Workshop 3

Learning From Big Data

ML Logistic Regression and K-Means Clustering

*Please do the exercises or tasks **without** "**Optional"** mark at first. After that, if you still have some time, please try the tasks with "Optional".

Part 1: Logistic Classification with *Scikit_learn*

In the area of ML, the most common supervised learning tasks are regression (*predicting values*) and classification (*predicting classes*). In this part, we will turn our attention to classification systems. Refer to the lecture of Session 3, you will have a number of tasks to carry out.

Task 1: Binary Classification and Performance Measure

1). Try to understand the example, Binary_Classifier.ipynb is given in …/Labs/Lab3. This sample is based on the **MNIST dataset**, which is a set of 70,000 small images of digits handwritten by high school students and employees of the US Census Bureau. Each image is labeled with the digit it represents, as explained in the lecture.

2). Two important performance measures, **''Confusion Matrix''** and **''Receiver Operator Curve (ROC)",** make sure of you understanding the concepts.

Task 2: Performance Evaluation and Multiclass Classification

1). Study the first part of codes titled **''Performance''**, given in Performance_Evaluation_and_Multiclass_Classification.ipynb, located in …/Labs/Lab3. See how to use **''confusion matrix''** and **''ROC''**.

2). (Optional)

- In the second part of this file, titled **''Multiclass Classification''**, the coding example demonstrates how use "multiclass classifiers" rather than "binary classifier", which I did not introduce in the lecture time. I have prepared some text in directory Lab3, called "Multiclass Classification.pdf", to assist you in understanding the codes.

Part 2: Logistic Regression with *Spark*

Logistic Regression is commonly used to estimate the probability that an instance belongs to a particular class (e.g., what is the probability that this email is spam?).

If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labeled "1"), or else it predicts that it does not (i.e., it belongs to

the negative class, labeled "0"). This makes it a **binary classifier**.

So, this is another way to have a "binary classifier".

Sample for Logistic Regression with Spark

This is a sample in real world, where the famous titanic dataset was employed. I have put more detailed comments inside of the codes. You will work with *categorical columns* and see how to use Spark *LogisticRegression* model, *pipeline* and its *evaluator*.

Actions:

Study the sample given in the notebook, called Titanic_Regression.ipynb, located in …/Labs/Lab3/Logistic_regression.

Part 3: ML – Clustering

As lectured, *K*-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.
Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically.

We will use Spark MLlib for our exercises, given as below.

Task 1: Example for Classification with Clustering

We'll be working with a real data set about seeds, from UCI repository:

https://archive.ics.uci.edu/ml/datasets/seeds.

The detailed specification of this task is presented in the notebook, Classification_with_clustering.ipynb, located in …/Labs/Lab3/Clustering.

I have added comments to the codes for assistance of your study.

Actions:

Study this example and see how to make a classification with Spark Clustering API.

Task 2: (Optional) A Project for Clustering Exercise

This is a real project for identifying 3 potential hackers, by using clustering technique. The data source - hack_data.csv, and the project specification - Clustering_Exercise.ipynb, are presented in …/Lab/Lab3.

Actions:

1) Try to solve this problem (by writing your Python codes with a notebook. Hint: review the lecture slides of Session 3 and see how to find the optimal K value for grouping the data.)

2) You may refer to its solution, given in Clustering_Exercise_SOLUTION.ipynb, for your learning.