Question 1 **(1+1+1+1+1 = 5 marks)**

Suggest plots that would be appropriate to explore datasets of the following types:

(A) A single continuous variable (e.g. height of a student).
(B) A single categorical variable (e.g. the days of a week).
(C) A single continuous variable (e.g. personal income) and a single categorical variable (e.g. gender).
(D) Two continuous variables (e.g. height and weight of a student).
(E) Two categorical variables (e.g. highest qualification and gender).

Question 2 **(1+1+1+2 = 5 marks)**

(A) Discuss the connections and differences between partitional clustering and hierarchical clustering.

(B) Explain self-organizing map and its "topology preserving properties".

(C) Training self-organizing map needs to specify several parameters. Name three parameters and explain their purpose.

(D) Given a set of points, students A and B apply k-means clustering to cluster these points into M clusters, respectively. Due to various reasons, their clustering results are not identical. You are invited to determine whose clustering result is better. Please describe your solution.

Question 3 **(2+2+2 = 6 marks)**

(A) Describe the main steps of the Apriori algorithm for mining association rules. Explain how the algorithm generates the sets of candidate itemsets and how the algorithm prunes the candidate itemsets.

(B) Consider the following set of items {A, B, D, F, H}. Create a set of transactions such that the association rule {A, D} => {F, H} would have support 0.3 and confidence 0.6.

(C) The measure "confidence" is commonly used to evaluate the interestingness of a mined association rule. However, sometimes a high confidence value does not necessarily mean a rule is indeed interesting. Discuss the potential issue of the measure "confidence" and explain how this issue is addressed in association analysis.

Question 4 **(2+2+2+3 = 9 marks)**

(A) K-nearest neighbour (k-NN) classifier is a simple and effective classifier. Suppose you are given a set of M samples and the class label of each sample is also provided to you. Meanwhile, another set of N samples are hidden from you and they will only be used as a test set to evaluate the k-NN classifier that you have developed. Describe the procedure that you will follow in order to obtain a k-NN classifier that can achieve the highest classification accuracy on the test set (i.e., the N samples).

(B) Given a training set with the following properties:
Number of samples: 1900
Dimension of features in each sample: 45
Dimension of the target value for each sample: 3
Assume that this dataset is being used to train an MLP which has a single hidden layer with 20 neurons, and that the network is being trained for 400 iterations. What is the total number of weights (weight parameters) in this MLP? Show and explain how you derived your answer.

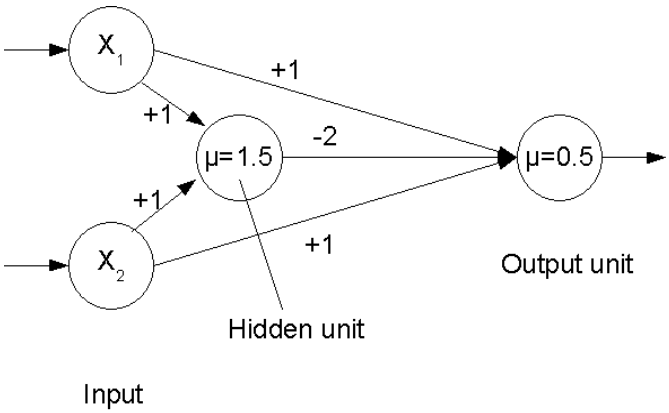
(C) Given a 2-layer MLP as is depicted below. The MLP depicted consists of 1 hidden layer neuron, one output layer neuron, and 5 weights. The value of each of the weights is indicated by a numeric value that is attached to a link (for example, the weight between input x_1 and the output neuron is +1). Assume that the activation function for both, the hidden layer neuron and the output layer neuron is a threshold function defined as

$$f(x) = \begin{cases} 1 & \textbf{if } x > \mu \\ 0 & \textbf{else,} \end{cases}$$

where μ is the threshold value, and x is the sum of all weighted inputs to a given neuron. Thus, for example, if the threshold of a neuron is μ = 0.5, and the sum of its weighted inputs is 0.35 then this neuron will produce 0 as an output.

Given an input set that contains the following four samples:
Sample1: x_1=1.5, x_2=1.2
Sample2: x_1=0, x_2=1.2
Sample3: x_1=0.5, x_2=0.5
Sample4: x_1=1.6, x_2=0

Compute the output produced by this network for each of these samples. You need to show the key steps of calculation.



(D) Using a nonlinear activation function (such as sigmoid, Tanh, and ReLU) in a hidden layer is important for MLP networks to model complex relationship between input and output variables. Prove that for an MLP network with an arbitrary number of hidden layers, if the linear activation function f(x) = x is used for all the neurons in this MLP network, the relationship between the input and output of this network will remain linear.

Question 5 **(2+1+2+1+2 = 8 marks)**

(A) Explain hold-out and 10-fold CV and their strengths and possible weaknesses.
(B) Explain the overfitting problem.
(C) Explain random forest method and how it can improve upon the standard decision tree for regression problems.
(D) State one obvious reason why linear regression is not appropriate for binary/categorical response variable.
(E) Below is the optimisation problem for soft margin classifier

$$\underset{\beta_0,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \; M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right) \geq M\left(1 - \epsilon_i\right)$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

where C is a non-negative tuning parameter.

(i) Explain the role of variable M.
(ii) Explain the relationships between C and the variables $\epsilon_i, i = 1, \ldots, n$, where n is the number of observations in the training dataset.

