STAT 6240 - HW 1 - Due ~~1/31~~ 2/7

Due Date: ~~January 31st~~ February 7th, Thursday, 6:10 PM

Your homework submission should contain: (i) Your R code, (ii) outputs from R (including figures), (iii) answers to the questions.

You should use *R Markdown* or *R Notebooks*. **Do not print out your homework.** Save it as a pdf (or an html) file and upload it to Blackboard.

Question 1 (60 Points)

- Install the `titanic` package from CRAN and load the `titanic_train` dataset, and check its help file to learn what the dataset contains.
- Remove the `PassengerId`, `Name`, `Ticket` and `Cabin` columns, and transform the `Fare` variable by taking its log.
- Choose three variable pairs (for instance, "Sex" and "Survived" is one pair) and plot their distribution in the dataset by using appropriate plots. You can try mosaic plots, densities, histograms. You are asked to provide at least one plot for each pair.
- Create a model matrix of the dataset that only contains numbers (no factors!) by using the `model.matrix` function. Then, remove the `Survived` variable from this dataset.
- Fit a PCA to your matrix from the previous step. Plot the scores of the observations (use only the first 2 dimensions) and color them according to the `Survived` variable.
- Repeat the previous question with NMF (non-negative matrix factorization). Use `rank=2` for the fit. Note which variables were chosen.
- Finally, load the `titanic_test` dataset. Using your fitted PCA from the previous stages, obtain the 2 dimensional projections of the test dataset.
- Obtain the `Survived` variable for the test dataset by combining the test dataset with the full dataset at the following link: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls.
- Plot the scores with respect to the `Survived` variable.

Question 2 (40 Points)

- Choose a dataset from Kaggle, any of the R packages or generate your own dataset by capturing photos from Amazon.com with the "Image Downloader" extension of Google Chrome as we did in class. If the dataset size is too large, randomly choose 2000 samples. Feel free to discard any variables that are not numbers or factors.
- Use PCA and at least one other method (this could be Logistic PCA, Sparse PCA, NMF or any other factorization method) to obtain a low dimensional representation of the data. Provide plots of the low dimensional representations.
- Analyze the fitted models in detail. Which variables are chosen or have a larger magnitude? Do the 2-dimensional plots suggest anything specific about the dataset? Write a summary of your results; your summary should have at least 150 words. **We will discuss these solutions in class.**