STAT 542, Homework 3

October 12, 2017

Due date: Nov 1 (Wed), 11:59 pm to Compass

Requirements: You should submit your report and *R* code(s), preferably in separate files. Your report should be in PDF/MS Word format. Font size should be 12pt and plots need to be clearly labeled. Your report should include necessary explanations and should not be a simple output file of the *R* code. The *R* code should include comments to help our grading process. There is a 15 page limit to the report. This homework worth 100 points total. Late submission penalty is 5 points for each day (round up) of delay.

Question 1: [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.

a) [10 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 primal of the linear separable SVM optimization problem into a form that can be solved by `solve.QP`. 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) [10 Points] Repeat question a) by using the dual form of linear separable SVM.

c) [20 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`. Figure out how to calculate the separating line. Plot and compare with `e1071`. For this question, you should set a **reasonable** C value, however, you are not required to tune it.

d) [20 Points] Perform analysis of the South Africa Heart Disease data (use `data(SAheart)` after loading the `ElemStatLearn` package) using your SVM code. The goal is to predict variable `chd` using other covariates. You should perform cross validation to select the tuning parameter and compare your results to an existing package.

Question 2: [20 points] In this question, we want to construct the natural cubic spline basis from the cubic spline basis. The truncated power series basis representation for cubic splines with K interior knots $\xi_1, \xi_2, \ldots, \xi_K$ is given by

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$$

To construct the natural cubic spline, we set the following constrains:

$$f(X) \text{ is linear for } X \leq \xi_1 \text{ and } X \geq \xi_K$$

Utilize the constrains (the corresponding derivatives f' and f''), show the following:

1). β_2 and β_3 are both 0

2). $\sum_{k=1}^{K} \theta_k = 0$

3). $\sum_{k=1}^{K} \theta_k \xi_k = 0$

With these results established, show that the power series representation can be rewrote as

$$f(X) = \beta_0 + \beta_1 X + \sum_{k=1}^{K-2} \alpha_k (d_k(X) - d_{K-1}(X)),$$

where

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k} \quad \text{and} \quad \alpha_k = \theta_k(\xi_K - \xi_k).$$

Your report should contain a rigorous proof with clear logic.

Question 3: [20 points] Download the "Mass Shootings Dataset Ver 2.csv" file from Kaggle:

https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years/data

There are many features in this dataset and there is no clear defined outcome, however, several questions were proposed (see the "Inspiration" section on the webpage). You are free to choose one of these questions (must be something that requires model fitting, so the "Visualize " question does not count) or form your own and then use the information in the data to either support or oppose that. Fully demonstrate the evidence in your report. However, you are limited to using SVM or splines (for multiple variables, you can use an additive model with splines). Grading will be based on the following (each consists of 5 points):

- Is the model implementation correct and appropriate
- Are you able to incorporate as much (relevant) information as you can from the dataset
- Does the evidence you found convincingly support your conclusion
- Overall presentation

This question should not take more than 5 pages.