STAT 542, Homework 1

September 5, 2017
Due date: Sep 18, 11:59 pm to Compass

Requirements: This homework consists of 3 problems. You should submit your report and *R* code, preferably in two 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. This homework worth 100 points total. Late submission penalty is 5 points for each day of delay.

Question 1: [40 points] Download the R markdown file from our course website (at the bottom of the google site, or click here). Follow the exact same code (line 22 - 30) to generate X and Y (do not change the random seed), then perform the following tasks. In this question, you are **NOT** allowed to use any additional `R` package (except the "`MASS`" package which is already used for generating the multivariate normal samples).

a) [10 points] Calculate the sample variance-covariance matrix $\widehat{\Sigma}$ of X (using the maximum likelihood estimator, not the unbiased version). Then calculate $\widehat{\Sigma}^{-1/2}$.

b) [15 points] We want to perform a 5-NN estimation at the target point $x = (0.5, 0.5, 0.5, 0.5)^{\mathrm{T}}$. To do this, lets first write a function `mydist <- function(x1, x2)` that will return the Euclidean distance between any two vectors `x1` and `x2`. Calculate the distance from all sample points to the target point x and output the row numbers of the closest 5 subjects. Use their Y values to obtain the 5-NN estimation at the target point.

c) [10 points] Write another function `mydist2 <- function(x1, x2, s)` that returns the Mahalanobis distance between any two vectors `x1` and `x2` using any covariance matrix `s`. Redo the steps in b) to find the 5-NN estimation based on this Mahalanobis distance with `s` = $\widehat{\Sigma}$.

c) [5 points] Which estimator seems to perform better? Can you give any explanation?

Question 2 [40 points]: You already know how to perform kNN on any target point x. Now, perform a simulation study to estimate the degrees of freedom of a k-nearest neighbor method for regression. The degrees of freedom of a fit is defined as $\sum_{i=1}^{n} \text{Cov}(\widehat{y}_i, y_i)/\sigma^2$. You should proceed as follows:

a) [10 points] If we are interested in using $k = 5$, derive the degrees of freedom of this model using the given formula.

b) [20 points] Perform the simulation study:

   - Generate a design matrix $\mathbf{X}$ from independent standard normal distribution with $n = 200$ and $p = 4$. Now, **Fix these X values for the rest of this problem.**
   - Define an appropriate true model $f(X)$ (choose whatever function you want) as the mean of Y.
   - Using your model, generate the response variables for these 200 observations by adding an independent standard normal noise ϵ. Fit 5-nearest neighbor to the data (you can use existing package if you like). Obtain $\widehat{y}_i$'s.
   - To get a good estimate of $\text{Cov}(\widehat{y}_i, y_i)$, you need to perform this experiment multiple times and calculate a sample covariance. Repeat the previous step 20 times to calculate the estimation. Keep in mind that you do not change X values, only re-generate Y's for each run.
   - Compare your estimated degrees of freedom with the theoretical value that you derived in (a).

c) [10 points] Consider the a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and the fitted value from linear regression $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}\mathbf{y}$. For simplicity, lets assume that ϵ_i's are i.i.d. normal with mean 0 and variance σ^2. Recall the alternative definition of the degrees of freedom:

$$\text{df}(\widehat{f}) = \frac{1}{\sigma^2}\text{Trace}\big(\text{Cov}(\widehat{\mathbf{y}}, \mathbf{y})\big)$$

What is the theoretical degrees of freedom for this linear regression?

Question 3 [20 points]: Load the `SAheart` dataset from the `ElemStatLearn` package. Consider kNN model using two variables, `age` and `tobacco` to model the binary outcome `chd` (coronary heart disease). Use 10-fold cross validation to select the best k. Also report your training error and plot the averaged cross-validation error curve for difference choices of k. Note: you can find some examples in our `intro.r` file, but feel free to improve it.