HW 5

1. Conceptual Questions

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

	spam = 1	**spam = 0**
prize = 1	N_{11} = 150	N_{10} = 1000
prize = 0	N_{01} = 10	N_{00} = 15000

$$I(prize, spam) = \frac{N_{11}}{N} log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} log_2 \frac{NN_{00}}{N_{0.}N_{.0}}$$

$$N = 150 + 1000 + 10 + 15000 = 16160$$
$$N_{1.} = 150 + 1000 = 1150$$
$$N_{.1} = 150 + 10 = 160$$
$$N_{0.} = 10 + 15000 = 15010$$
$$N_{.0} = 1000 + 15000 = 16000$$

$$I(prize, spam) = \frac{150}{16160} log_2 \frac{16160*150}{1150*160} + \frac{10}{16160} log_2 \frac{16160*10}{15010*160} + \frac{1000}{16160} log_2 \frac{16160*1000}{1150*16000} + \frac{15000}{16160} log_2 \frac{16160*15000}{15010*16000}$$
$$= 0.03296011876395397$$

	spam = 1	**spam = 0**
hello = 1	N_{11} = 155	N_{10} = 14000
hello = 0	N_{01} = 5	N_{00} = 2000

$$I(hello, spam) = \frac{N_{11}}{N} log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} log_2 \frac{NN_{00}}{N_{0.}N_{.0}}$$
$$= 0.0007839352232340266$$

$$N = 155 + 14000 + 5 + 2000 = 16160$$
$$N_{1.} = 155 + 14000 = 14155$$
$$N_{.1} = 155 + 5 = 160$$
$$N_{0.} = 5 + 2000 = 2005$$
$$N_{.0} = 14000 + 2000 = 16000$$

Since $I(prize, spam) > I(hello, spam)$, **prize** is more informative for deciding whether or not the email is spam.

(b) Given two distributions, $f_0 = N(0, 1), f_1 = N(1.5, 1.1)$, explicitly derive what the CUSUM statistic should be.

From page 22 the Anomaly Detection lectures notes:

$$W_t = max(W_{t-1} + log\frac{f_1(X_t)}{f_0(X_t)}, 0)$$

Completing the PDF of Normal Distributions:

$$f_1(X_t) = N(1.5, 1.1) = \frac{1}{\sqrt{2\pi 1.1}} e^{\frac{-1}{2}(\frac{x-1.5}{\sqrt{1.1}})^2}$$
$$f_0(X_t) = N(0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}x^2}$$

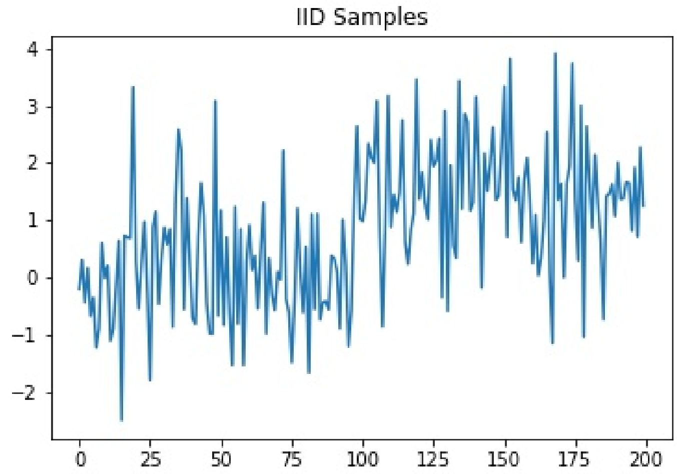
Simplifying:

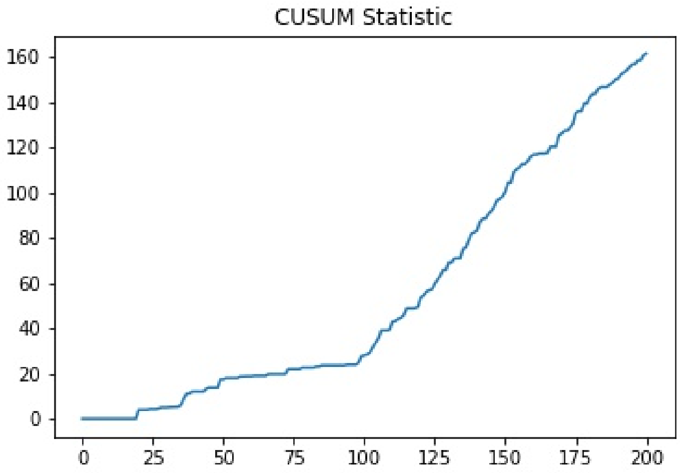
$$\frac{f_1(X_t)}{f_0(X_t)} = \frac{1}{\sqrt{1.1}} e^{\frac{1}{2}(x^2-(\frac{x-1.5}{\sqrt{1.1}})^2)}$$

