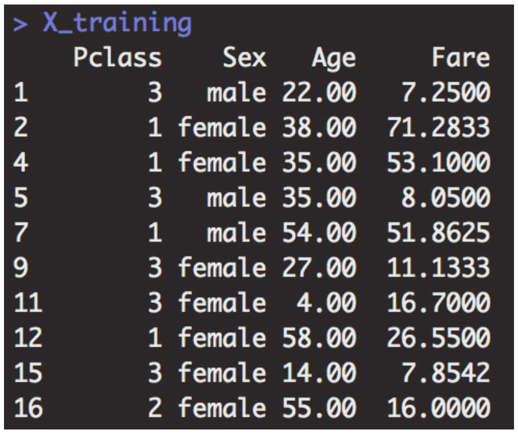
Code: (Run in R)

```
library(e1071)
library(ROCR)

setwd(‘~/Downloads/SVM') # You can set your own work directory

# First I standardized the original dataframe.
df = read.table('titanic-train.csv', header = TRUE, sep = ',')
head(df)
df2 = df[ , -which(names(df) %in%
c("PassengerId","Name","SibSp","Ticket","Cabin","Embarked","Parch"))]
head(df2)
df3 = df2[complete.cases(df2),]
head(df3)

#Then I split df3 into two sets
set.seed(43)
tst_idx = sample(714, 200, replace = FALSE)
test = df3[tst_idx,]
training = df3[-tst_idx,]

X_training = training[, -which(names(training) %in% c("Survived"))]
y_training = training[,c(“Survived")]
X_test = test[, -which(names(test) %in% c("Survived"))]
y_test = test[,c("Survived")]

#Then I did the AUC test.
aucs = c()
plot(x = NA, y = NA, xlim = c(0, 1), ylim = c(0, 1), ylab = 'True Positive Rate', xlab = 'False
Positive Rate', bty = 'n')

lvls = c(0, 1)
for (type.id in 1:2) {
  type = as.factor(training$Survived == lvls[type.id])
  model = model <- svm(as.formula(paste("type", "~ .", sep = " ")), data = training[,-1], kernel
= "linear", probability = T, type = "C-classification",cost = 1)

  Survived = predict(model, test[,-1], decision.values = T, probability = T)
  score = attr(Survived, "probabilities")[,2]

  actual.Survived = test$Survived == lvls[type.id]
  pred = prediction(score, actual.Survived)
  perf = performance(pred, "tpr", "fpr")
  roc.x = unlist(perf@x.values)
  roc.y = unlist(perf@y.values)
  lines(roc.y ~ roc.x, col = type.id + 1, lwd = 2)
  nauc = performance(pred, "auc")
  nauc = unlist(slot(nauc, "y.values"))
  aucs[type.id] = nauc
}
legend("bottomright", legend = c("1 vs others", "2 vs others"), col = 2:3, lty = 1, lwd = 2)
lines(x = c(0, 1), c(0, 1))
cat("AUC", fill = T)
aucs
```

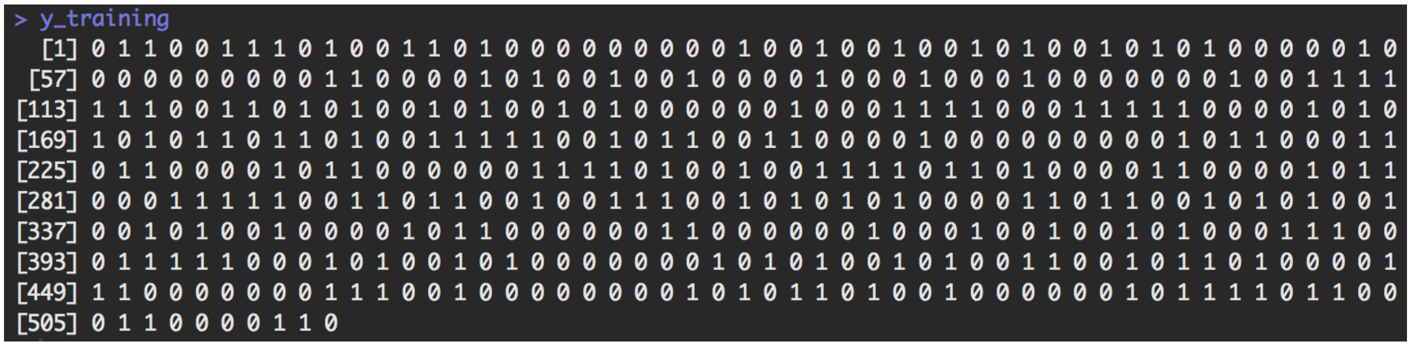
ScreenShots:



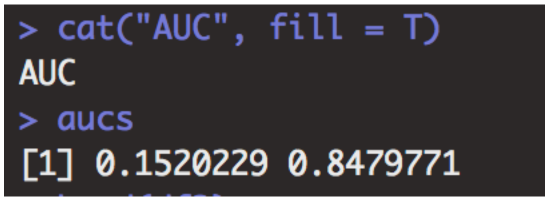
```
> head(df3)
  Survived Pclass    Sex Age    Fare
1        0      3   male  22  7.2500
2        1      1 female  38 71.2833
3        1      3 female  26  7.9250
4        1      1 female  35 53.1000
5        0      3   male  35  8.0500
7        0      1   male  54 51.8625
```



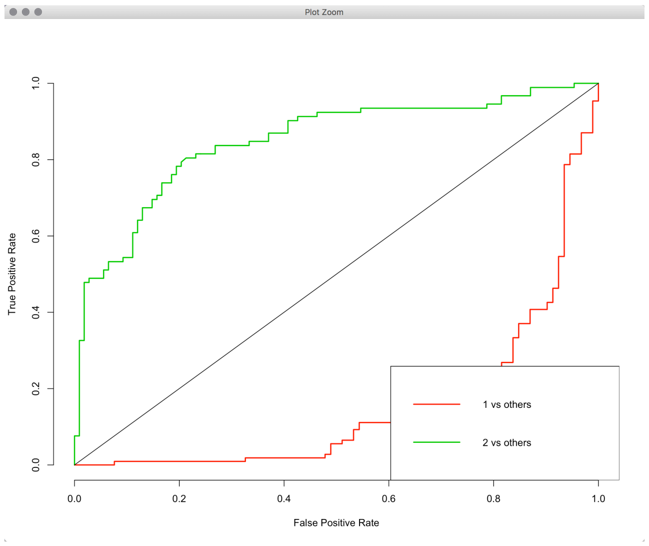
```
> X_training
      Pclass    Sex   Age      Fare
1          3   male 22.00   7.2500
2          1 female 38.00  71.2833
4          1 female 35.00  53.1000
5          3   male 35.00   8.0500
7          1   male 54.00  51.8625
9          3 female 27.00  11.1333
11         3 female  4.00  16.7000
12         1 female 58.00  26.5500
15         3 female 14.00   7.8542
16         2 female 55.00  16.0000
```



```
> y_training
  [1] 0 1 1 0 0 1 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 0
 [57] 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 1 1
[113] 1 1 1 0 0 1 1 0 1 0 1 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 1 1 1 0 0 0 1 1 1 1 1 0 0 0 0 1 0 1 0
[169] 1 0 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 0 0 1 0 1 1 0 0 1 1 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 1
[225] 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 0 0 1 1 1 1 0 1 0 0 1 0 0 0 0 1 0 0 1 1 1 1 0 1 1 0 1 0 0 0 0 1 1 0 0 0 0 1 0 1 1
[281] 0 0 0 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 1 0 0 1 0 1 0 0 1 0 1 0 1 0 1 0 0 0 0 1 1 0 1 1 0 0 1 0 1 0 1 0 0 1
[337] 0 0 1 0 1 0 0 1 0 0 0 0 1 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 1 1 0 0
[393] 0 1 1 1 1 1 0 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 1 0 1 0 0 1 1 0 0 1 0 1 1 0 1 0 0 0 0 1
[449] 1 1 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 1 1 1 0 1 1 0 0
[505] 0 1 1 0 0 0 0 1 1 0
```



```
> cat("AUC", fill = T)
AUC
> aucs
[1] 0.1520229 0.8479771
```



Thank you so much for your effort and help in this semester!

Thank you professor!