This homework uses the same dataset with over 20,000 recipes from the website Epicurious as Homework 5A.

Support Vector Machines

1. Using the Epicurious dataset, use support vector machines to predict whether or not a recipe is tagged as cake. You may experiment with the features (i.e., use different combinations, use the principal components from last homework, etc.). Report your accuracy, sensitivity, specificity, etc.

2. Plot the ROC curve for your support vector machine model. Add the ROC curve from your best model from Homework 5A to same plot. Find the areas under the curve for both models. Which one performs better?

k-Means Clustering

1. Apply k-means clustering for a range of k to all of the Epicurious features. Most of the features are categorical and should already be coded as indicator variables (0's and 1's), but there are a few numeric variables corresponding to the nutrition facts for each recipe. Make a plot of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the residual within cluster variance as a function of k and choose the best k.

2. Using the k that you found in the previous step, produce a visualization that explains how some of the clusters differ using the following steps:

   - So that you can visualize the clusters in two dimensions at a time, perform Principal Component Analysis (a.k.a., singular value decomposition). This should be very similar to your analysis from Homework 5A, Part 2.
   - Create scatter plots of the first two principal components, PC1 and PC2. Label the axes with the fraction of the variance explained by PC1 and PC2.
   - Color each point according to the cluster labels.
   - Repeat the previous two steps for the following pairs of components: PC3 and PC4, PC5 and PC6, PC7 and PC8, and PC9 and PC10. You should not have to re-perform PCA on the features.

3. Try to come up with a "name" or description for at least a few of the clusters–are they easily interpretable?