$$log(\frac{f_1(X_t)}{f_0(X_t)}) = log(\frac{1}{\sqrt{1.1}} e^{\frac{1}{2}(x^2-(\frac{x-1.5}{\sqrt{1.1}})^2)}) = log(\frac{1}{\sqrt{1.1}}) + log(e^{\frac{1}{2}(x^2-(\frac{x-1.5}{\sqrt{1.1}})^2)}) = log(\frac{1}{\sqrt{1.1}}) + \frac{1}{2}(x^2 - (\frac{x-1.5}{\sqrt{1.1}})^2) =$$

$$\longrightarrow W_t = max(W_{t-1} - log(\sqrt{1.1}) + \frac{1}{2}(x^2 - (\frac{x-1.5}{\sqrt{1.1}})^2), 0)$$

Plot the CUSUM statistic for a sequence of randomly generated samples that are i.i.d $x_1, \ldots, x_{100} \backsim f_0, x_{101}, \ldots, x_{200} \backsim f_1$





2. House Price Dataset

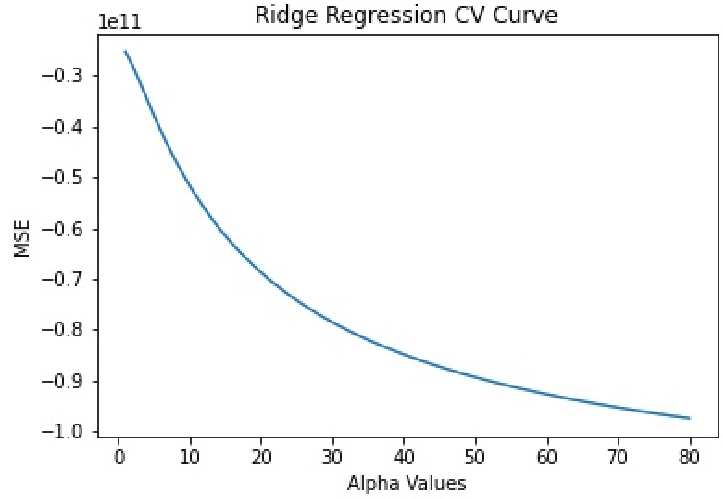
The HOUSES dataset contains a collection of recent real estate listings in San Luis Obispo county and around it. The dataset is provided in RealEstate.csv. You may use “one-hot-keying” to expand the categorical variables.
The dataset contains the following useful fields (You may exclude the Location 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) 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. The suggested search range for the regularization parameter is from 1 to 80, and the suggested seed is 2.

The chosen α is 1 based on the negative mean squared error scoring metric.

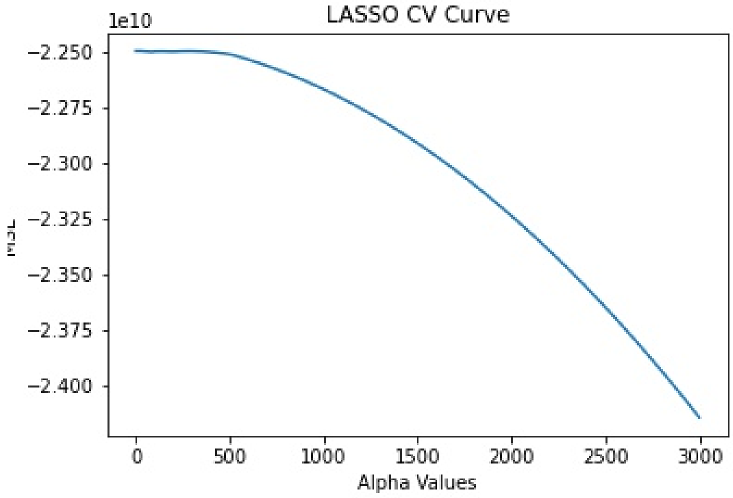


The fitted model coefficients are:
Intercept = -314766.34852713783
Bedrooms = 24264.47041987
Size = 1579560.26950984
Price/SQ.Ft = 1849370.30103488
Foreclosure = -15396.50807562
Regular = 40911.75979374
Short Sale = -25515.25171813

