Big Data Analytics (COMP4434)

Assignment two (20 marks in total)

(Due on 20 March 2018)

12 March 2018

1, [10 marks] Write a Scala program to train a linear regression model with the given data set. Use this model to make prediction for every data point and print all actual and predicted labels.
You are required to upload a Scala program file for this question.

Tips 1, the code to parse the dataset is provided：

```
val data = sc.textFile("Path/A2.data")
val parsedData = data.map { line =>
val parts = line.split(',')
LabeledPoint(parts(0).toDouble, Vectors.dense(parts(1).split(' ').map(_.toDouble)))
}.cache()
```

Tips 2, the method called LinearRegressionWithSGD can be used to train this model.

2, [10 marks] Table shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75

1992	75.7
1997	76.4
2002	76.9
2006	77.7
2010	78.5

You are required to upload a document to **answer** the following questions：

Based on the whole data, can you predict the life expectancy for an individual born in 2015? If The life expectancy in 2015 is 78.7, how to improve your prediction?