Part 1: short questions.

(1) For the classification trees which two loss functions are the most similar

(A) Gini & Entropy

(B) Misclassification & Gini

(C) Entropy & Misclassification

2 marks

(2) Which of the following statements is FALSE?

(A) In bagging trees the individual trees are independent of each other.

(B) Bagging is the method for improving the performance by aggregating the results of weak learners.

(C) In boosting trees the individual weak learners are independent of each other.

(D) In random forest the individual tree is built on a subset of the features.

2 marks

(3) What are the selections for the algorithm parameters that will lead to similar trees being produced by the random forest and by bagging.

2 marks

(4) Think of the random forest method performed on dataset with large number of correlated predictors. The selected m (number of selected variables) should be:

(A) relatively small.

(B) relatively large.

(C) selected by the following role — $m = \sqrt{p}$ for classification and $m = p/3$ for regression problem.

(D) chosen via cross validation.

2 marks

(5) Is the following statement TRUE or FALSE? A regression spline allows nonlinear fitting using linear regression. Explain your answer.

2 marks

(6) Examine the code below and explain what it does.

(A) locfit(y ~ lp(x, h = 0.1))

(B) ns(x, knots = quantile(x, probs = c(0.2, 0.4, 0.6, 0.8)), intercept = TRUE)

(C) bs(x, knots = c(25,40,60))

2 + 2 + 2 = 6 marks

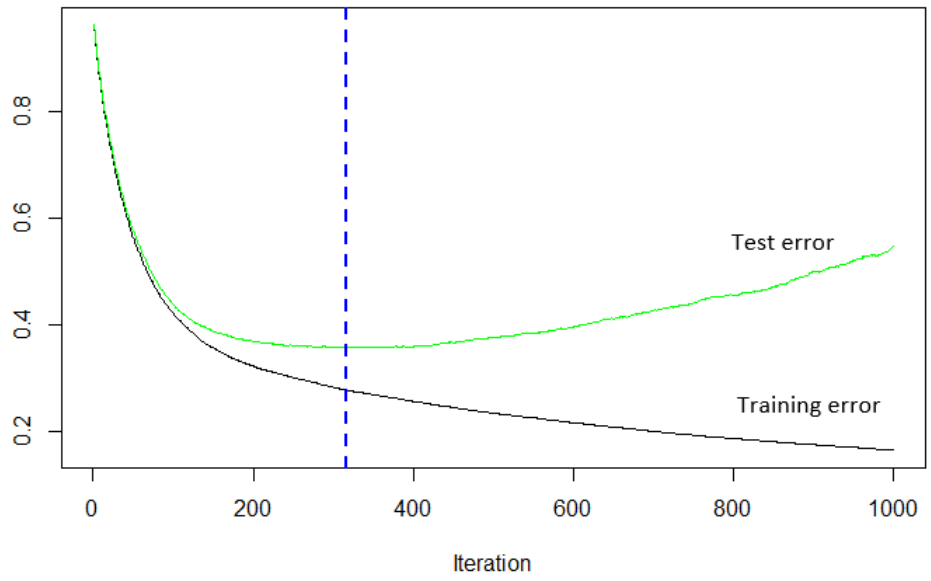
Part 2: long questions.

(7) Daniel wrote the following code.

```
gbmcv <- gbm(Y ~., data=spam[train==1,], cv.folds =5, distribution = "adaboost",
n.trees=1000, interaction.depth=3,shrinkage=0.03, verbose=F)
best.iter <- gbm.perf(gbmcv,method="cv")
```

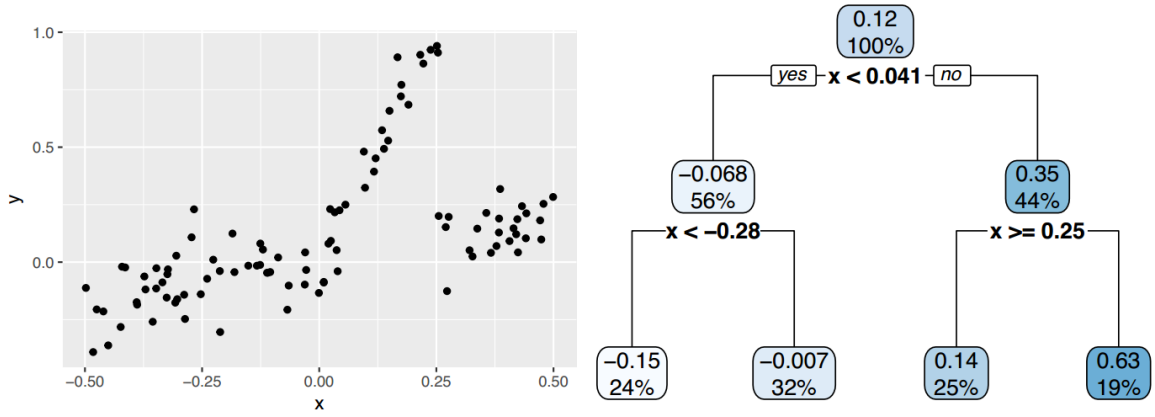
(i) Explain what the code does.