Sum of Squares Residuals = 17,003,281,203,675

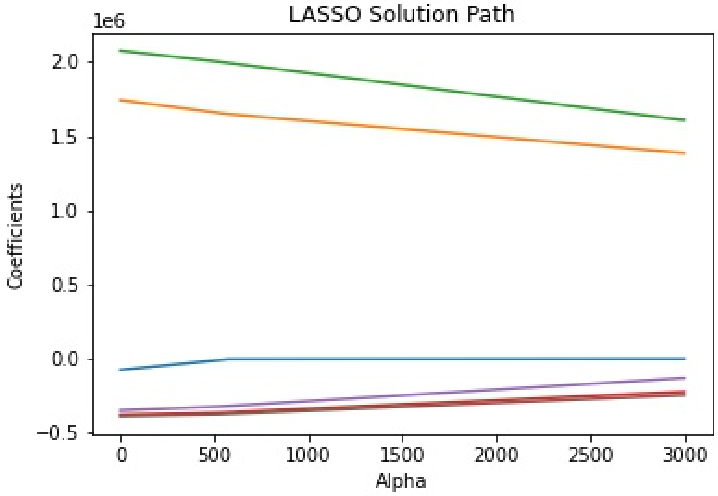
(b) Use lasso to select variables. Use 5-fold cross validation to select the regularizer optimal parameter, and show the CV curve. Report the fitted model (i.e., the parameters selected and their coefficient). Show the Lasso solution path. The suggested search range for the regularization parameter is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

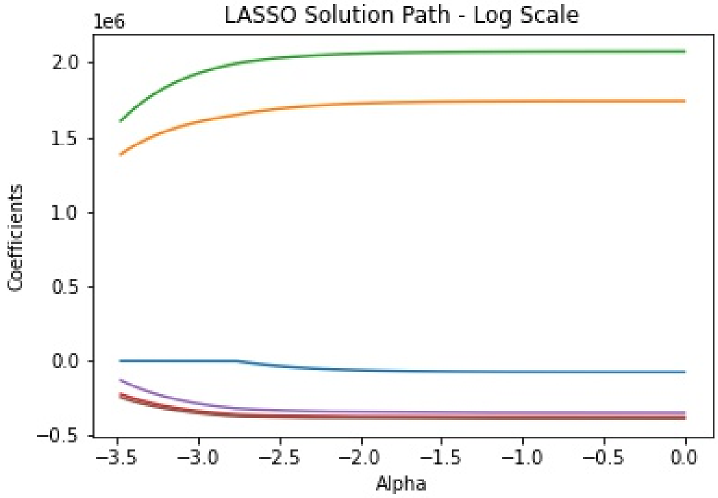
The chosen α is 11 based on the negative mean squared error scoring metric.



The fitted model coefficients are:
Intercept = -380005.4024880027
Bedrooms = -71614.75646417
Size = 1741200.2885389
Price/SQ.Ft = 2074627.53267139
Foreclosure = 4610.26062518
Regular = 33538.84775991
Short Sale = -4583.3327755

