FSRM 588: Financial Data Mining Homework 3: Chapter 4

Fall 2015

Due 6:40pm on October 26

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:

- HW3 is intended to complement the materials in Chapter 4 and help you familiarize with all of the algorithms introduced in class.
- To this end, HW3 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.
- Use set.seed(1) every time you start a new CV. We consider K-fold CV with $K = 5$.
- 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 "hw3 + your last name + your first name". For instance, "hw3 Yang Dan". Only .r files are allowed and the title of the attached document should be "hw3 + your last name + your first name.r". For instance, "hw3YangDan.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. *Linear regression for classification.* Consider the linear regression on indicator matrix method, which is used for classification. $\mathbf{Y}$ is the $N \times K$ indicator response matrix, $\mathbf{X}$ is the $N \times (p+1)$ design matrix where the first column are all 1's and the rest columns are all centered (zero mean), and $\mathbf{B}$ is the $(p+1) \times K$ coefficients matrix. The estimate we obtain for $\mathbf{B}$ is

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Using this estimate, we can get the following estimate for any new input $x \in \mathbb{R}^p$

$$\hat{\mathbf{f}}(x) = \hat{\mathbf{B}}^T \begin{pmatrix} 1 \\ x \end{pmatrix} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}$$

Show that

$$\sum_{k=1}^{K} \hat{f}_k(x) = 1.$$

Hint, write the design matrix as $\mathbf{X} = (\mathbf{1}_N \ \ \widetilde{\mathbf{X}})$. Note that all of the p predictors are centered implies that

$$\widetilde{\mathbf{X}}^T\mathbf{1}_N = \mathbf{0}_p,$$

and that the fact that the sum of each row of $\mathbf{Y}$ is one means

$$\mathbf{Y}\mathbf{1}_K = \mathbf{1}_N.$$

And, finally $\sum_{k=1}^K \hat{f}_k(x) = \mathbf{1}_K^T \hat{\mathbf{f}}(x)$.

2. *The similarity of linear regression and LDA.*

   (a) Consider the minimization of the least squares criterion

   $$\|\mathbf{y} - \beta_0 \mathbf{1}_N - \mathbf{X}\beta\|_2^2$$

   where $\mathbf{X}$ is the $N \times p$ design matrix and $\mathbf{y}$ is a vector of length N with {0,1} entries with $y_i = 1$ if $g_i = 2$ and $y_i = 0$ if $g_i = 1$. Show that the solution $\hat{\beta}$ satisfies

   $$((N-2)\hat{\Sigma} + N\hat{\Sigma}_B)\beta = \frac{N_1 N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)$$

   after simplification, where $\hat{\Sigma}_B = \frac{N_1 N_2}{N^2}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T$ and $\hat{\Sigma} = (\sum_{g_i=1}(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T + \sum_{g_i=2}(x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^T)/(N-2)$ is what we defined in class.

   (b) Hence show that $\hat{\Sigma}_B\beta$ is in the direction $\hat{\mu}_2 - \hat{\mu}_1$ and thus

   $$\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1).$$

   Combining with the fact that LDA for binary classification classifies to class 2 if

   $$x^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^T\hat{\Sigma}^{-1}\hat{\mu}_1 + \log(N_1/N_2),$$

   we have proved that the least-square regression coefficient is identical to the LDA coefficient up to a scaler multiple.

3. Start thinking about the project.

PART II: Computer

4. *Logistic regression and linear discriminant analysis.* In this problem, you will compare the performance of logistic regression and linear discriminant analysis in simulation experiments. Suppose that the input $X \in \mathbb{R}^p$ and output $G \in \{0, 1\}$ are distributed so that

   $$\begin{aligned} P(G=1) &= \pi, \\ P(G=0) &= 1-\pi, \\ X|(G=0) &\sim N\{(\Delta/2)e_1, I_p\}, \\ X|(G=1) &\sim N\{-(\Delta/2)e_1, I_p\}, \end{aligned}$$

   where π, Δ, p will be further specified below and $e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^p$ is the vector whose first entry is one and other entries are zero.

   (a) For each combination of $\pi \in \{0.25, 0.5\}$, $\Delta \in \{0, 1, 2, 3\}$, and $p \in \{1, 3, 5, 10\}$, repeat the following steps 1000 times:

   (i) Generate 300 training and 1000 test observations from the distribution specified above.

   (ii) Fit a logistic regression model to the training set (using all of the predictors) and record the estimated coefficients. Using these coefficients, record the misclassification rate on the training set and the misclassification rate on the test data.

(iii) Using the sample means and covariance from the training data, implement linear discriminant analysis and record the misclassification rate on the training set and the misclassification rate on the test set.

(b) For each combination of π, Δ, p specified in part (a), report the average misclassification rate (computed over the 1000 datasets) on the training and test datasets for logistic regression and linear discriminant analysis. This information should be presented clearly, in a table. Comment on your results.

(c) Can you think of some joint distribtion of (X, G) for which logistic regression works *much* better than LDA? Why or why not? Devise some simulations to help illustrate your conclusion.

5. *Classification and dimension reduction via PCA and FDA.* Consider the zipcode data, the 2's, 3's, 8's. The zipcode data are available from Sakai. Use the following code to load the data: load("hw3.RData"). The names of the variables are self-explanatory. Do the following exercises.

   Note:

   - *Do not use the built-in functions from R, such as prcomp or lda; you can still use other functions*
   - *As we explained in class, it did not make sense to make the norms of all predictors the same, because they were already on the same scale.*
   - *There is no need to exclude V16 any more.*

   (a) To avoid confusion, let us first center the data, both the training and the test data. Note that we should use the sample mean from the training data, rather than from the test data to center both the training and the test data. Perform the following by using the centered data.

   (b) Perform PCA on the centered data, retain the leading two scores; no need to report anything.

   (c) Perform FDA on the centered data, retain the leading two scores; no need to report anything.

   (d) Perform OLS, logistic regression, LDA with all of the input variables. What are the training and test errors for these three methods?

   (e) Perform OLS, logistic regression, LDA with the leading two PC scores. What are the training and test errors for these three methods? Plot the first two PC scores and the decision boundaries obtained from these three methods with three different colors.

   (f) Repeat (e) with the first two PC scores replaced by the first two FDA scores

   (g) Compare your results from (d.e.f). What do you find?

   (h) Perform CV on the number of PC scores to use in OLS. Plot the test errors against the number of PC scores retained. How many PCs should you keep? If you keep that many number PCs, what are the training and test errors? Use the following to randomly split the data.

```
nfolds=5
folds <- split(sample(n.train),rep(1:nfolds,length=n.train))
```

   (i) Repeat (h) with OLS replaced by LDA

   (j) Potentially you could repeat (h) with OLS replaced by logistic regression. However, there will be errors. So you do not need to try this out.

   (k) Combine your observations from (f.h.i.j). What do you find?