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).
Q1 Answer(Variable - Plot Type)
A. Single continuous variable - histogram 单个连续变量
B. Single categorical variable - Pie chart 单个类别变量
C. Single continuous variable and single categorical variable - Box plots 单个连续变量和单个类别变量 - 箱形图
D. Two continuous variables - Scatter plot 两个连续变量 - 散点图
E. Two categorical variables - two-bar plot 两个分类变量

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

(A) Discuss the connections and differences between partitional clustering and hierarchical clustering.
a) Answer首先说它们的联系：
它们都是聚类算法，它们最终都会把数据分为多个组，每个组之间具有相似性。

然后说不同：
... 概念题

(B) Explain self-organizing map and its “topology preserving properties”.
b) Answer参考资料：https://zhuanlan.zhihu.com/p/73534694

SOM算法基于无监督的竞争性学习。它提供了一个拓扑，保留了从高维空间到映射单位的映射。

拓扑保留属性是啥?

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

```
som_model <-
  som(data_train_matrix, # 原始数据
      grid = som_grid, # 初始化的网格图
      rlen = 1000, # 随机遍历一遍样本的次数
      alpha = c(0.7, 0.01), # 学习率的范围，逐渐从大到小
      mode = "online",
      normalizeDataLayers = false,
      keep.data = TRUE # 如果FALSE的话 最后固然会有数据，但是看不了
     ))
```

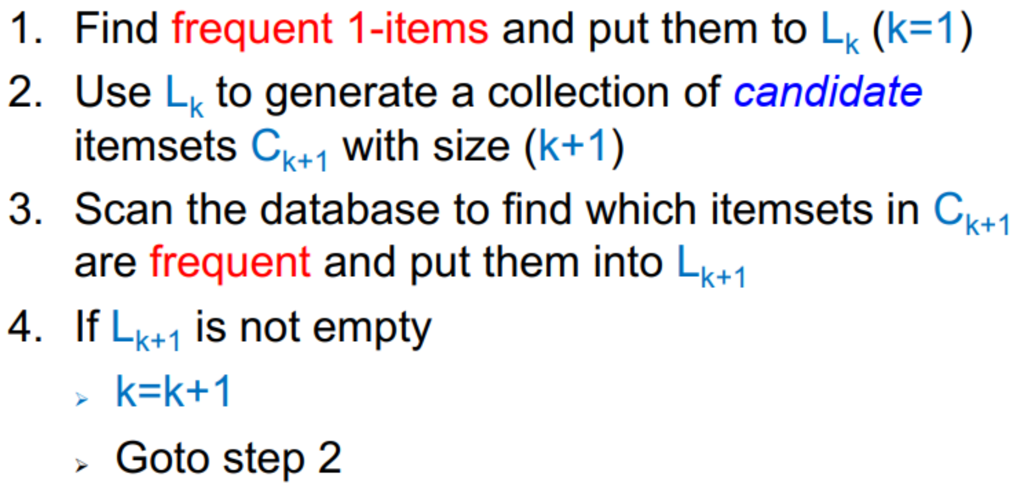
grid、rlen、alpha
grid refers to the initialized grid diagram
rlen refers to the the number of times the sample was traversed
alpha refers to the the range of learning rate

(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.
翻译：给定一组点，学生 A 和 B 应用 k 均值聚类分别将这些点聚类为 M 聚类。由于各种原因，它们的聚类结果并不相同。请您确定谁的聚类分析结果更好。请描述您的解决方案。

计算所有簇的误差平方和，比较两种算法的模型的SSE。SSE小的聚类结果更好。
Calculate the sum of the error squares of all clusters and compare the SSE of the models of the two algorithms.
The smaller the SSE value, the better the clustering results

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.



1. Find frequent 1-items and put them to L_k (k=1)
2. Use L_k to generate a collection of *candidate* itemsets C_{k+1} with size (k+1)
3. Scan the database to find which itemsets in C_{k+1} are frequent and put them into L_{k+1}
4. If L_{k+1} is not empty
   - k=k+1
   - Goto step 2

怎么生成候选项集?

