Stat 154 - Fall 2017, Midterm 1

October 20, 2017

Name: _____, **Lab:** _____

- There are 6 problems with a total of 40 points, and 1 problem with 2.5 extra credit pts.
- The exam is closed book and will be 50 minutes long (not applicable to DSP students).
- You must not discuss the exam with anyone who is scheduled to take the exam at a later time.
- We reserve the right to NOT answer questions about the problems. So make your best guess
- Keep your eyes on your paper. If we see you looking at someone else's paper, we'll take that as cheating.
- Please be quiet. If we see you moving your lips, we'll take that as cheating.
- The exam will be collected promptly at 2:00pm. Please stop when time is called.
- The test begins on the next page. Good luck!

"I WILL ABIDE BY THE CAL HONOR CODE." Signature: _____

Problem	Maximum score	Score achieved
1	10	
2	5	
3	8	
4	8	
5	5	
6	4	
Total	40	
Extra	*2.5*	

Problem 1 [10 pts]

Indicate whether each of the following statements is TRUE or FALSE.

a) When the matrix $(\mathbf{X}^\mathsf{T}\mathbf{X})$ is not invertible, the least squares estimator exists and it is not unique.

TRUE

b) In Principal Components Regression (PCR), the first principal component $\mathbf{z_1}$ is such that the inner product with the response $\mathbf{z_1^\mathsf{T}y}$ is maximum

FALSE

c) If an $n \times p$ matrix $\mathbf{X}$ has full column rank p, fitting a PCR with all the components gives coefficient estimates equivalent to fitting an ordinary least squares regression.

TRUE

d) The optimal penalty parameter λ of Ridge Regression can be obtained with an analytical equation.

FALSE

e) When the data matrix $\mathbf{X}$ has more columns than rows, PCA via EVD of $\mathbf{X}\mathbf{X}^\mathsf{T}$ is computationally preferable than PCA via EVD of $\mathbf{X}^\mathsf{T}\mathbf{X}$

TRUE

f) In Lasso, when the penalty parameter λ is (sufficiently) large, some coefficients may be set exactly to zero.

TRUE

g) In Partial Least Squares Regression (PLSR), the first PLS component $\mathbf{z_1}$ is such that the inner product with the response $\mathbf{z_1^\mathsf{T}y}$ is maximum.

Either TRUE and FALSE are ok.

However, the actual answer for this answer should be FALSE, because $\mathbf{z_1}$ is obtained by maximizing the squared covariance with the response. However, APM gives an explanation that gives the impression that PLS maximizes the covariance.

h) The matrices $(\mathbf{X}^\mathsf{T}\mathbf{X})$ and $(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})$ have the same eigenvectors.

TRUE

i) The Ordinary Least Squares solution is scale invariant.

TRUE

j) If an $n \times p$ matrix $\mathbf{X}$ has column rank $r < p$, fitting a PLSR with r components gives coefficient estimates equivalent to fitting a PCR with r components.

TRUE

Problem 2 [5 pts]

One of your friends has a data set with $n = 100$ rows, and $p = 15$ predictors. In order to study the predictive performance of a least squares model, your friend has devised the following cross validation scheme:

1. Divide the training data set in 10 folds
2. For each fold:

   2.1 Use the fold to fit a model

   2.2 Use the left 9-folds to compute the fold-MSE
3. Average the 10 fold-MSE's to obtain a CV-MSE

Your friend wants to know your opinion. Comment on your friend's CV scheme. If you had to propose an alternative CV procedure, what would you suggest and why?

We expect that you comment on the following issues:

With 100 rows and 15 columns, each fold will be formed by 10 rows and 15 variables. Trying to fit a LS model won't work: there are more columns than rows.

In addition, using 10% of the data to fit a model while leaving the rest 90% to measure the performance will result in poor predictive measures with high MSEs.

We also expect that you suggest a more appropriate CV-scheme:

It would be better to use normal 10-fold CV, in which each fold is held-out while using the rest of observations to fit a model.

Likewise, with the size of the data (100 observations), you can also try to perform leave-one-out CV.

Problem 3 [8 pts]

a) [2] Compare and contrast PCR and PLSR. Discuss briefly similarities and differences.

Both PCR and PLSR are dimension reduction regression methods. Also, they both involve forming a set of components as linear combinations of the predictors. Instead of directly regressing the response on the predictors, we regress the response on the extracted components.

The main difference between PCR and PLSR relies in the way each method extracts the components. In PCR components are extracted in such a way that they maximize the variance in the predictors, without taking into account the variability of the response.

In contrast, PLS components are obtained by reaching a compromise between summarizing some variation of the predictors, while at the same time taking into account the covariance with the response.

b) [2] Why methods such as ridge regression, lasso, or elastic net are referred to as penalized methods?

Simple answer: These methods are known as penalized regression methods because the minimization criterion includes a penalty term, or *shrinkage penalty* that is used to control the size of the coefficients.