(ii) Based on the following plot, what number of trees should be used? Why?



3 + 2 = 5 marks

(8) Regression trees are fit to data by recursively partitioning it into subsets. Below is a scatter plot of the data set (x, y) and the resulting fitted regression tree.



(i) What value of x does it define the first partition by?

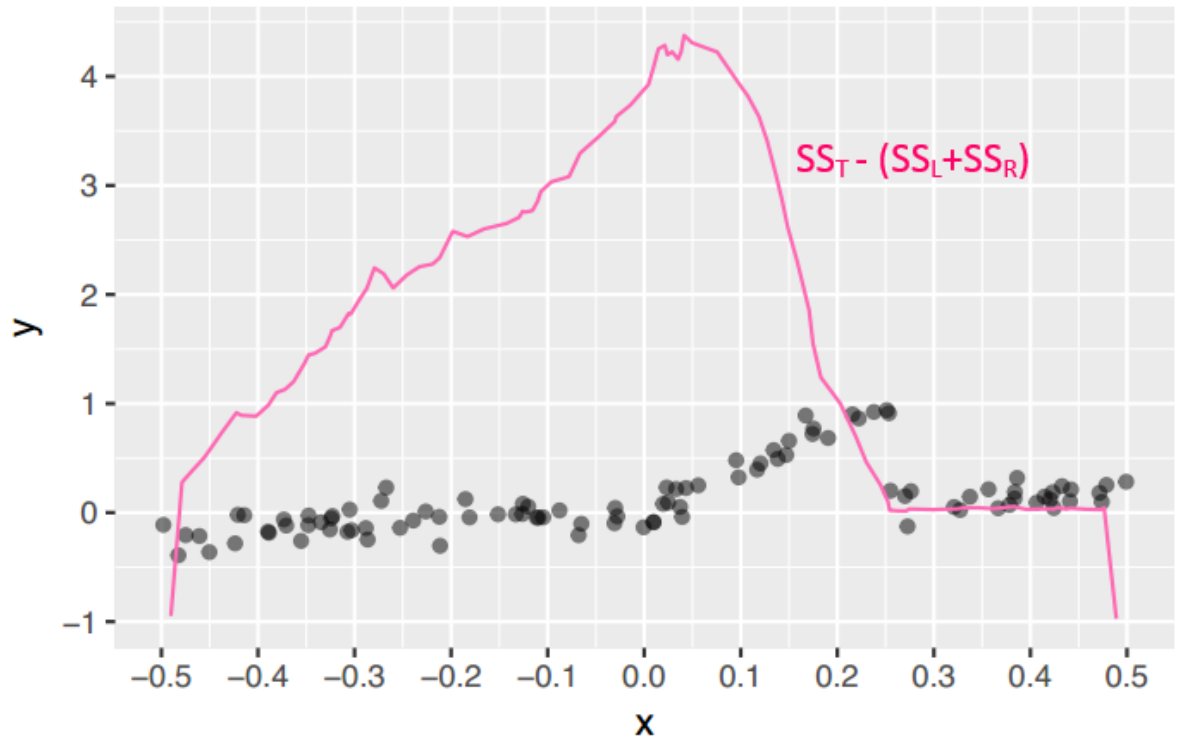
(ii) How many terminal nodes are there in the tree?

(iii) Write down the decisions that need to be followed to obtain fitted values for the model. What would this fitted value be?

(iv) Partitions are decided by optimising the criterion,

$$SS_T - (SS_L + SS_R) \;\; where \;\; SS_T = \sum_{i=1}^{\#\, before\ split} (y_i - \hat{y})^2 \,,$$

where SS_L and SS_R are the equivalent sum of squares for the left and right partition. This is a plot of this partition criterion function, showing the possible criterion values for the first split in order to decide on the best split. (The dots on the plot represent the observations.)



Which value of x corresponds to the optimal value of the function?

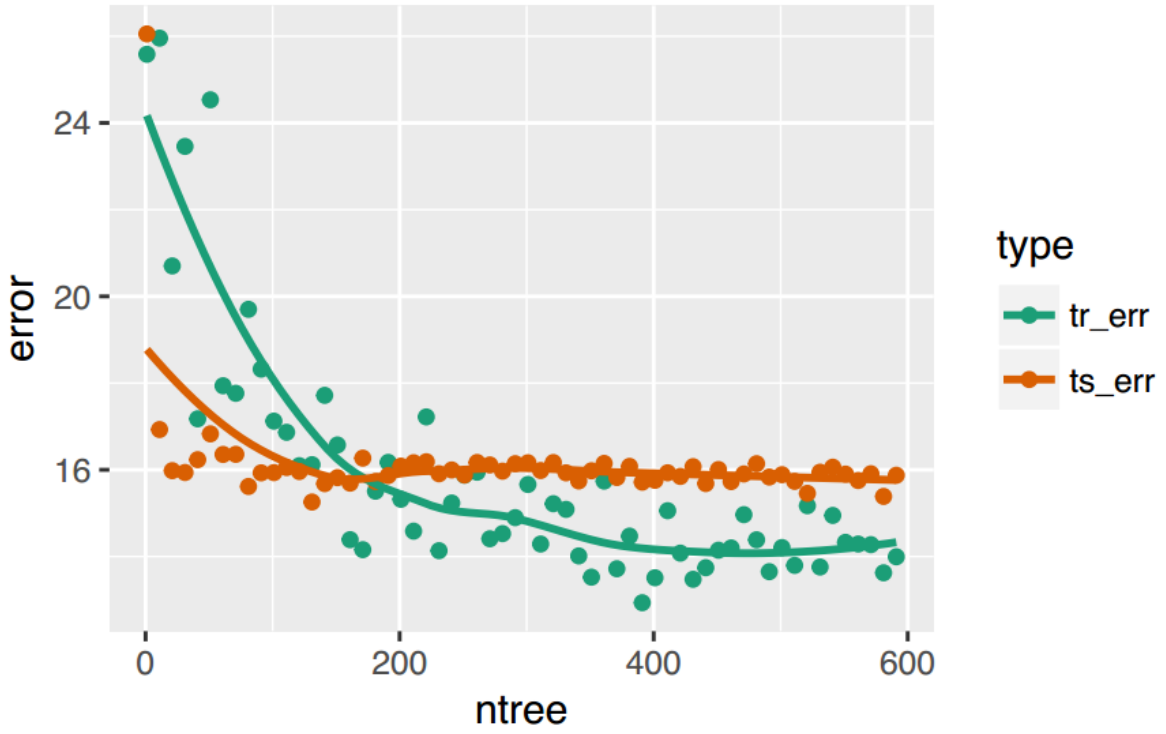
$1 + 1 + 2 + 2 =$ 6 marks

(9) This question is about fitting a random forest model to Melbourne property auction prices.

(i) Explain what a bootstrap sample is.

(ii) Error in the model is computed using "out-of-bag" cases for each tree. What does "out-of-bag" mean?

(iii) The plot below shows the training and the test errors (*tr_err* and *ts_err*, respectively) for forests built on different numbers of trees. Explain why the test error gets higher than the training error as the number of trees increases. What is the recommended size of the forest?



$1 + 2 + 3 =$ 6 marks

(10) This question is about classification trees.

(i) The plot below shows the pruned decision tree for the fgl data set. This data contains information on fragments of glass collected in forensic work (146 rows and 10 columns overall). `type` is glass type (window float glass, WinF or window non-float glass, WinNF) in addition to 8 measurements that are percentages by weight of oxides (`Na, Mg, Al, Si, K, Ca, Ba, Fe`) and a refractive index (`RI`)

Draw the data domain decision diagram corresponding to this tree.

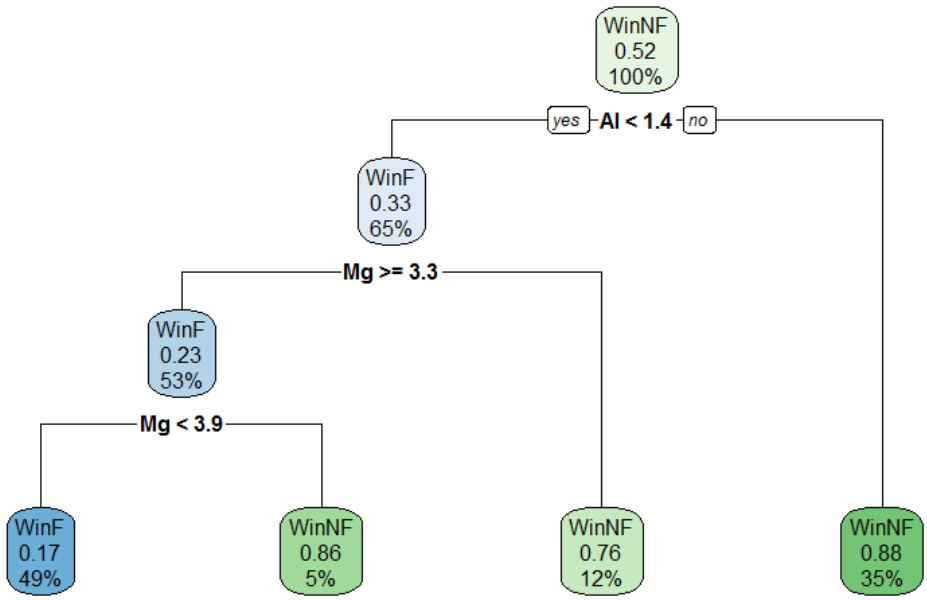
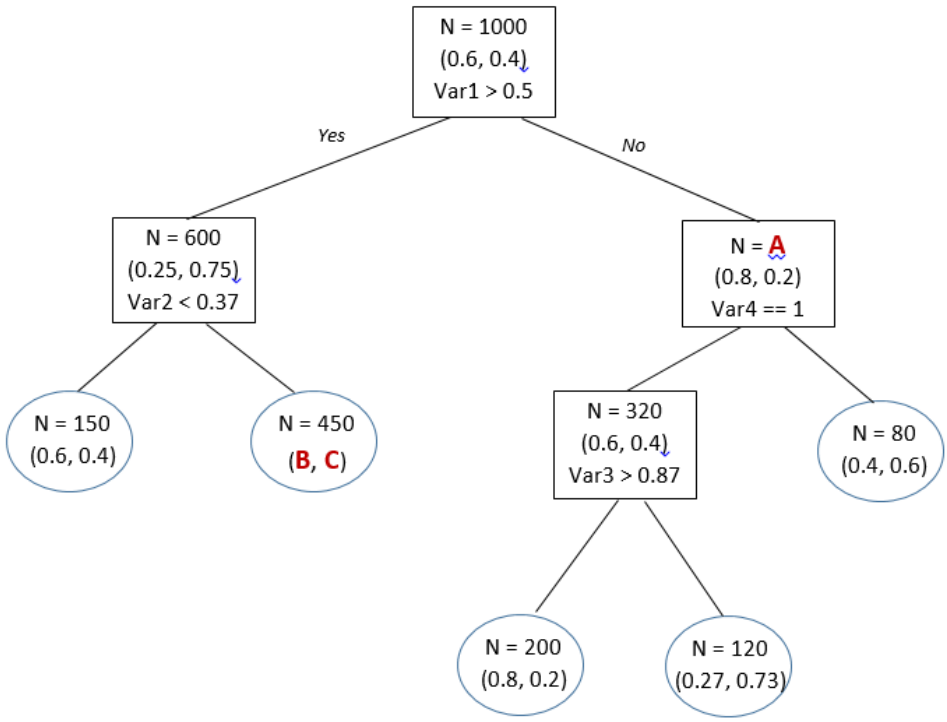


Figure 1: Decision tree (pruned) for fgl data

(ii) Consider the following classification tree, with each node containing the number of observations and the observed class probabilities:



(a) Find the value of A in the tree above.

(b) Find the values of B and C in the tree above.

(c) Suppose we want to prune the tree to have 4 terminal nodes instead of 5. Which nodes should be removed so that the total Gini index loss for the tree is increased as little as possible? (*Note*: Gini index loss is $2\hat{p}(1-\hat{p})$ per observation for each node in a binary classification tree.)

$3 + 7 =$ 10 marks

(11) Let X be a continuous random variable from a half-normal distribution having the pdf

$$p(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}$$

for $x \geq 0$

(i) Suggest an envelope function function $M(x)$ for this problem.