- 将所有 频繁k项集进行两两组合，生成候选 k+1 项集。
- Combine all frequent k-itemsets in pairs to produce candidate k+1 itemsets.

怎么剪枝候选项集?

- 计算所有候选项集的支持度，去掉支持度小于阈值的候选项集。
- Calculates support for all candidate sets, eliminating candidate itemsets whose support is less than the threshold.

(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.

Transaction 1: A,D,F,H
Transaction 2: B,D,F,H
Transaction 3: A,D,F,H
Transaction 4: D,B,F,H
Transaction 5: A,B,F,H
Transaction 6: A,B,D,H

支持度 Support of (AD, FH) = 2/6 = 0.3 没错了
Confidence of ADFH = Support of ADFH/Support of AD = 0.3/0.5 =0.6
也没错了

(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.
度量“置信度”通常用于评估已开采关联规则的吸引力（关联性）。但是，有时高置信度值并不一定意味着规则确实很吸引力。讨论度量“置信度”的潜在问题，并解释如何在关联分析中解决此问题。
3.c 暂时不知道！！！

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).
K-最近邻 （k-NN） 分类器是一个简单有效的分类器。假设您获得了一组 M 个样本，并且还向您提供了每个样本的类标签。同时，另一组 N 个样本对您隐藏，它们将仅用作测试集来评估您开发的 k-NN 分类器。描述您将要遵循的过程，以便获得可以在测试集（即N个样本）上实现最高分类精度的k-NN分类器。(就是描述Knn的过程)
Q 4.1 AnswerWe are given with **M** samples of data along with their labels.There will be **N** samples of data will be used to evaluate the k-NN model.
knn算法描述以下就行。

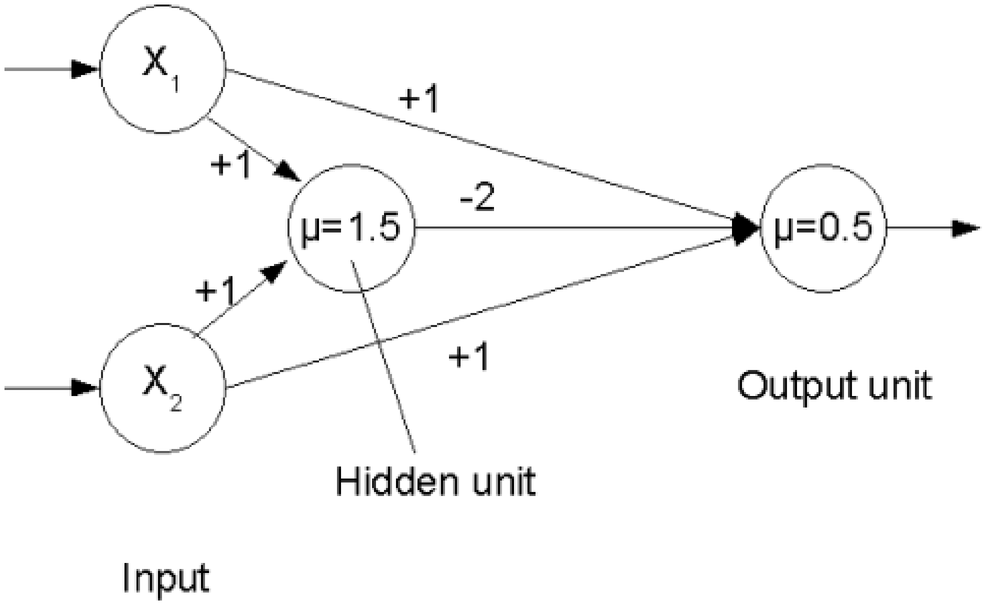
(B) Given a training set with the following properties:
Number of samples: 1900
Dimension of features in each sample: 45
Dimension of the target value for each sample: 3
Assume that this dataset is being used to train an MLP which has a single hidden layer with 20 neurons, and that the network is being trained for 400 iterations. What is the total number of weights (weight parameters) in this MLP? Show and explain how you derived your answer.
假设此数据集用于训练具有具有 20 个神经元的单个隐藏层的 MLP，并且该网络正在训练 400 次迭代。此 MLP 中的权重（权重参数）总数是多少？展示并解释您如何得出答案。
就是问有单个隐藏层的神经网络（即总共三层）的权重参数的总数。神经元之间有几个连接就有几个权重参数。简单的一批。

45 * 20 + 20 * 3 = 960

(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 $\mu = 0.5$, 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: x1=1.5, x2=1.2
Sample2: x1=0, x2=1.2
Sample3: x1=0.5, x2=0.5
Sample4: x1=1.6, x2=0
Compute the output produced by this network for each of these samples. You need to show the key steps of calculation.



