BIOMEDICAL DATA SCIENCE HOMEWORK #2

DUE DATE 10/16/2018

- Please turn your homework code through a GitHub Repository (see below).
- No late homework will be accepted (you will not be able to submit the repository link to Canvas after the due date). Do not change the repository after the due date; your homework will not be graded.
- 20 points total

Use the accompanied dataset for this homework. Read the dataset description below carefully and make sure you understand the dataset features and values.

> *Dataset description: This dataset is composed of a range of biomedical voice measurements from patients with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.*
>
> *The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording; the name of the patient is identified in the first column.*

a) Load the dataset in an iPython notebook [2 point].

b) Split the dataset for training a knn classifier in a holdout + cross-validation fashion. Run the k-nn (k-nearest neighbors) algorithm on this dataset and find the optimal k using Grid search, where k={4, 5, 6, 7, 8, 9, 10} (Hint: use the GridSearchCV operator in sklearn). [10 points]

c) Plot the learning curve, as well as the validation curve [6 points].

d) Create a GitHub repository and share your code via GitHub with the instructor by submitting the link on Canvas [2 points].