STAT 542, Homework 3

February 28, 2018

Due date: Mar 16 (Fri), 5 pm to Compass

Requirements: This homework consists of three problems. You should submit your **report and `R` code** as separate files to Compass2g. The report should be in PDF/html/MS Word format, with no more than 10 pages, including plots. Your report should focus on describing the analysis strategy, demonstrating and interpreting the main result. No excessive `R` code should be included in the report. Font size should be 12pt and plots need to be clearly labeled. If you use R markdown and need a template, there is an example source file provided at our course website. This homework worth 100 points total. Late submission penalty is 5 points for each day (round up) of delay.

Question 1: [15 points] Use the hand written digit dataset from the `ElemStatLearn` dataset. The dataset is already separated into train and test. Take only digits 4 and 9 from the dataset, and perform the following:

a. [5 Points] Use ONLY the training data: Fit SVM with different tuning parameters (cost) and three different kernels (your choice), and the parameters (if any) for the kernel. Use cross-validation to select the best combination.

b. [5 Points] Apply your selected model to the testing data and report the performance.

c. [5 Points] Comment on the following: i). What are the advantages and disadvantages of different kernels that you used. ii) Is cross-validation effective on this analysis? Is there any potential pitfall that you could think of?

Question 2: [25 points] Use the Fashion MNIST dataset that we analyzed in HW1. Preform the following:

a. [10 Points] Write your own linear discriminate analysis (LDA) code following our lecture note, and fit the model to the training data. Report the performance of your model on the testing data.

b. [10 Points] Use the regularized QDA method described in page 34 of lecture note "Class":

$$\widehat{\Sigma}_k(\alpha) = \alpha\widehat{\Sigma}_k + (1-\alpha)\Sigma$$

Again, you do not need to tune the α parameter using cross-validation. Directly apply your model to the testing data to select the best tuning.

c. [5 Points] Is there any benefit using the regularized QDA compared with the regularized LDA method? Comment on the advantages and disadvantages of these three methods: LDA, QDA, and regularized QDA

Question 2: [60 points] Install the `quadprog` package and utilize the function `solve.QP` to solve SVM. The `solve.QP` function is trying to perform the minimization problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} - d^{\mathrm{T}}\boldsymbol{\beta} \\ \text{subject to} \quad & \mathbf{A}^{\mathrm{T}}\boldsymbol{\beta} \geq b_0 \end{aligned}$$

For more details, read the document file of the `quadprog` package on CRAN. One difficulty you may have in this question is that the package requires $\mathbf{D}$ to be positive definite, while it may not be true in our problems. A workaround is to add a "ridge", e.g. $10^{-5}\mathbf{I}$, to the matrix, making it invertible. However, by doing this, you will also face the problem of numerical accuracy of the solution, i.e., they may not be exact. Figure out a way to deal with them if you think they cause trouble for your estimation.

a) [25 Points] Generate a set of separable data using the following code:

```
set.seed(1); n <- 40; p <- 2
xpos <- matrix(rnorm(n*p,mean=0,sd=1),n,p)
xneg <- matrix(rnorm(n*p,mean=4,sd=1),n,p)
x <- rbind(xpos,xneg)
y <- matrix(c(rep(1,n),rep(-1,n)))
```

Then formulate the dual problem of the linear separable SVM optimization problem into a form that can be solved by `solve.QP`. You should define explicitly what are $\mathbf{D}$, $\mathbf{A}$, d and b_0. Obtain the support vectors and decision line from your result. Compare your solution to the results produced by `e1071` package. Use plots if necessary.

b) [25 Points] Generate a set of nonseparable data by yourself (preferably in two dimensions and plot them). Formulate the dual form of linear SVM so that it can be solved by `solve.QP`. You should define explicitly what are $\mathbf{D}$, $\mathbf{A}$, d and b_0. To calculate the separation line (especially the intercept term), you may want to read page 421 of the textbook. Plot and compare your results with the `e1071` package. For this question, you should set a **reasonable** C value, however, you are not required to tune it.

c) [5 Points] Recall that our dual form of the problem requires

$$\boldsymbol{\beta} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

If instead of the original x, we use basis expansions $\Phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_m(x))^{\mathrm{T}}$ as the new covariates, derive the new decision function $f(x)$ (using β_0, α's, y's and $\Phi(x)$'s) based on this expansion. After doing this, utilize the relationship between the kernel function and the basis expansions to rewrite $f(x)$ using the kernel functions.

c) [5 Points] After realizing this connection, start from page 39 of the lecture note "SVM", and formulate the dual problem of the linear separable SVM optimization problem into a form that can be solved by `solve.QP`. You should use only the kernel functions, not the basis expansions. You should define explicitly what are $\mathbf{D}$, $\mathbf{A}$, d and b_0. You are not required to solve it, only need to state the problem.