以sample1为例：

隐藏层：

- $weight\ sum = x1 * 1 + x2 * 1 = 1.5 * 1 + 1.2 * 1 = 2.7 > \mu = 1.5$
- so, 隐藏层 值为 1

输出层：

- $weightsum = 1.5 * 1 - 2.7 * 2 + 1.2 * 1 = 0 < \mu = 0.5$
- so, output value = 0

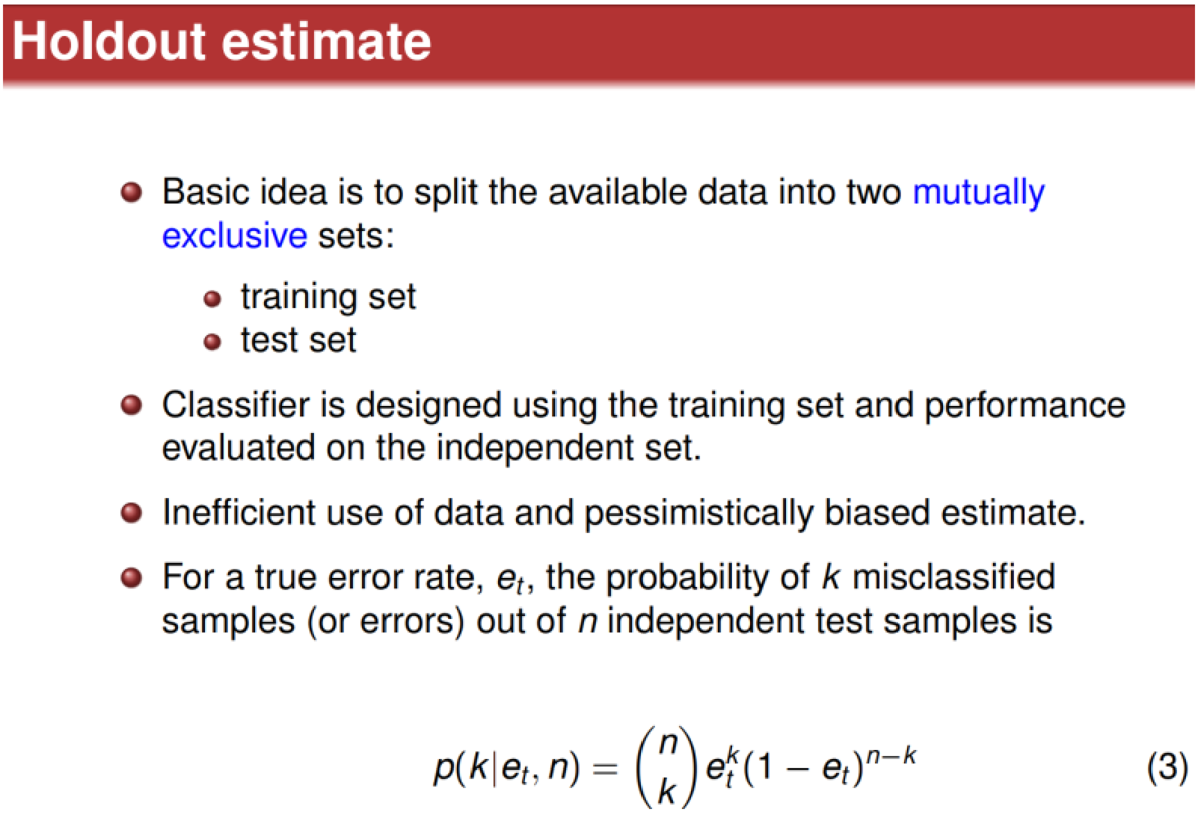
其他sample同理。

(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.
在隐藏层中使用非线性激活函数对于MLP网络模拟输入和输出变量之间的复杂关系非常重要。证明对于具有任意数量隐藏层的MLP网络，如果线性激活函数$f（x） = x$用于该MLP网络中的所有神经元，则该网络的输入和输出之间的关系将保持线性。
4.4 Answer这咋证明啊，，，这不是很显然的嘛~

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

(A) Explain hold-out and 10-fold CV and their strengths and possible weaknesses.

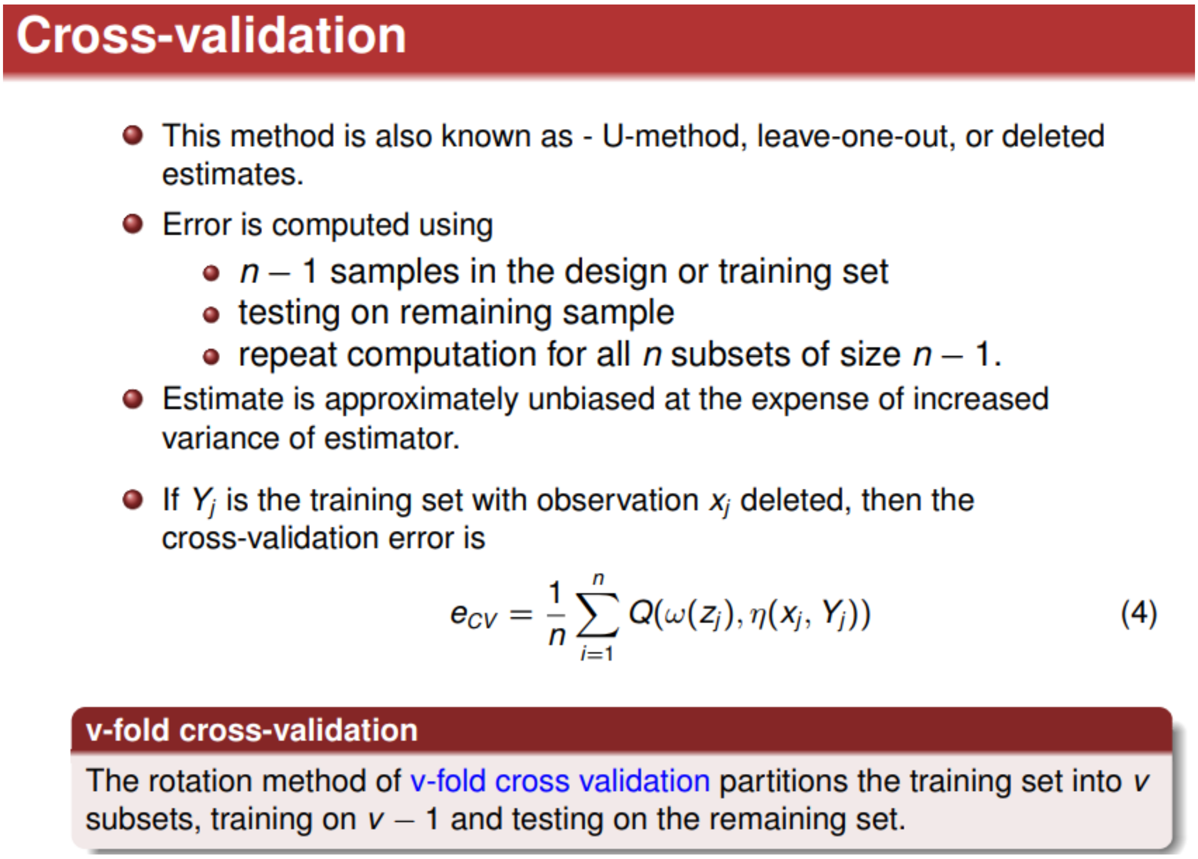
机器学习PPT中：



Holdout estimate

- Basic idea is to split the available data into two mutually exclusive sets:
  - training set
  - test set
- Classifier is designed using the training set and performance evaluated on the independent set.
- Inefficient use of data and pessimistically biased estimate.
- For a true error rate, e_t, the probability of k misclassified samples (or errors) out of n independent test samples is

$$p(k|e_t,n) = \binom{n}{k} e_t^k (1-e_t)^{n-k} \qquad (3)$$



Cross-validation

- This method is also known as - U-method, leave-one-out, or deleted estimates.
- Error is computed using
  - $n-1$ samples in the design or training set
  - testing on remaining sample
  - repeat computation for all n subsets of size $n-1$.
- Estimate is approximately unbiased at the expense of increased variance of estimator.
- If Y_j is the training set with observation x_j deleted, then the cross-validation error is

$$e_{CV} = \frac{1}{n}\sum_{i=1}^{n} Q(\omega(z_j),\eta(x_j,Y_j)) \qquad (4)$$

v-fold cross-validation

The rotation method of v-fold cross validation partitions the training set into v subsets, training on $v-1$ and testing on the remaining set.

留出法优点：

- 确定性：实验没有随机因素，整个过程是可重复的。

缺点：

- 结果不如交叉验证准确

交叉验证计算最繁琐，但样本利用率最高。适合于小样本的情况。

(B) Explain the overfitting problem.
概念题。。。

(C) Explain random forest method and how it can improve upon the standard decision tree for regression problems.
Random Forest (RF) is an algorithm that integrates multiple trees through the idea of integrated learning.
When a new input sample enters the forest, let each decision tree in the forest judge it separately, see which category the sample should belong to (for classification algorithms), and then use **the minority obedience majority method** to see which category is chosen the most, and predict which category the sample is.

(D) State one obvious reason why linear regression is not appropriate for binary/categorical response variable.
In the regression task, the model fits a real value. Regression models are often not directly used for classification tasks, essentially because the relationship between the fitted real values and the categories to be classified is not explicitly defined.

(E) Below is the optimisation problem for soft margin classifier
不会。。。

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

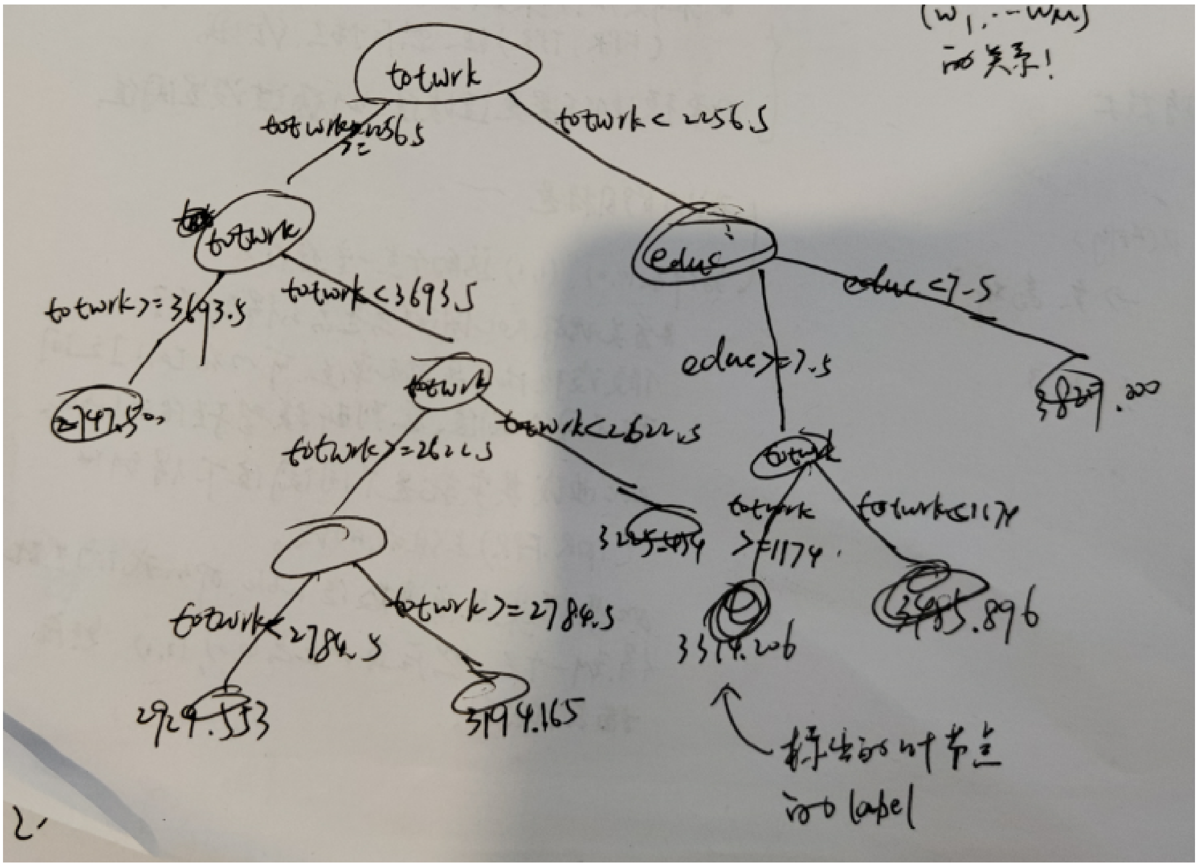
```
       n= 706

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.

翻译：睡觉很重要，它对工资有影响。然后给出一个包含706个样本的数据集。
我们要做的就是根据决策树的输出画出决策树。



6.2 Answer

1. 3225.434
2. 2747.500
3. 3314.206

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

$$SSE = \sum_{i=1}^{n_{test}} (y_i - \hat{y_i})^2 = (3220 - 3225.434)^2 + (2745 - 2747.500)^2 + (3320 - 3314.206)^2 = 69.349$$

$$MSE = \frac{SSE}{n_{test}} = \frac{69.349}{3} = 23.12$$
$$MAE = \frac{|3220 - 3225.434| + |2745 - 2747.500| + |3320 - 3314.206|}{3} = 4.576$$

(D) For the support vector machine classification methods, explain clearly the important differences between the support vectors and usual observations.
翻译：对于支持向量机分类方法，请清楚地解释支持向量与常规观测值之间的重要区别。
支持向量是离分割超平面最近的点。
而usual observation离分割超平面的距离都大于支持向量到超平面的距离。
The support vector is the point closest to the split hyperplane. The distance of usual observation from the split hyperplane is greater than the distance from the support vector to the hyperplane.

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

7.1 Answer

$$Presicsion = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{P}$$

for classifier1 :

- $$precision = \frac{230}{265} = 0.868$$
  $$recall = \frac{230}{262} = 0.878$$