(ii) Describe steps of the acceptance-rejection sampling algorithm relevant to this problem, for generating a random number from $p(x)$.

$4 + 6 =$ 10 marks

(12) Consider a regression model $Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2 + \beta_3 X_2^2 + \varepsilon$. We are interested in the parameter $\theta = \beta_2 + \beta_3$ and we want to use bootstrap to make inference on θ.

Suppose $\mathcal{X} = \{(y_1, x_{11}, x_{21}), \cdots, (y_n, x_{1n}, x_{2n})\}$ are the observed data generated from an unknown cdf F. It is easy to see that a good estimator of θ is $\hat{\theta} = \hat{\beta}_2 + \hat{\beta}_3$ where $\hat{\beta}_2$ and $\hat{\beta}_3$ are the least squares estimators of β_2 and β_3 respectively.

(i) Explain how to use the "Bootstrap the residuals" approach to generate a nonparametric bootstrap sample $\mathcal{X}^*$.

(ii) Suppose $R(\mathcal{X}, F) = \frac{T(\hat{F}) - T(F)}{\sqrt{V(\hat{F})}} = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}}$ is an approximate pivot, with $T(F) = \theta$ and $\hat{F}$ the empirical cdf. Let $R(\mathcal{X}^*, \hat{F}) = \frac{T(\hat{F}^*) - T(\hat{F})}{\sqrt{V(\hat{F}^*)}} = \frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{\text{Var}(\hat{\theta}^*)}}$ be a bootstrap replicate of $R(\mathcal{X}, F)$. Denote by $\hat{G}$ and $\hat{G}^*$ the cdfs of $R(\mathcal{X}, F)$ and $R(\mathcal{X}^*, \hat{F})$ respectively.
Derive a 95% studentized bootstrap confidence interval for θ.

$5 + 5 =$ 10 marks

(13) Observations for a continuous response variable Y and its corresponding predictions from a model $\hat{y}_i = f(x_i)$ are given in the following table.

y_i	7	4.5	6.5	2
$f(x_i)$	8	6.5	4.5	3

(i) Complete the following table:

$f(x_i)$	y_i	CumSum	CumSum%	Data%
8	7			
6.5	4.5			
4.5	6.5			
3	2			

(ii) Sketch the gains chart for the model

(iii) Calculate the associated area under the curve (AUC) score

$3 + 4 + 2 =$ 9 marks

(14) This question is about re-sampling.

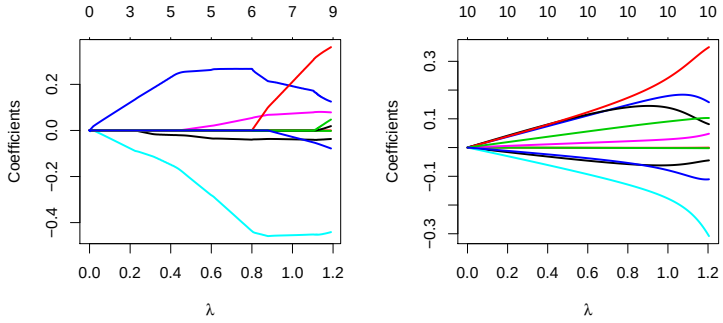
(A) Briefly describe the bias and variance tradeoff when using different k values for the k-fold cross-validation.

(B) If we have $n = 3$ points, $x_1 = 15, x_2 = 36, x_3 = 24$, what is the probability that x_1 does NOT appear in a bootstrap sample?

$3 + 3 =$ 6 marks

(15) This question is about Lasso and Rigde Regression methods.

(A) Explain what is shown in the plots below.
Which one corresponds to the Lasso method and which one to the Ridge method?



(B) Explain why coefficients for some variables may be exactly zero for one of the methods but not for the other.

