**Take-home Final Exam**

Ishwar Bhat Kattur

**IFT598: Analyzing Big Data**

Prof. Robert Rucker

05/04/2017

**Question 1 Part a**

*Before starting this regression question, it would help to be clear about the components and their meaning and uses: Using the components of Q1 to illustrate your answers:*

*0. entities ( these may be implicit as in the array below), observations, features*

*1. Independent and dependent variables*

*2. Categorical versus numerical features*

*3. Labels*

*4. Labeled versus unlabeled observations*

*5. Describe the connection between a regression algorithm and a regression model, and illustrate this with the classes used in the analysis below..*

**Entities** in Machine Learning are the crux of the problem for which we are trying to find a solution or build a model for. For example, in the Spam-Ham classification problem, Emails are the entities. Also, each entity is associated with real-valued label or a binary label (depending on whether we are working on a linear regression or logistic regression problem).

The number of training and test data we have make up the **observations**. For the given problem in 1b, fileMat has 5 **observations**. Usually, the more observations we have (assuming they are accurate), the better we can train the data.

Wikipedia says that a **feature** is an individual measurable property of the phenomenon being observed. In simple terms, features are inputs and **label** is the output. We use features to train our model to predict the **labels**.

**Categorical features** have categories (say “Gender”, “Sexual Orientation” or “Marital Status”) as features. **Numerical features** have numbers (like “age” which has to be a positive number) as feature values.

**Labels**, as mentioned above, are the outputs. For email classification, “Spam” and “Ham” are the two possible labels. **Labeled observations** have their observations labeled. These are used in supervised learning, where the model is trained with both inputs and the actual output. **Unlabeled observations** on the other hand are used in unsupervised learning, like clustering. We (or the ML model) have to figure out how to group the data, so that when a new input arrives, we can classify it to the correct group.

A **regression algorithm** is an algorithm for predicting relationship among variables. A **regression model** is an actual implementation of the algorithm, which when trained with the training dataset, can predict the output given a test dataset.

In this question down below, LinearRegressionWithSGD() is the algorithm (the set of rules which lay down the steps for building the model). model is the regression model (which we train by regression.run(parsedData)). This **model** is later used to predict output for custom input (val prediction = model.predict(point.features))

**Question 1 Part b**

We have the data as below:



First step is cleaning the data.



Next step: Split and get the X and Y vectors



Slope and intercept are calculated using First Principles:



**Question 1 Part c**

Constructing the RDD:



Scaled parsedData:



weights and intercepts using LinearRegressionWithSGD():



Testing with custom as well as given data:



**Question 2: DataFrames and Datasets**

The bodies.csv input file:



The case class required:



Defining the StructType:



DataFrame:



Using UDF to create dataset:



**Part b:** Surface area column added:



**Question 3: Sentiment Analysis**

Guller pg. 199



Using the SparkContext and SparkConf to create RDDs:



Parsing data:



DataFrame:



Making sure data is populated:



Tokenizer:



HashingTF:



LogisticRegression Pipeline:



Training and Test:



aucTraining and aucTest:



ParamGridBuilder:



Cross Validator:





Best Model, which can be used to classify new reviews:

