STAT 8330: **Homework 7**

Due on December 9, 2020

Consider the `footwear` dataset that is on the Canvas website. This dataset has 18000 images of various footwear (types of shoes). Each image is a 28×28 grayscale image, but is vectorized as a 784 element vector (so, the dataset is a 784×18000 matrix in ASCII text format – very large!). Here is how you can read the data in and orient the first image properly (notice, it looks like a shoe):

```
> library(readr)
> setwd("~/Box Sync/Courses/Stat_DAIII/Fall2020/Homework/HW7")
> shoes <- read_table('footwear.txt', col_names=F)
> shoes <- t(shoes)
> rotate <- function(x) t(apply(x,2,rev))
> image(rotate(matrix(shoes[1,], nrow=sqrt(ncol(shoes)))))
```

1. Make 28×28 image plots of the mean and standard deviation of all the images.
2. Perform a PCA on the data using the singular value decomposition – make sure you remove the overall spatial (pixel) mean from each image first.
   (a) How much variance is accounted for by each of the first 5 PCs?
   (b) Plot the weights associated with the first 2 PCs (these will be images) and comment on what you see.
   (c) Use the randomization approach discussed in lecture to determine the number of principal components that are not "noise."
3. Using the first four principal components as features, use a K-means classifier with $K = 3$ to classify each image into 3 groups.
   (a) Use a scatter plot to plot the group classification (1,2,3) by color in 2 dimensions, with x-axis corresponding to the first PC and the y-axis corresponding to the second PC. Are the groups well separated in 2 dimensions?
4. Use the `kpca` function in the `kernlab` package to perform kernel PCA on these data. Try different choices of kernels and parameters and report your results. Use the first 4 kernel PCs to repeat the K-means classification from above, and make another scatterplot as above.
5. Use the `NMF` package to perform non-negative matrix factorization on these data. Plot the first 6 weight matrices (as images) and describe how these compare to the PCA weight matrices from above.
6. Use the `LLE` package to perform local linear embedding (LLE) on these data. What can you say about the first two dimensions of the new features found from this procedure? Repeat the K-means clustering ($K = 3$) from above (using 2 dimensions from the LLE) and repeat the scatterplot as above. **Extra Credit:** Produce a plot similar to the one shown on page 18/19 in the lecture 24 notes and interpret it. (Note, the images don't need to be within the main scatterplot.)