for classifier2:

- $$precision = \frac{253}{285} = 0.888$$
  $$recall = \frac{253}{262} = 0.966$$

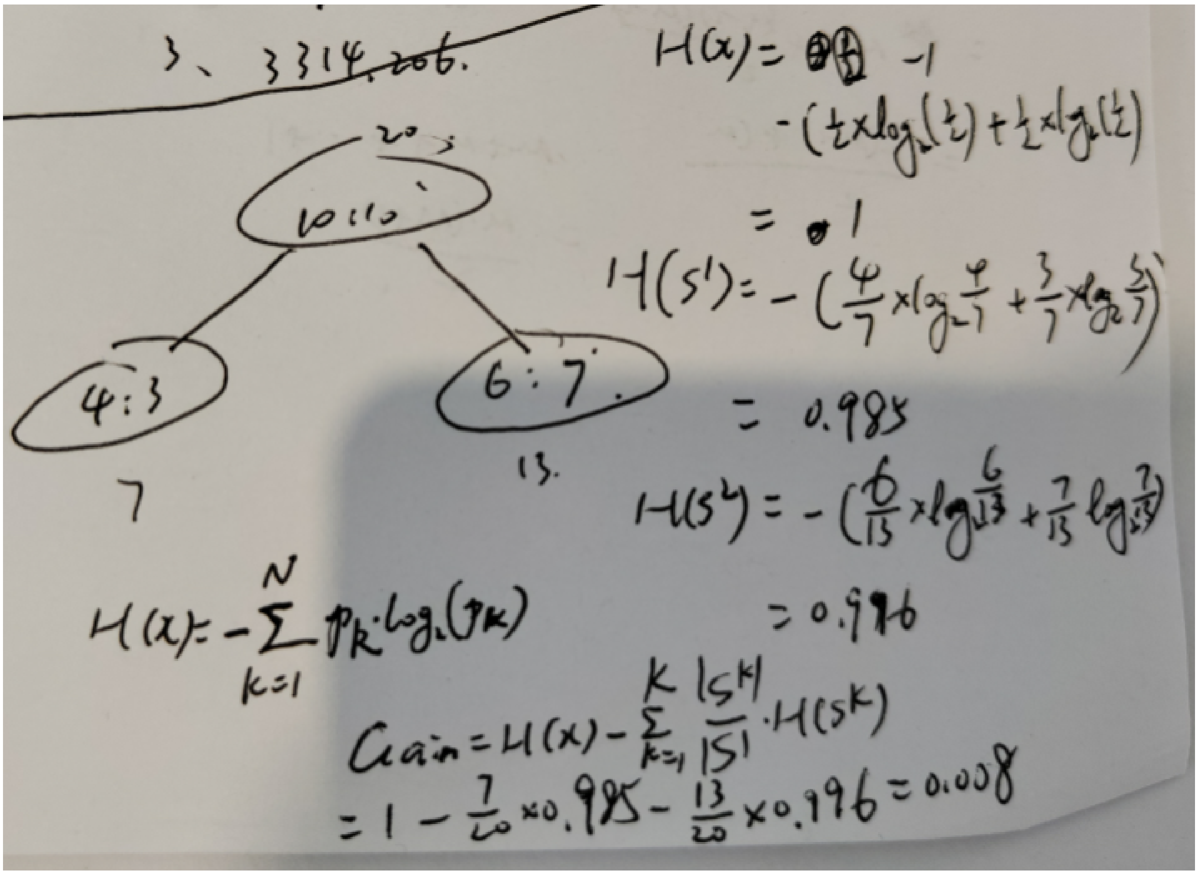
for classifier3:

- $$precision = \frac{180}{192} = 0.938$$
  $$recall = \frac{180}{262} = 0.687$$

这咋比较?

(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.
对于给定的决策树，假设在根节点中，有 10 个观测值属于类 1，10 个观测值属于类 2。然后，在第一次拆分之后，您有 • 左节点：4 个观测值（1 类）和 3 个观测值（2 类） • 右节点：6 个观测值（1 类）和 7 个观测值（2 类）。计算第一次拆分的信息增益。

7.2


3、3314.206.
20
10:10
4:3
6:7
7
13.

$H(x) = -1$
$-(\frac{1}{2}\times\log_2(\frac{1}{2})+\frac{1}{2}\times\log_2(\frac{1}{2}))$
$= 1$

$H(S^1) = -(\frac{4}{7}\times\log_2\frac{4}{7}+\frac{3}{7}\times\log_2\frac{3}{7})$
$= 0.985$

$H(S^2) = -(\frac{6}{13}\times\lg\frac{6}{13}+\frac{7}{13}\lg\frac{7}{13})$
$= 0.996$

$$H(x) = -\sum_{k=1}^{N} p_k \cdot \log_2(p_k)$$

$$Gain = H(x) - \sum_{k=1}^{K} \frac{|S^k|}{|S|} \cdot H(S^k)$$
$$= 1 - \frac{7}{20}\times 0.985 - \frac{13}{20}\times 0.996 = 0.008$$

(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.
是否有证据表明男性和女性每周夜间睡眠分钟数存在一些差异？使用显著性水平 0.05 验证结果的合理性。

(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?

7.5 Answer

sleep = 3640.23 - 0.1657 * torwrk + 2.01 * age + 87.55 * male - 11.77 * educ + 4.78 * kid + ε

For 6.B test1 :

- sleep = 3640.23 - 0.1657 * 2400 + 2.01 * 30 + 87.55 * 1 - 11.77 * 5 + 4.78 * 1 + ε = 3336.33 + ε

然后计算其他测试数据
再求MAE、MSE 完事儿