Alternative answer: the estimators obtained by these methods are the solutions of a penalized least-squares problem. We not only minimize the RSS (residual sum of squares), but also there is constraint for the size of the coefficients.

$$min\{\text{RSS} + \lambda\, norm(\boldsymbol{\beta})\}$$

The tuning parameter serves to control the size of the coefficients.

c) [2] Why is recommended to standardize data as a preprocessing step during a Principal Components Analysis (PCA)?

The general advise when performing a PCA is to standardize variables. This is done in order to harmonize the variability of the data.

In general, most data sets will have variables measured in different units. Standardizing the variables will prevent obtaining results that are dominated by those variables with larger variance magnitudes.

d) [2] Why is the coefficient β_0 excluded from the minimization criterion in Ridge Regression and Lasso?

The shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$ but not to the intercept β_0. This is because we don't want to shrink the intercept, which is simply a measure of the mean value of the response when the predictors are equal to zero.

Problem 4 [8 pts]

Let $\mathbf{X}$ be an $n \times p$ matrix of standardized predictors; and $\mathbf{y}$ be a centered vector for the response variable. Recall the ridge regression estimate $\hat{\boldsymbol{\beta}}_{RR}$ is given by

$$\hat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

The singular value decomposition (SVD) of the predictors matrix is: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$ where:

- $\mathbf{D}$ is a $p \times p$ diagonal matrix whose entries are the singular values of $\mathbf{X}$.
- $\mathbf{U}$ is an $n \times p$ matrix whose columns are an orthonormal basis for the p-dimensional space spanned by the columns of $\mathbf{X}$.
- $\mathbf{V}$ is a $p \times p$ matrix whose columns are an orthonormal basis for the n-dimensional subspace spanned by the rows of $\mathbf{X}$.

Obtain the SVD form of $\hat{\boldsymbol{\beta}}_{RR}$ and $\hat{\mathbf{y}}$ (in Ridge Regression).

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{RR} &= (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \\
&= \left((\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T})^\mathsf{T}\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T} + \lambda\mathbf{I}\right)^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T})^\mathsf{T}\mathbf{y} \\
&= (\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T} + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^\mathsf{T} + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \quad \text{(simplified form)} \\
&= \mathbf{V}(\mathbf{\Lambda} + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \quad \text{(this step is optional)}
\end{aligned}$$

where $\mathbf{\Lambda}$ is the matrix of eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}$

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= \mathbf{U}\mathbf{D}(\mathbf{\Lambda} + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y}
\end{aligned}$$

- Give 2 pts for pluging-in the SVD form of $\mathbf{X}$
- Give 4 pts for simplification and no algberaic mistakes in $\hat{\boldsymbol{\beta}}_{RR}$
- Give 2 pts for correct and simplified SVD form of $\hat{\mathbf{y}}$

Problem 5 [5 pts]

Figure 6.6 in the textbook (ISL) shows the results of fitting lasso regression of *Balance* on several predictors using the `Credit` data set. Describe *briefly* what we observe in each panel separately as a result of varying the tuning parameter.

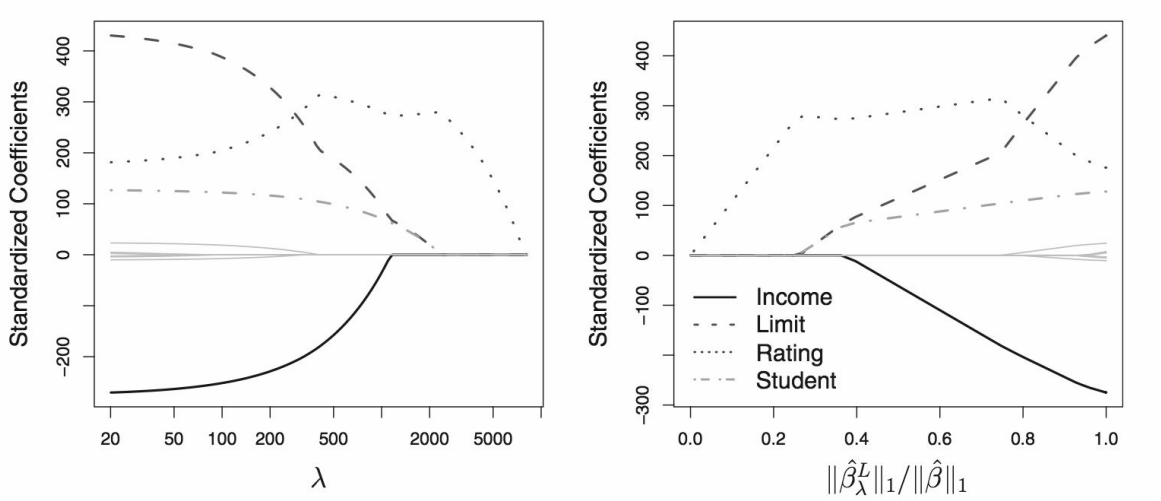

FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.

- Give 2.5 pts for each correct/sufficient description of the figure.

When $\lambda = 0$, then the lasso simply gives the least squares fit, and when λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates equal zero.