Question 6 **(2+2+2+1 = 7 marks)**

(A) There is a belief that sleep is important because of its impact on wages through the labour-market productivity. The dataset contains 706 individuals and is a subset of the data used by Biddle and Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy*, 98, 1, pp 922-943. The list of variables is:

sleep	minutes sleep at night per week
totwrk	minutes worked per week
age	age in years
male	=1 if male
educ	years of schooling
kid	=1 if there present children under 3 years of age
inlf	=1 if in labour force

You next use rpart to build a regression tree for these data using the following command:

tree.sleep <-rpart(sleep ~ (totwrk+age+factor(male)+educ+factor(kid)),data=sleep.data)

The resulting regression tree returned by R is the following

```
       n= 706

node), split, n, deviance, yval
      * denotes terminal node

 1) root 706 139239800 3266.356
   2) totwrk>=2256.5 369  70902640 3151.081
     4) totwrk>=3693.5 20   4127513 2747.500 *
     5) totwrk< 3693.5 349  63330890 3174.209
      10) totwrk>=2622.5 174  35897580 3122.690
        20) totwrk< 2784.5 47  11624900 2929.553 *
        21) totwrk>=2784.5 127  21870690 3194.165 *
      11) totwrk< 2622.5 175  26512260 3225.434 *
   3) totwrk< 2256.5 337  58064950 3392.576
     6) educ>=7.5 324  54736770 3375.145
      12) totwrk>=1174 209  32660830 3314.206 *
      13) totwrk< 1174 115  19889250 3485.896 *
     7) educ< 7.5 13    776314 3827.000 *
```

Draw the regression tree corresponding to this output and clearly label the prediction at each leaf node.

(B) Three records in the test dataset are

1. Totwrk: 2400 minutes worked per week.
   Educ: 5 years of schooling
   Male: 1
   Kid: 1
   Age: 30 years of age.
   Sleep: 3220 minutes worked per week

2. Totwrk: 4000 minutes worked per week.
   Educ: 5 years of schooling
   Male: 0
   Kid: 0
   Age: 40 years of age.
   Sleep: 2745 minutes worked per week

3. Totwrk: 1500 minutes worked per week
   Educ: 8 years of schooling
   Male: 1
   Kid: 1
   Age: 25 years of age.
   Sleep: 3320 minutes worked per week.

Compute the predicted sleep per week for each record above using the output of the regression tree in part **(A)** of this question.

(C) Compute the Mean Square Error (MSE) and Mean Absolute Error (MAE) in the test set.

(D) For the support vector machine classification methods, explain clearly the important differences between the support vectors and usual observations.

Question 7 **(2+2+2+2+2 = 10 marks)**

(A) For a given classification task, the confusion matrix from fitting a linear SVM with the training data and predicting the test data (positive = 1, negative = 0) is

	predicted		
actual	0	1	Row Total
0	275	35	310
1	32	230	262
Column Total	307	265	572

The confusion matrix from fitting a nonlinear SVM with radial kernel and scale parameter (positive = 1, negative = 0) $\gamma = 1$ is

	predicted		
actual	0	1	Row Total
0	278	32	310
1	9	253	262
Column Total	287	285	572

The confusion matrix from fitting a nonlinear SVM with radial kernel and scale parameter (positive = 1, negative = 0) $\gamma = 500$ is

	predicted		
actual	0	1	Row Total
0	298	12	310
1	82	180	262
Column Total	380	192	572

Calculate the precision and recall for each classifier and identify which SVM classifier you judge to be the best overall.

(B) For a given decision tree, suppose that in the root node, you have 10 observations belong to class 1 and 10 observations belong to class 2. Then, following the first split, you have

- Left node: 4 observations with class 1 and 3 observations with class 2
- Right node: 6 observations with class 1 and 7 observations with class 2.

Compute the information gain from the first split.

(C) Using the same sleep data in **Question 6(A)**, we now consider the following linear regression model that relates the number of minutes sleeping to the total number of minutes spent on working and other factors that may be affecting sleep.

$$Sleep = \beta_0 + \beta_1 totwrk + \beta_2 age + \beta_3 male + \beta_4 educ + \beta_5 kid + \varepsilon$$

When running the commands lm and summary in R, we obtain the following results.

```
> summary(sleep.linear <-lm (sleep ~ (totwrk+age+factor(male)+
educ+factor(kid)), data= sleep.data))

Call:
lm(formula = sleep ~ (totwrk + age + factor(male) + educ + fac
tor(kid)),
    data = sleep.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2349.05  -241.44     5.82   265.69  1345.44

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   3640.23376  114.33203  31.839   <2e-16 ***
totwrk          -0.16569    0.01801  -9.202   <2e-16 ***
age              2.00994    1.52083   1.322   0.1867
factor(male)1   87.54557   34.66501   2.525   0.0118 *
educ           -11.76532    5.87132  -2.004   0.0455 *
factor(kid)1     4.78424   50.01991   0.096   0.9238
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 418 on 700 degrees of freedom
Multiple R-squared:  0.1216, Adjusted R-squared:  0.1153
F-statistic: 19.38 on 5 and 700 DF,  p-value: < 2.2e-16
```

Using a 10% significance level, decide whether the predictor totwrk helps explaining how long people sleep.

(D) Is there evidence that there are some differences in the minutes sleep at night per week between men and women? Justify your results using significance levels of 0.05.

(E) Compute the predicted sleep per week for each record in the test set in **Question 6(B)** using the output of the linear regression in **Question 7(C)** and compute the MAE and MSE for the test set. Does linear regression perform better than regression tree for this task?

--End of this document--