ISYE 6740 Homework 5
Spring 2022
Total 100 points.

1. **Conceptual questions.** (30 points)

   (a) (15 points) Consider the mutual information based feature selection. Suppose we have the following table (the entries in table indicate counts) for the spam versus and non-spam emails:

	"prize" = 1	"prize" = 0
"spam" = 1	150	10
"spam" = 0	1000	15000

	"hello" = 1	"hello" = 0
"spam" = 1	155	5
"spam" = 0	14000	1000

   Given the two tables above, calculate the mutual information for the two keywords, "prize" and "hello" respectively. Which keyword is more informative for deciding whether or not the email is a spam?

   (b) (15 points) Given two distributions, $f_0 = \mathcal{N}(0,1)$, $f_1 = \mathcal{N}(1.5, 1.1)$ (meaning that we are interested in detecting a mean shift of minimum size 3), explicitly derive what the CUSUM statistic should be (i.e., write down the CUSUM detection statistic).

   Plot the CUSUM statistic for a sequence of randomly generated samples, $x_1, \ldots, x_{100}$ are are i.i.d. (independent and identically distributed) according to f_0 and $x_{101}, \ldots, x_{200}$ that are i.i.d. according to f_1.

2. **House price dataset.** (30 points)

The HOUSES dataset contains a collection of recent real estate listings in San Luis Obispo county and around it. The dataset is provided in RealEstate.csv. You may use "one-hot-keying" to expand the categorical variables.

The dataset contains the following useful fields (You may exclude the Location and MLS in your linear regression model).

You can use any package for this question.

Note: We suggest you scale the independent variables (but not the dependent variable). We also suggest you use our suggested seeds, as this dataset is particularly seed dependent.

- Price: the most recent listing price of the house (in dollars).
- Bedrooms: number of bedrooms.
- Bathrooms: number of bathrooms.
- Size: size of the house in square feet.
- Price/SQ.ft: price of the house per square foot.
- Status: Short Sale, Foreclosure and Regular.

(a) (15 points) Fit the Ridge regression model to predict Price from all variable. You can use one-hot keying to expand the categorical variable Status. Use 5-fold cross validation to select the regularizer optimal parameter, and show the CV curve. Report the fitted model (i.e., the parameters), and the sum-of-squares residuals. You can use any package. The suggested search range for the regularization parameter is from 1 to 80, and the suggested seed is 2.

(b) (15 points) Use lasso to select variables. Use 5-fold cross validation to select the regularizer optimal parameter, and show the CV curve. Report the fitted model (i.e., the parameters selected and their coefficient). Show the Lasso solution path. You can use any package for this. The suggested search range for the regularization parameter is from 1 to 3000, and the suggested seed is 3.

3. **Medical imaging reconstruction.** (40 points)

In this problem, you will consider an example resembles medical imaging reconstruction in MRI. We begin with a true image image of dimension 50×50 (i.e., there are 2500 pixels in total). Data is cs.mat; you can plot it first. This image is truly sparse, in the sense that 2084 of its pixels have a value of 0, while 416 pixels have a value of 1. You can think of this image as a toy version of an MRI image that we are interested in collecting.

Because of the nature of the machine that collects the MRI image, it takes a long time to measure each pixel value individually, but it's faster to measure a linear combination of pixel values. We measure $n = 1300$ linear combinations, with the weights in the linear combination being random, in fact, independently distributed as $\mathcal{N}(0, 1)$. Because the machine is not perfect, we don't get to observe this directly, but we observe a noisy version. These measurements are given by the entries of the vector

$$y = Ax + n,$$

where $y \in \mathbb{R}^{1300}$, $A \in \mathbb{R}^{1300 \times 2500}$, and $n \sim \mathcal{N}(0, 25 \times I_{1300})$ where I_n denotes the identity matrix of size $n \times n$. In this homework, you can generate the data y using this model.

Now the question is: can we model y as a linear combination of the columns of x to recover some coefficient vector that is close to the image? Roughly speaking, the answer is yes.

Key points here: although the number of measurements $n = 1300$ is smaller than the dimension $p = 2500$, the true image is sparse. Thus we can recover the sparse image using few measurements exploiting its structure. This is the idea behind the field of *compressed sensing*.

The image recovery can be done using lasso

$$\min_x \|y - Ax\|_2^2 + \lambda \|x\|_1.$$

(a) (20 points) Now use lasso to recover the image and select λ using 10-fold cross-validation. Plot the cross-validation error curves, and show the recovered image.

(b) (20 points) To compare, also use ridge regression to recover the image:

$$\min_x \|y - Ax\|_2^2 + \lambda \|x\|_2^2.$$

Select λ using 10-fold cross-validation. Plot the cross-validation error curves, and show the recovered image. Which approaches give a better recovered image?