Assignment 1

CS564 : Foundation of Machine Learning
Department of CSE, IIT Patna

Date : 11-Aug-2019
Marks – 20

(Read all the instruction carefully and adhere to them.)
Instructions:

1. All the assignments should be completed and uploaded by ***18-Aug-2019, 11.00 pm.***

2. Markings will be based on the correctness and soundness of the outputs. Marks will be ***deducted in case of plagiarism***.

3. Be precise for your explanations in the report. Unnecessary verbosity will be penalized. Prepare a Detailed report of the assignment.

4. Code should be done in ***Python***.

5. You should zip all the required files and name the zip file as **rollno1_rollno2_rollno3_assignment1.zip**, e.g., 1811cs01_1811cs02_1811cs03_assignment1.zip.

6. Upload your solution(zip file) to the following link:

   https://www.dropbox.com/request/eWE7CiUXKsTma79iwf43

Questions:

(1) The crucial task before applying any machine learning algorithms is to understand the given data, i.e., a thorough data analysis cum data visualization is always necessary. As the part of this assignment, you are given a dataset, from which the following informations are to be extracted.

Dataset : **stackOverflow.csv**

Information to be extracted out:

1. Find out the no. of questions asked with respect to the given Tags.
2. Find out the most commonly used tags and what is the trend in Data Science Tags.
3. The average time is taken to answer a question.
4. Numbers of views related to the number of Answers.
5. Tags get highest/lowest rating in Questions.
6. Tags get highest/lowest rating in Answers.
7. Find out the most Active/Inactive in answering the questions.
8. Which tags draws the highest/lowest views?

Point to be noted :

1. You need to infer the above imformations using proper graph, wherever necessary.
2. You must do the code stuff in Python only.

Dataset is to be downloaded from the below mentioned link:

https://drive.google.com/file/d/0B1AC_DBfxZmWS0pMbWsyNUJrV083akMtVV81NmViRjcxbmhj/view?usp=sharing

(2) Consider the training dataset **data.csv**, which has 8 variables, as follows.

"NumPreg","PlasmaGlucose", "DiastolicBP", "TricepSkin", "BodyMassIndex" ,"Pedigree" "Age", "Diabetic"

The target is to fit a logistic regression model to predict the "Diabetic" variable based on the other 7 variables. In this connection, please answer the following questions, in given sequence.

1. Develop the best model to predict the categorical response variable "Diabetic" in case of the given dataset? Justify your choice for best model.

2. Suppose you have chosen a threshold t to classify P(Diabetic | X) > t as "Diabetic" = Yes. How would you choose the optimal threshold t such that the aforesaid classification achieves maximum accuracy for your best model? Justify your choice.

This dataset is to be downloaded from the below mentioned link:

https://drive.google.com/file/d/0B1AC_DBfxZmWNkZ2QXVSVnVRbXQzVldQNFJsTnloRVlvN0Rv/view?usp=sharing