DATA 221
due Thursday, February 9 at 11:59pm **Homework 5A** Winter 2023

Problem 1

A Kaggle user has created a dataset with over 20,000 recipes from the website Epicurious. Loosely speaking, there are a few groups of variables in the dataset:

- The outcome of interest (`cake`, whether a recipe is tagged as a cake)
- The nutritional variables (`calories`, `protein`, `fat`, `sodium`)
- Ingredient tags (`almond`, `amaretto`, `anchovy`, and so on)
- Place tags (`alabama`, `alaska`, `aspen`, `australia`, and so forth)
- Other tags (`advance.prep.required`, `anthony.bourdain`, etc.)

You may want to carry out your own exploratory dataset to a) become familiar with the dataset, and b) carry out some pre-processing on the dataset. You will have to make choices during pre-processing–these choices are up to you. We are happy to offer suggestions during office hours, but as long as you feel that you have made a reasonable choice, we will not take points off.

Once you have pre-processed the data. . .

(a) Logistic Regression with ℓ_1 Regularization

(a) Find the logistic regression coefficients for this dataset with ℓ_1 regularization for several different values of t (at least 10–it's probably easiest to write a loop here and evaluate the function over a range of t values).

(b) Plot the regression coefficients as a function of the (logarithm of the) regularization parameter t–example on the second page of the assignment. If you can get `glmnet` to work, you can use the functions the authors of the book used and create an identical plot, but Python installation is a bit of a headache.

(c) Find the optimum regularization parameter t by optimizing for minimum error on the test set.

(a) Principal Component Analysis

(a) Perform Principal Component Analysis (a.k.a., singular value decomposition) on all the features of the dataset.

(b) Display scatter plots of the first two principal components, PC1 and PC2, for the two response variable classes. Label the axes with the fraction of the variance explained by PC1 and PC2.

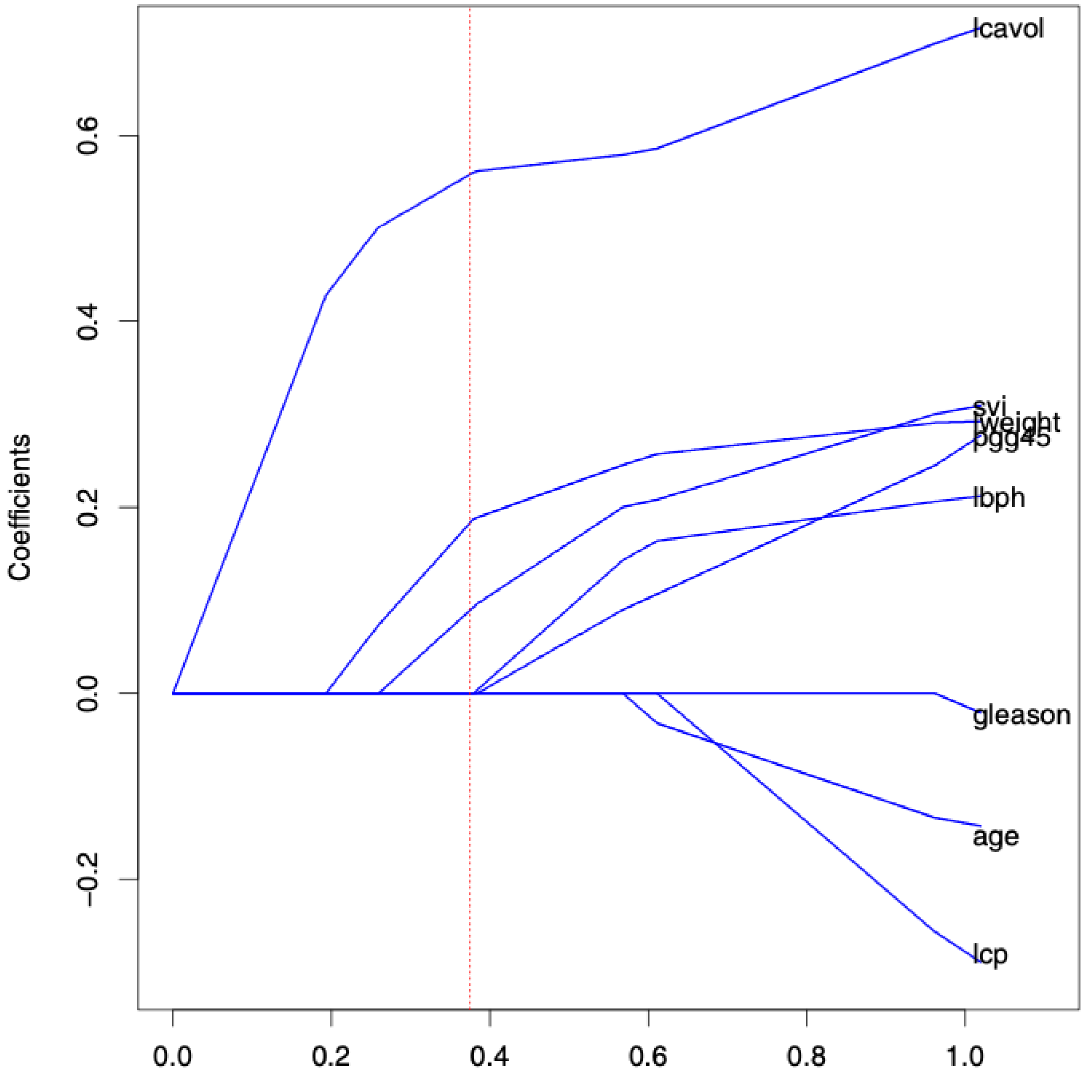
(c) Compute a table with the fraction of the variance in each of the first few principal components. How many principal components would you pick?

(a) Logistic Regression and Principal Components

(a) Using the principal components as features, fit an ℓ_1-regularized logistic regression for several different values of t.

(b) Plot the feature coefficients as a function of the logarithm of the regularization parameter t.

(c) Find the optimum regularization parameter t by minimizing error on the test set. Does the optimum include more or fewer features than the model from Part A?


Coefficients
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp