FSRM 588: Financial Data Mining Homework 2: Chapters 7 and 14

Fall 2015

Due 6:40pm on October 12

NOTES: Submit all your code with your assignment. Homework must be neatly written-up or typed for submission. Please read and follow the instructions carefully. I reserve the right to refuse homework that is deemed (by me) to be excessively messy.

Instructions:

- HW2 is intended to complement the materials in Chapters 7 and 14 and help you familiarize with all of the algorithms introduced in class.
- To this end, HW2 contains two parts: Part I is theoretical and Part II is computer exercises. Your answers to Part II MUST be typed up and printed out but your answers to Part I can be either typed up or written up. Submit the answers to both Part I and Part II in class.
- The purpose of most exercises in Part II is to perform CV for all of the methods we have learned so far. Since CV involves randomization, your answer may vary every time you run the code. Therefore, it is generally a good idea to set a random seed when performing an analysis such as cross-validation that contains an element of randomness, so that the results obtained can be reproduced precisely at a later time.
  - *Use set.seed(1) every time you start a new CV.*
  - *Throughout this exercise, we consider K-fold CV with $K = 5$.*
  - *As we explained in class, it did not make sense to standardize the predictors, because they were already on the same scale.*
  - *Run "apply(x.train,2,var)". The output implies that V16 has almost zero variance and should be excluded from the model.*
- Submit the R code you use to the email address fsrm588@gmail.com as an attachment before class. The subject of the email should be "hw2 + your last name + your first name". For instance, "hw2 Yang Dan". Only .r files are allowed and the title of the attached document should be "hw2 + your last name + your first name.r". For instance, "hw2YangDan.r"
- The content of the email can be empty. There is no need to say thank me and I won't reply to the emails sent to fsrm588@gmail.com. fsrm588@gmail.com is only used for the purpose of submission of homework, project, and exam. Please reach me via dyang@stat.rutgers.edu if you have questions or need help.

PART I: Theory

1. *Effective degrees of freedom.* Suppose that the training data consists of independent (output, input) pairs $(y_1, x_1), ..., (y_N, x_N)$. Additionally assume that $y_i = f(x_i) + \epsilon_i$, where x_i and ϵ_i are independent, $f(x_i)$ is some regression function, and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Let $\mathbf{y} = (y_1, ..., y_N)^T \in \mathbb{R}^N$. Suppose that you derive a prediction rule $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is an $n \times n$ matrix that depends on the data only through the inputs $x_1, ..., x_N$. Show that

$$\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = \sigma^2 \mathrm{tr}(\mathbf{S}),$$

where the Cov's above are computed conditional on the inputs $x_1, ..., x_N$.

2. *AIC with unknown variance.* Consider a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with Gaussian errors $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)$. Let $\mathbf{x}_j$ denote the j-th column of $\mathbf{X}$, so that $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_p)$. Assume that the parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$ are unknown. Let $\alpha \subseteq \{1, ..., p\}$ be a subset that indicates the predictors to be included in a regression model (e.g. if $\alpha = \{1, 3, 4\}$, then $\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4$ would be included in the model) and define the corresponding parameter space $\Theta_\alpha = \{\boldsymbol{\beta} \in \mathbb{R}^p;\ \mathbf{b}_j = 0 \text{ if } j \notin \alpha\}$ (i.e. Θ_α is the collection of $\boldsymbol{\beta}$'s whose only nonzero coordinates correspond to elements of α). In this setting, the AIC criteria is defined by

$$\mathrm{AIC}(\alpha) = -\frac{2}{N}\left[\sup_{\sigma^2 > 0, \boldsymbol{\beta} \in \Theta_\alpha} \log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})\right] + \frac{2d(\alpha)}{N},$$

where $d(\alpha)$ is the size of α and

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2\right\}$$

is the likelihood of the data. Prove that minimizing $\mathrm{AIC}(\alpha)$ is equivalent to minimizing

$$\log\left(\frac{1}{N}||\mathbf{y} - \mathbf{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha||_2^2\right) + \frac{2d(\alpha)}{N},$$

where $\mathbf{X}_\alpha = (\mathbf{x}_j)_{j \in \alpha}$ is the $N \times d(\alpha)$ matrix whose columns are given by the columns of $\mathbf{X}$ that correspond to elements of α and $\hat{\boldsymbol{\beta}}_\alpha = (\mathbf{X}_\alpha^T \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^T \mathbf{y}$.

3. *Simple algebra.* Verify that

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{j=1}^p d_j \mathbf{u}_j \mathbf{v}_j^T$$

$$S := \frac{1}{N}\mathbf{X}^T\mathbf{X} = \frac{1}{N}\sum_{i=1}^N x_i x_i^T$$

where x_i is the ith row of $\mathbf{X}$. And furthermore,

$$\mathbb{E}S = \Sigma,$$

where Σ is the population covariance matrix assuming that X has mean zero.

4. *SVD.* First show that if we optimize the following objective functions partially for μ and λ_i,

$$\min_{\mu, \lambda_i, \mathbf{V}_q} \sum_{i=1}^N \|x_i - \mu - \mathbf{V}_q \lambda_i\|_2^2,$$

we obtain

$$\hat{\lambda}_i = \mathbf{V}_q^T (x_i - \bar{x})$$
$$\hat{\mu} = \bar{x}.$$

Secondly, show that $\hat{\mu}$ is not unique, and characterize the family of equivalent solutions.
Lastly, suppose centered $\mathbf{X}$ has SVD $\mathbf{U}\mathbf{D}\mathbf{V}^T$, find the relationship between $\mathbf{V}_q$ and $\mathbf{V}$ and the relationship between λ_i and $\mathbf{U}, \mathbf{V}, \mathbf{D}$

5. *PC regression.* Consider PC regression with the leading q principal components and suppose the design matrix $\mathbf{X}$ has SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Show that

$$\hat{\mathbf{y}}^{PC} = \sum_{j=1}^q \mathbf{u}_j \mathbf{u}_j^T \mathbf{y}$$

and it has the following bias and variance

$$\begin{aligned}
\|\mathbb{E}\hat{\mathbf{y}}^{PC} - \mathbb{E}\mathbf{y}\|_2^2 &= \sum_{j=q+1}^{p} (\mathbf{u}_j^T \mathbb{E}\mathbf{y})^2 \\
\mathbb{E}\|\hat{\mathbf{y}}^{PC} - \mathbb{E}\hat{\mathbf{y}}^{PC}\|_2^2 &= \sigma^2 q
\end{aligned}$$

PART II: Computer

Consider the zipcode data, only 2's and 3's. The zipcode data are available from Sakai. Use the following code to load the data: load("hw1.RData"). The names of the variables are self-explanatory. Do the following exercises.

6. *PCA.*

   (a) Perform PCA on the training data for digit 3's, i.e., `train3`. Show the following plots: the second score vs the first score, the image of the mean of 3's, the image of the first loading vector, and the image of the second loading vector. The goal is to reproduce Figure 1.

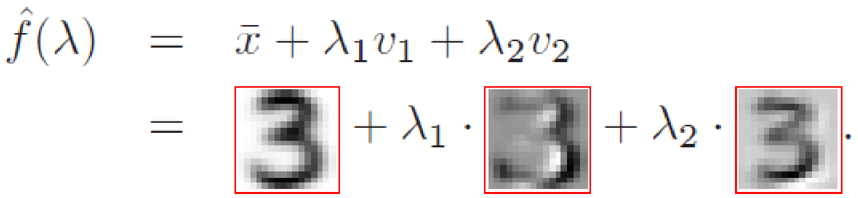
$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$
$= \quad + \lambda_1 \cdot \quad + \lambda_2 \cdot \quad .$

Figure 1:

   What do the leading two loading components mean?

   (b) Repeat 1(a) with digit 2's instead of 3's.

   (c) Combine the training data for both digit 2 and 3 as follows. Perform PCA on `x.train`. Show the following plots and color them according to whether they are digit 2 or 3: the plot of first score, the plot of the second score, and the plot of the second score vs the first score. What do you observe? Show the following plots as well: the image of the mean, the image of the first loading vector, and the image of the second loading vector. How do they make sense to you?

```
x.train <- rbind(train2,train3)
```

7. *CV for OLS*

   (a) Use CV to obtain an estimate of the test error of OLS purely based on the training data via function `cv.glm` without changing the `cost` argument. What is CV error without adjustment?

   (b) Note that the default loss function in `cv.glm` is mean squared error, which is not appropriate for classification, set the cost to be

```
cost = function(y, y.hat) mean((y.hat>.5)!=y)
```

   Redo part (a). What is the CV error now?

   (c) What is the true test error? Is CV from part (b) producing a reasonable estimate?

8. *CV for KNN.* Consider the following choices of K, the number of neighbors, for KNN (Hint: you should change the loss function from the code in class from MSE to misclassification or 0-1 loss again.)

   ```
   klist <- seq(1,21,by=2)
   ```

   (a) Plot the test error, training error, and CV error against K, where K is the number of neighbors *in the same figure.*

   (b) According to the CV error, what is the best K? What are the CV error and test error corresponding to that K? Highlight the point on the plot in (a).

9. *CV for Ridge.* Strictly speaking, the default loss function is MSE if we use the built-in CV function `cv.glmnet` the same fashion as before when we run Ridge regression. To make everything right, one needs to run logistic regression, which we have not learned yet. For now, there are still two choices left for us. (Note that you need to change the default number of folds to 5, `nfolds=5`.)

   (a) Choice 1: ignore the fact that we should use misclassification error when performing CV and bear with what we have (that is equivalent to say we are implementing regression instead of classification). Plot the CV error against λ, the tuning parameter in Ridge using the built-in plot function, what is value of λ that has the smallest cross-validated MSE? If one use that λ, what is the true test misclassification error? What is the best λ according to "One-Standard Error" Rule? If one uses that λ, what is the true test misclassification error?

   (b) Choice 2: Brutal-force way. Write your own CV function (Hint: mimic the spirit from K-Fold CV for Subset Selection in code4.r). Suppose you record the CV result from part (a) and let it be `cv.out`. Obtain potential choices of λ values by `cv.out$lambda`. Run a 5-fold CV on these values and obtain CV errors, the misclassification error as the loss function this time, for all of these values. Plot the CV error against λ. what is value of λ that has the smallest cross-validated misclassification error? If one uses that λ, what are the CV error and the true test misclassification error?

   (c) Note that the computation required in part (a) and (b) should be approximately the same. Why does part(b) take much longer time to run than (a)? Can you guess the reason?

10. *CV for LASSO.* LASSO has the same problem as Ridge regression. Do the following by setting `alpha=1` instead of `alpha=0`.

    (a) Repeat 9(a) for LASSO

    (b) Repeat 9(b) for LASSO

11. *CV for PC regression.* The default CV function for PC regression uses MSE as well. If you can do 9(b) well, the ideas are the same. Get the training, test, and CV errors for PC regression with 1 principal component, 2 components, ..., till 256 components. Plot the errors in one graph. How many components should you choose according to the picture? What are the training and test errors with that many components selected. Highlight the point in the graph.

12. *CV for Subset Selection.* Get the training, test, and CV errors for all of the nested models generated by forward stepwise regression $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$. Plot the errors in one graph. How many predictors should you choose according to the picture? What are the training and test errors with that many predictors selected. Highlight the point in the graph.