Moving from left to right in the right-hand panel of Figure 6.6, we observe that at first the lasso results in a model that contains only the rating predictor. Then *Student* and *Limit* enter the model almost simultaneously, shortly followed by *Income*. Eventually, the remaining variables enter the model. Hence, depending on the value of λ, the lasso can produce a model involving any number of variables.

The right-hand panel of Figure 6.6 displays the same lasso coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, it now displays $\|\hat{\beta}_\lambda^L\|_2/\|\hat{\beta}\|_2$, where $\hat{\beta}$ denotes the vector of least squares coefficients.

As λ increases, the L_1 norm of $\hat{\beta}_\lambda^L$ will always decrease, and so will $\|\hat{\beta}_\lambda^L\|_2/\|\hat{\beta}\|_2$. The latter quantity ranges from 1 (when $\lambda = 1$) to 0 (when $\lambda = \infty$)

Problem 6 [4 pts]

Consider 2 functions: `my_lasso()`, and `my_predict()`.

You can invoke `my_lasso()` as:

```
lasso_fit <- my_lasso(Preds, resp, lambda = 0.05)
```

where `Preds` is a matrix of predictors, `resp` is a vector of response, and `lambda` is a given λ value. This function returns an object of class `"lasso"` which is basically a list with:

- `$coefficients`: vector of coefficient estimates
- `$fitted.values`: vector of predicted values

The function `my_predict()` is invoked as follows:

```
y_pred <- predict(obj, newx = Preds[new,])
```

where the first argument is an object of class `"lasso"`, and `newx` is a matrix of predictors with new observations. The output is a vector of predicted values.

Assume that you have the following objects:

- `X` = a matrix of standardized predictors.
- `y` = a vector of standardized response.
- `training` = a vector of indices for the observations of the training set (to be used in the model building process).
- `test` = a vector of indices for the observations of the test set (not to be used in the model building process).
- `best_lam` = a number for the best λ obtained from 10-fold CV on the training set.

Use the provided functions and objects to write R code that computes the training MSE, and the test MSE, associated to the best λ `best_lam`.

```
lasso_fit <- my_lasso(X[training, ], y[training, ], lambda = best_lam)

# training MSE
train_pred <- my_predict(lasso_fit, newx = X[training, ])
mean((y[training, ] - train_pred)^2)

# (equivalently)
# mean((y[training, ] - lasso_fit$fitted.values)^2)


# test MSE
test_pred <- my_predict(lasso_fit, newx = X[test, ])
mean((y[test, ] - test_pred)^2)
```

- 1 point for fitting a model with training data
- 1.5 points for correctcomputation of training MSE

- 1.5 points for correct computation of test MSE
- Other point adjustments listed in gradescope

Problem 7: Extra Credit [2.5 pts]

Explain the panels of Figure 6.7 from the textbook (ISL).

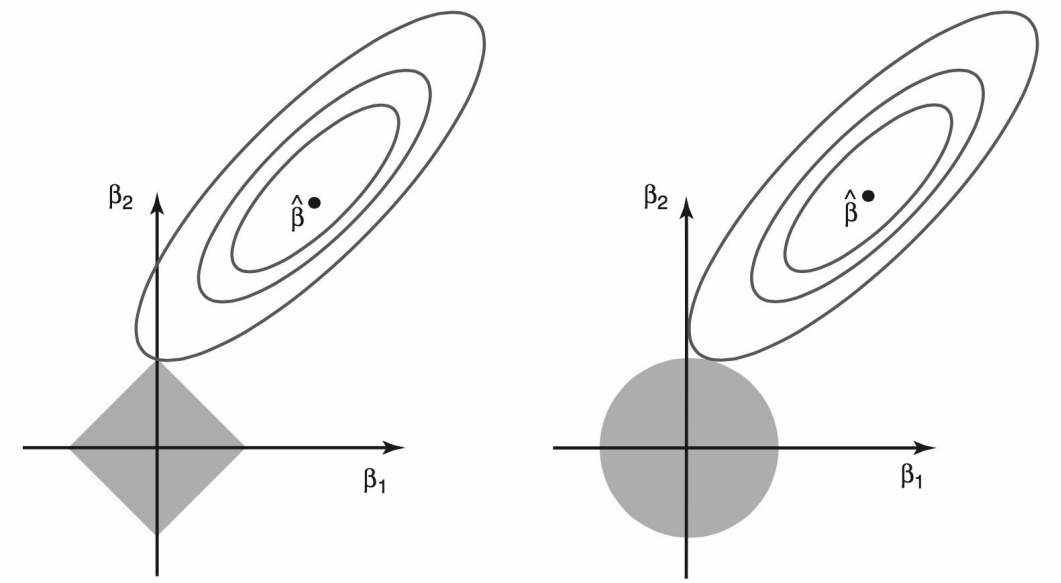

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right).

We are looking for descriptions of each panel mentioning:

- the shape of the regions corresponding to the different constraint norms: L_2-norm for ridge, and L_1-norm for lasso.
- the form of the solution, that is, the point at which an ellipse contacts the constraint region: lasso may set some coefficients to zero (i.e. sparse solutions) while the ridge solution typically includes all coefficient estimates (i.e. dense solution that uses all variables).