Midterm Exam

Advanced Data Mining

Baruch College

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the college's educational mission and the students' personal and intellectual growth. Baruch students are expected to bear individual responsibility for their work, to learn the rules and definitions that underlie the practice of academic integrity, and to uphold its ideals. Ignorance of the rules is not an acceptable excuse for disobeying them. Any student who attempts to compromise or devalue the academic process will be sanctioned.

This is a take home exam. The result you upload should only be your work.

Present your results in a crisp, clear, concise and readable fashion. For theoretical questions, present all the necessary steps to obtain the result but not more or less than what is needed to make your proof/calculation/argument complete and correct.

1. Suppose we observe a vector $y \in \mathbb{R}^n$ of observations from the model,

$$y = X\beta^* + \epsilon,$$

where $X \in \mathbb{R}^{n \times p}$ is a fixed matrix of predictor variables, $\beta^* \in \mathbb{R}^p$ is the true unknown coefficient vector that we would like to learn, and $\epsilon \in \mathbb{R}^n$ is a random error vector, with

$$\mathrm{E}[\epsilon] = 0, \quad \mathrm{Cov}(\epsilon) = \sigma^2 I.$$

a. Prove that $\mathrm{E}[y] = X\beta^*$ and $\mathrm{Cov}[y] = \sigma^2 I$.

b. Suppose that $\hat{\beta} \in \mathbb{R}^p$ is a minimizer of the least squares problem:

$$\hat{\beta} \;=\; \arg\min_{\beta} \|y - X\beta\|_2^2 . \tag{1}$$

Show that if $v \in \mathbb{R}^p$ is a vector such that $Xv = 0$, then $\hat{\beta} + c.v$ is also a minimizer of the least squares problem, for any $c \in \mathbb{R}$.

c. If the columns of X are linearly independent, then what vectors $v \in \mathbb{R}^p$ satisfy $Xv = 0$?

d. Suppose that $p > n$. Show that there exist a vector $v \neq 0$ such that $Xv = 0$. Argue based on part b, that there are infinitely many linear regression estimates. Furthermore argue that there is a variable $i \in \{1, \ldots, p\}$ such that the regression coefficient of variable i can have different signs, depending on which estimate we choose. Comment on this.

e. Recall the ridge regression estimate,

$$\hat{\beta}^{\mathrm{ridge}} \;=\; \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{2}$$

Show that $\hat{\beta}^{\mathrm{ridge}}$ is simply the vector of linear regression coefficients from regressing the response $\tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{n+p}$ onto the predictor matrix $\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \in \mathbb{R}^{(n+p) \times p}$, where $0 \in \mathbb{R}^p$, and $I \in \mathbb{R}^{p \times p}$ is the identity matrix.

f. Show that the matrix $\tilde{X}$ always has full column-rank, i.e. its columns are always linearly independent, regardless of the columns of X. Hence argue that the ridge regression estimate is always unique, for any matrix predictors X.

g. Write out an explicit formula for $\hat{\beta}^{\mathrm{ridge}}$ involving X, y, λ.

h. Let X have singular value decomposition $X = UDV^\top$, where $U \in \mathbb{R}^{n \times r}, D \in \mathbb{R}^{r \times r}, V \in \mathbb{R}^{p \times r}$, U, V have orthonormal columns, and D is diagonal with elements $d_1 \geq \ldots \geq d_r \geq 0$. Rewrite your formula for the ridge regression solution $\hat{\beta}^{\mathrm{ridge}}$ by replacing X with $UDV^\top$, and simplifying the expression as much as possible.

i. In statistics, the bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. Let $\tilde{x} \in \mathbb{R}^p$. Prove that the estimate $\tilde{x}^\top \hat{\beta}^{\mathrm{ridge}}$ is indeed a biased estimate of $\tilde{x}^\top \beta^*$, for any $\lambda > 0$.