Sum of Squares Residuals = 16,330,843,510,219

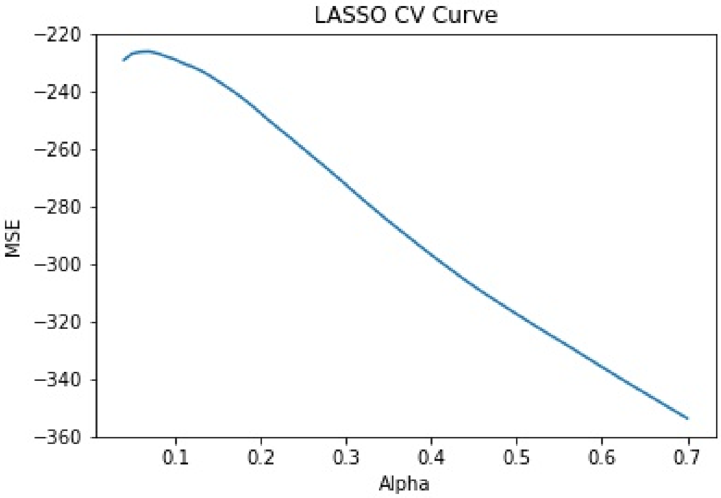




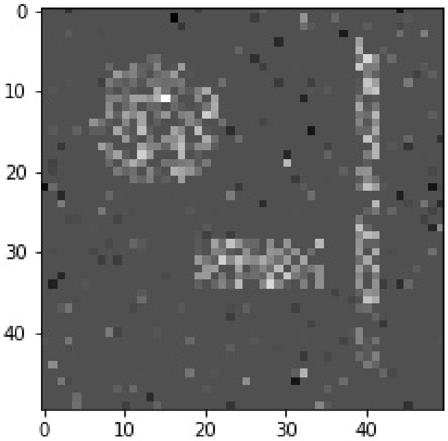
3. Medical Imaging Reconstuction

(a) 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.

To find lambda, I tested multiple ranges and found that smaller values (< 1) of λ worked best. The ideal λ is 0.07.

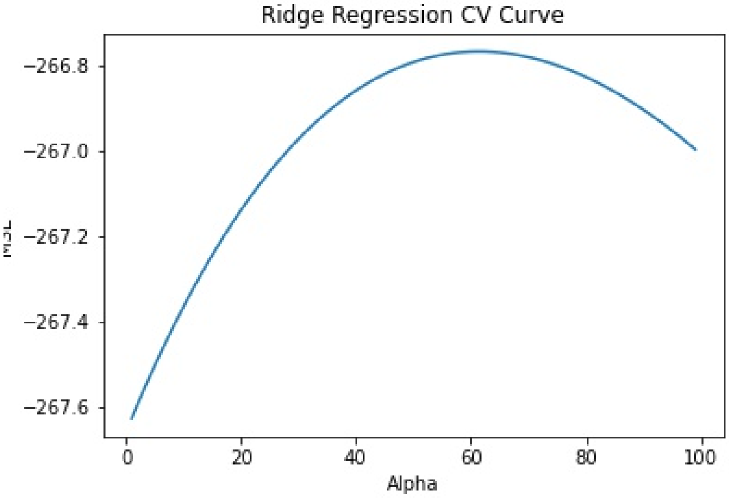


Reconstructed Image:

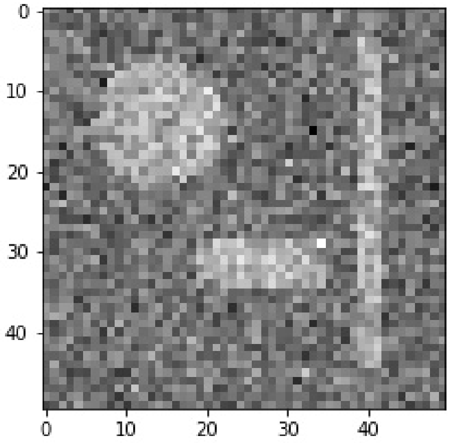


(b) To compare, use ridge regression to recover the image. Use λ using 10-fold cross validation. Plot the cross-validation error curves and show the recovered image.

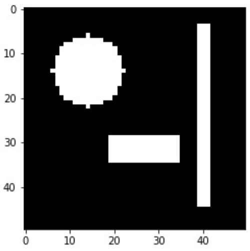
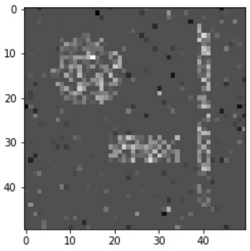
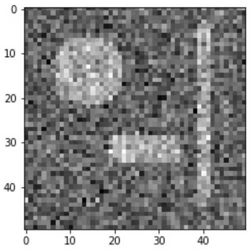
For ridge regressiosn, larger vaues of λ worked best. The ideal λ is 61.



Reconstructed Image:



Which approach gives a better recovered image?

Original	LASSO	Ridge Regression
		

Ridge regression provides a more noisy reconstructed image with varying shades of grey but you can clearly see all 3 shapes. LASSO provides a less noisy reconstruction but is missing parts of each shape. For the purpose of an MRI, I would say ridge regression provides the better recovered image.