4 + 4 = 8 marks

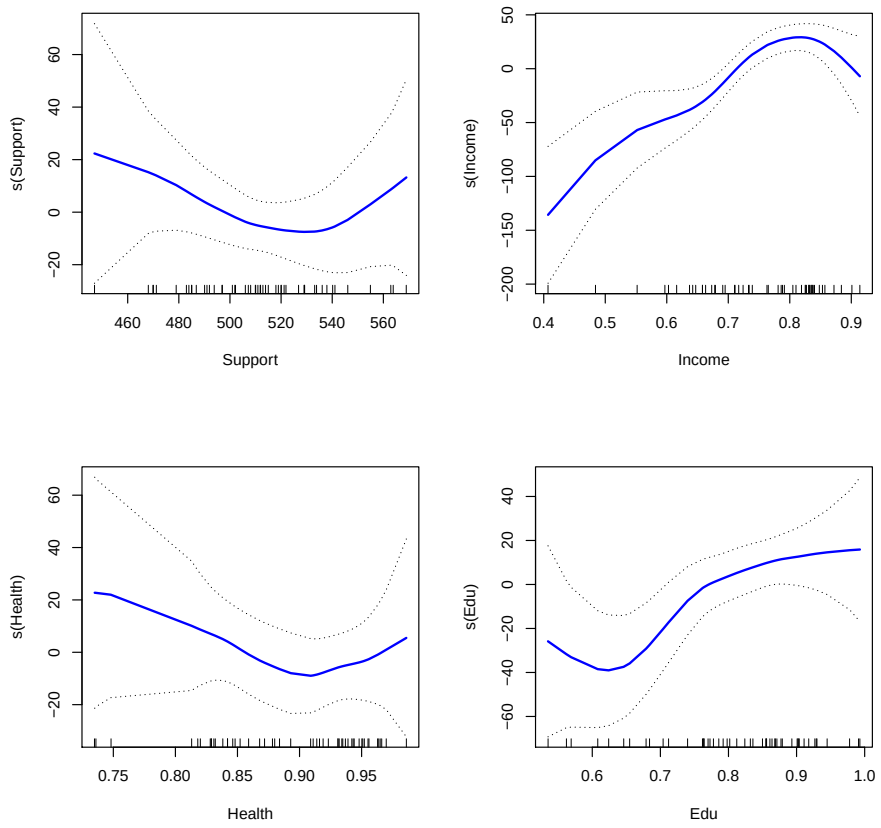
(16) Keren analysed the PISA2006 data using GAM model.

The data set PISA2006 has been constructed using average Science scores by country from the Programme for International Student Assessment (PISA) 2006, along with a few other variables. Each observation is for a country. The key variables that Keren analysed are as follows:

- Overall Science Score - average score for 15 year olds.
- Health Index
- Income Index
- Support for scientific inquiry
- Education Index

Keren ran the following R code:

```
gam.pisa <- gam(Overall ~ s(Support) + s(Income) + s(Health) + s(Edu) , data=pisa)
par(mfrow=c(2,2))
plot(gam.pisa, se=TRUE ,col="blue")
```



(A) Explain what type of splines Keren used in the GAM model.

(B) Based on the plots, explain the marginal influence of each of the explanatory variables on the overall science score.

$2 + 2 =$ 4 marks

(17) Gene XCO-1856 is suspected to be correlated with the Crohn's Disease. To examine this a team of researchers collected a sample of this gene expression from 60 people diagnosed with Crohn's Disease and another 40 people free of Crohn's Disease. Unfortunately, by the time the data become available to a statistician for the analysis, the labels indicating whether the sample came from the person with or without the Crohn's Disease were lost. So the statistician just left with the sample of size $n = 100$, a knowledge of the existence of the two groups but without clear ability to distinguish which record belongs to a person with the disease.

It is reasonable to believe that the distribution of the gene expressions from people with and without the disease can be approximated by Normal distribution. It is also reasonable to believe that the variance for both groups is the same and equals to 1. To test if Gene XCO-1856 correlates with the Crohn's Disease, the means of both normal distributions must be estimated.

This question is about the estimation of the means via the EM algorithm.

(A) What is the vector of the parameters that need to be estimated?

(B) Assuming that z_i is an indicator whether a person has the disease (unobserved), what are the conditional pdf's of Y_i given Z_i?

(C) Write down the observed-data log-likelihood function $\ln L(\theta|\mathbf{y}_n)$ where $\mathbf{y}_n = (y_1, \cdots, y_n)^T$ are observations of $Y_1, \cdots, Y_n$ and $n = 100$.

(D) Write down the complete-data log-likelihood function $\ln L(\theta|\mathbf{y}_n, \mathbf{Z}_n)$ where $\mathbf{Z}_n = (Z_1, \cdots, Z_n)^T$. (You may use the fact that $f(y_i, z_i|\theta) = f(y_i|z_i, \theta)f(z_i)$)

(E) Explain how you will construct the EM algorithm (please answer qualitatively, no formula is needed)

$1 + 1 + 1 + 4 + 3 =$ 10 marks

End of Exam. Total marks = 100