MATH 4397, Project: Final Report.

Instructor: Andrey Skripnikov

**DUE: Friday, May** 3**rd, at 11:59PM.**

**At least** 6 **pages, no more than** 10 **pages.**

**Please write the names of all members for your project group. Submission remarks:**

* Provide only the most critical pieces of code in this report, without including too much output and clogging up the report. Simply attach the full *R* source code to your submission in order for me to have access to the whole thing.
* For each section/subsection, make sure to specify the names of students who mostly worked on that part of the report. E.g. ”2. Hierarchical Clustering (Johnny Doey, Brenda Parker)”.

1. Overall intro, including:
   1. A very short description of your data (main ideas from your ”Data Description” project assignment).
   2. The questions/tasks you are posing about this data (main excerpts from your ”Formulating Questions/Models” project assignment).
2. Then, for at least one of the overall data questions that you have (it’s fine, and probably even **preferred**, if you just have **one main overall question**, which deals with one response variable; try to **avoid** asking multiple questions, each about a **different response variable**) :
   1. Formulate the question. Outline the two models you are using to answer it, alongside the reasons why, and what are the expected advantages/disadvantages of each model.
   2. For each of the two models, proceed to:
      1. Write down the model formula. If it’s KNN - write how we calculate the predictions for the observation based on its neighbors; SVM - write the hyperplane equations corresponding to selected kernels; *K*-Means clustering write the explicit formula for main optimization criteria; Hierarchical clustering - explain how clusters are merged depending on the linkage used.

1

* + 1. Walk me through the thought process of your considerations while fitting the model. E.g. would you consider excluding some predictors from consideration, and why? Some predictors might not make sense from domain knowledge, or they might be highly correlated with other predictors. What about scaling the data - is it necessary, and why?
    2. Select the optimal tuning parameter value for your model. E.g. select optimal *K* for KNN or *K*-Means; select optimal values of *cost,γ* or *d* for your SVM; select optimal linkage for your hierarchical clustering (by selecting optimal *K* for each linkage beforehand).
    3. If you are performing **supervised** learning (KNN, SVM), then
       1. randomly subdivide your full data set in 80% for training, and 20% for testing (**Note:** make sure to use *set.seed*() in order for both models to work on the train/test data subdivision.)
       2. Proceed to train your model on the 80% training data, and then record its prediction error on left out 20% testing data.

If you are performing **unsupervised** learning (clustering), then

A. Make sure to calculate the silhouette coefficient for the optimal clustering.

1. For **supervised** learning: compare the prediction test error resulting from part (b) for both models.

For **unsupervised** learning: compare optimal silhouette coefficients for both clustering methods; use external class labels (if available) to check if the clustering makes sense (see some of HW problems on *iris* and gene expression data).

1. Proceed to fit your **best** model from part (c) on the **full data set** (no train/test subdivision), and
   1. Output the results, providing the most important model summaries and images (if applicable) that resulted from your model fitting. For KNN summary of your KNN object, its classification performance on full data. For SVM - summary of SVM object, # of support vectors in each class, a plot of your fitted boundary with respect to some **pairs of predictors** (see one of HW assignments). For *K*-means clustering - use *eclust*() with its 2*D* plot; for hierarchical clustering - output dendrogram.
   2. Provide the interpretation of results and your conclusions as it pertains to the original overall question. E.g. for **supervised** learning, did you achieve a respectable predictive performance? For **unsupervised**, what kind of groups you’ve obtained? Do they make sense?

2