CS 7641 CSE/ISYE 6740 Homework 5

- Submit your answers as an electronic copy on T-square.
- No unapproved extension of deadline is allowed. Zero credit will be assigned for late submissions. Email request for late submission may not be replied.
- For typed answers with LaTeX (recommended) or word processors, extra credits will be given. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- You are encouraged to work with some collaborators especially if you are not familiar with Matlab.
- Explicitly mention your collaborators if any.

1 Recover an image using compressed sensing [60 points]

In this problem, you will consider choosing the tuning parameters for both ridge regression and the lasso, using 10-fold cross-validation. First download the data cs.mat.

We begin with a true image image of dimension 50 × 50. 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.

Suppose that, 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 $\mathcal{N}(0, 1)$. These measurements are given by the entries of the vector A*double(img(:)) in our MATLAB code. Because the machine is not perfect, we don't get to observe this directly, but we see a noisy version of this. Hence, in terms of our code, we observe y = A*double(img(:)) + 5*randn(1300,1) 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, and the weights in a linear combination are i.i.d normal. This is the idea behind the field of *compressed sensing*.

The file model-selection.m is setup to perform ridge regression of y on x, and the lasso of y on x, with the tuning parameter for each method selected by cross-validation. You will fill in the missing pieces. It's helpful to read through the whole file to get a sense of what's to be accomplished. Try to understand all the parts, even if it doesn't seem related to what you have to fill in; this should be good practice for working with MATLAB in the future, etc. You may build your code based on model-selection.m.

Plot the cross-validation error curves for each of ridge regression and the lasso. For both ridge regression and the lasso, what value of λ has a smaller minimum cross-validation error?

Also plot the recovered images under the optimal choices of regularizing parameters for the ridge regression and lasso, respectively. Which one do you think recovers a better image?

2 House price estimation through ridge regression [40 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.

The dataset contains the following fields:

- MLS: Multiple listing service number for the house (unique ID).
- Location: city/town where the house is located. Most locations are in San Luis Obispo county and northern Santa Barbara county (Santa Maria-Orcutt, Lompoc, Guadelupe, Los Alamos), but there some out of area locations as well.
- 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: type of sale. Thee types are represented in the dataset: Short Sale, Foreclosure and Regular.

*Please ignore the "Location" variable below. It is a categorical variable. (Extra hint: To deal with categorical variable, we typically will convert them using "one-hot" keying and then decide to include "Location" or not, using group lasso. But you are not required to do this for this question.)

1. Fit ridge regression model to predict Price using remaining factors (except Status), for each of the three types of sales: Short Sale, Foreclosure and Regular, respectively. Find optimal λ using cross-validation.
2. Using lasso to select the 2 leading factors for Price, for each of the three type of sales, for each of the three types of sales: Short Sale, Foreclosure and Regular, respectively. Report the factors you found. Find optimal λ using cross-validation.

You may use built in MATLAB functions.