2. a. Given $y \in \mathbb{R}^n$, consider the ridge regression with predictor matrix $X = I_{n\times n}$, i.e.,

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2. \quad (3)$$

Show that the solution is

$$\hat{\beta}_i^{\text{ridge}} = \frac{y_i}{1+\lambda}, \quad i = 1, \ldots, n. \quad (4)$$

b. For the lasso with identity predictor matrix,

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \quad (5)$$

the solution (bonus problem) is

$$\hat{\beta}_i^{\text{lasso}} = \text{sign}(y_i).\max(|y_i| - \lambda/2, 0), \quad i = 1, \ldots, n. \quad (6)$$

For a fixed value of λ (e.g. $\lambda = 1$), draw $\hat{\beta}_i^{\text{lasso}}(y_i)$ and $\hat{\beta}_i^{\text{ridge}}(y_i)$ as a function of y_i. Describe the difference between these two coefficient functions.

c. Suppose that $X \in \mathbb{R}^{n\times p}$ is orthogonal, i.e. $X^\top X = I_{p\times p}$. Consider the ridge regression and lasso problems:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \quad (7)$$

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \quad (8)$$

Let $X_i \in \mathbb{R}^n$ denote the ith column of X. Show that the solutions $\hat{\beta}^{\text{ridge}}, \hat{\beta}^{\text{lasso}}$ are given by the same formulas as in the previous parts, but with $X_i^\top y$ in place of y_i (and p in place of n).

(Hint 1: if $U \in R^{n\times n}$ is an orthogonal and square matrix, recall that it preserves distances, i.e. $\|Uz\|_2 = \|z\|_2$ for any $z \in \mathbb{R}^n$.)

(Hint 2: using Hint 1 and the fact that X has orthonormal columns, show that

$$\|y - X\beta\|_2^2 = \|X^\top y - \beta\|_2^2 + c,$$

where c is a constant, meaning that doesn't depend on β.)

d. If $X \in \mathbb{R}^{n\times p}$ is orthogonal, what are the linear regression coefficients $\hat{\beta}$ of y on X? Given your answers for the ridge regression and lasso coefficients in the previous part, give a few sentences interpreting the ridge and lasso coefficients as functions of the linear regression coefficients.

3. 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 files "takeHome.R", "plotfuns.R", and "bstar.Rdata". The first line of the file "takeHome.R" has you install the package glmnet. Once you have done this (i.e., once you have installed this package), you can comment this line out.

We begin with a true signal bstar. Although this is stored as a vector of length $p = 2500$, bstar really represents an image of dimension 50×50. You can plot it by calling

```
plot.image(bstar).
```

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 $N(0, 1)$. These measurements are given by the entries of the vector

```
X %* % bstar
```

in our R 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 R code, we observe

```
y = X %*% bstar + rnorm(n,sd=5).
```

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 bstar? 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 vector bstar is sparse, and the weights in a linear combination are i.i.d normal. This is the idea behind the field of *compressed sensing*. Below, you can find several clips regarding the history, motivation and applications of compressed sensing:

- Robust Compressed Sensing: How Undersampling Introduces Noise and What We Can Do About It (minutes 2-16). https://www.youtube.com/watch?v=ThiAk_n-8HI
- Compressed Sensing: Recovery, Algorithms, and Analysis (first 4 minutes). https://www.youtube.com/watch?v=mgCIKnMgBmk
- Compressive Sensing (minutes 5-16). https://www.youtube.com/watch?v=RvMgVv-xZhQ

The file "takeHome.R" 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 R in the future, etc.

a. Fill in the missing parts. There are 3 missing parts marked by # TODO. When you're getting started, it might be helpful to read the documentation for the glmnet function, which you will use to perform ridge regression and the lasso.

b. Plot the cross-validation error curves for each of ridge regression and the lasso. You can do this using the function plot.cv, as demonstrated by the code at the end. For both ridge regression and the lasso, what value of λ is chosen by the usual rule? What

value is chosen by the one standard error rule? Which method, ridge regression or the lasso, has a smaller minimum cross-validation error?

c. Now run ridge regression and the lasso on the entire data set X,y, for the same tuning parameter values as you did before. Save the objects returned by glmnet as a.rid, a.las, respectively. Plot the coefficient images corresponding to the values of λ chosen by the usual rule and the one standard error rule, for each of ridge regression and the lasso. For this, you'll want to use the indices that you computed in parts (a) and (b), i1.rid, i1.las (usual rule) and i2.rid, i2.las (one standard error rule), as well as the coefficients a.rid$beta, a.las$beta that you just computed. What is the difference between the usual rule and the one standard error rule for the lasso? Which image looks better? What is the difference between the ridge regression images and the lasso images? Which do you think matches the true image bstar more closely?

d. Look at the squared error between the ridge regression and the lasso coefficients that you computed in (c), for both the estimate chosen by cross-validation and that from the one standard error rule, and the true coefficient vector bstar. What has the lowest squared error?

4. (Bonus) Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^\top x_i)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \beta^\top \tilde{x}_i)^2$, prove that

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})], \tag{9}$$

where the expectations are over all that is random in each expression.