Instructor: L.-P. Chen

SS4850G Assignment #2
(Due April 3 2020)

Please carefully read the following instructions:

- **There are 3 questions in this assignment. 4 marks for each subquestion.**
- **Show all your works in details and provide your code for computations. Also, well summarize your numerical results in each question.**
- **This assignment is an INDIVIDUAL work. Plagiarism will earn ZERO mark.**
- **ONLY R functions/packages mentioned in this course are allowed. Using other functions/packages that are not used in this course will lose marks.**
- **Submit your solutions to the Drop Box in the course site. Delayed submission will lose marks.**

1. **(Wine Data Set)** These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents (including `Alcohol`, `Malic acid`, `Ash`, `Alcalinity of ash`, `Magnesium`, `Total phenols`, `Flavanoids`, `Nonflavanoid phenols`, `Proanthocyanins`, `Color intensity`, `Hue`,`OD280/OD315 of diluted wines`, and `Proline`) found in each of the three types of wines. The sample size is 178. The dataset is available in the course site. The main interest of this dataset is to study multiclassification of the three types of wines. Let $\widehat{y}$ denote the predicted class of observations.

   (a) Use nominal logistic regression in Section 2.3 to examine the multiclassification. The R function is **multinom**. In addition, summarize the confusion table for y and $\widehat{y}$, use macro averaged metrics to evaluate recall, precision, F-measure, and then conduct performance of classification.

   (b) Use the methods in linear discriminant analysis and quadratic discriminant analysis to obtain $\widehat{y}$. In addition, summarize the confusion table for y and $\widehat{y}$, use macro averaged metrics to evaluate recall, precision, F-measure, and then conduct performance of classification.

   (c) Use the support vector machine method to obtain $\widehat{y}$. In addition, summarize the confusion table for y and $\widehat{y}$, use macro averaged metrics to evaluate recall, precision, F-measure, and then conduct performance of classification.

   (d) Summarize your findings in (a)-(c).

2. **(Simulation studies)** Consider the following linear model:

$$y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 - 4\sqrt{\rho}X_5\beta_5 + \epsilon, \tag{1}$$

where $X = (X_1, \cdots, X_p)$ is a p-dimensional vector of covariates and each X_k is generated from $N(0, 1)$. The correlations of all X_k except X_5 are ρ, while X_5 has the correlation $\sqrt{\rho}$ with all other $p - 1$ variables. Suppose that the sample size is $n = 200$.

(a) Show that X_5 is marginally independent of y.

(b) Now, consider $p = 1500$ and generate the artificial data based on model (1) for 1000 repetitions. Specifically, let $\beta_i = 1$ for every $i = 1, \cdots, 5$ and set $\rho = 0.7$. After that, use the SIS and iterated SIS methods to do variable selection and estimate the parameters associated with selected covariates. Finally, summarize the estimator in the following table:

Table 1: Simulation result for (b)

	$\|\Delta\beta\|_1$	$\|\Delta\beta\|_2$	#S	#FN
SIS				
Iterated SIS				

(c) Here we consider the scenario that is different from (b). Let $p = 40$ and $X \sim N(0, \Sigma_X)$ with entry (j, k) in Σ_X being $0.5^{|j-k|}$ for $j, k = 1, \cdots, p$. We generate the artificial data based on (1) for 1000 repetition with $\beta_i = 1$ for every $i = 1, \cdots, 5$. After that, use the lasso, adaptive lasso, and Elastic net (set $\alpha = 0.5$) methods to estimate the parameters. Finally, summarize numerical results in the following table.

Table 2: Simulation result for (c)

	$\|\Delta\beta\|_1$	$\|\Delta\beta\|_2$	#S	#FN
lasso				
adaptive lasso				
Elastic net ($\alpha = 0.5$)				

(d) Summarize your findings for parts (b) and (c), respectively.

Note: Let $\widehat{\beta}$ be the estimator, then $\Delta\beta$ is defined as $\Delta\beta = \widehat{\beta} - \beta$ with the ith component being $\widehat{\beta}_i - \beta_i$. Therefore, $\|\Delta\beta\|_1$ and $\|\Delta\beta\|_2$ are defined as

- $\|\Delta\beta\|_1 = \sum_{i=1}^{p} \left|\widehat{\beta}_i - \beta_i\right|$;
- $\|\Delta\beta\|_2 = \sqrt{\sum_{i=1}^{p} \left(\widehat{\beta}_i - \beta_i\right)^2}$.

3. In the class, we have discussed analysis of graphical models.

   (a) Show that a multivariate normal distribution with mean μ and covariance matrix Σ can be expressed as the Gaussian graphical model.

   (b) Comparisons with glasso and neighbourhood inference:

   Consider the sample size $n = 400$ and the dimensions of the Gaussian random vector $p = 10$ and 50. Please use the methods introduced in the class to generate *Lattice* and *Hub* graphical structures with 10 repetitions. Regarding the glasso method, choose $\rho = 0.001, 0.01$, and 0.1. Then compute the *specificity* (Spe) and *Sensitivity* (Sen). Finally, summarize numerical results in the following table, and provide your comments on these two methods.

Table 3: Numerical results for the estimators of Θ

p	Model	Method	ρ	Estimator of Θ	
				Spe	Sen
10	Lattice	1	0.001		
			0.010		
			0.100		
		2	×		
	Hub	1	0.001		
			0.010		
			0.100		
		2	×		
50	Lattice	1	0.001		
			0.010		
			0.100		
		2	×		
	Hub	1	0.001		
			0.010		
			0.100		
		2	×		

Method 1: glasso Method 2: neighbourhood inference.

Hint: Regarding simulation studies with 1000 repetitions.

In Question 2, you are asked to use simulation studies with 1000 repetitions to estimate the parameters. Specifically, based on the kth artificial data that are independently generated, you are able to obtain the estimator, denoted by $\widehat{\beta}^{(k)}$. As a result, with 1000 repetitions, the final estimator is given by $\widehat{\beta} = \frac{1}{1000} \sum_{k=1}^{1000} \widehat{\beta}^{(k)}$.