Q1

1. Boxplot
2. Bar plot or pie chart
3. Boxplot
4. Scatterplot
5. Mosaic plot

Q2

1.他们都是not overlap 的聚类（不确定）

区别是partitional 聚类没有层级关系，hierarchy聚类有层级关系，上级会包含下级的clusters。

2.不考

Som可以视为两层的神经网络，输入层和竞争层。它把高维的数据降低到两维同时保留拓扑结构。保留拓扑结构的意思是在高维中相近的在som mapping里面也相近。

3.迭代次数，cell连接的形状（正方形还是六边形），cell数量，几乘几，距离算法

4. the mean distance to the center of clusters. 用这个评价

Q3 假设abcde 五个item

生成5个itemset {a}… 计算{a}….的frequency。

If 低于门槛值 去掉该itemset（剪枝）

Merge:假如目前迭代次数是k那么用前k-1个item 相同的就合并。

合并完后检查子集是否有被剪枝 比如{a,b,c}由{a,b}{a,c}合并而来 但是{b,c}早被删掉了那么就是把{a,b,c}给删了。

检查目前itemset里面的frequency是否高于门槛值，这一步用哈希树来做。

2. ADFH ADFH ADFH ADH ADB A B D F H 这样support=3/10， confidence = 3/5

3. c老师回答：The potential issue with the measure "confidence" in association analysis is that it only considers the conditional probability of the consequent given the antecedent, without taking into account the overall frequency of the rule. This can lead to high confidence values for trivial or uninformative rules. To address this issue, additional measures such as support, lift, and conviction are used to assess the interestingness of association rules, providing a more comprehensive evaluation of their significance and usefulness.

Q4：

1. 在knn里面唯一能调整的就是k值。不知道这题啥意思。

C老师回答：To obtain a k-NN classifier with the highest classification accuracy on the test set, I will first preprocess the training set by normalizing or standardizing the features. Then, I will iterate through different values of k, train the k-NN classifier using the training set, and evaluate its performance using cross-validation techniques such as k-fold cross-validation. Finally, I will select the value of k that yields the highest accuracy and use it to classify the hidden test set samples. c老师的意思就是做预处理

1. 输入层没有权重参数，hidden layer 20个神经元每个45个权重参数 共900个。输出层要压缩到3个输出，MLP是全连接的(这个估计要着重解释一下) 每个20个权重参数。总共960个权重参数。
2. 前向传播。

Hidden unit 里面发生的计算：

Sample 1: 2.7 > 1.5

Sample 2 1.2<1.5

Sample 3 1<1.5

Sampler4 1.6>1.5

所以输出的是1 0 0 1

Output发生的计算：

Sample1 0.7 >0.5

Sampler2 1.2 >0.5

Sample 3 1>0.5

Sample 4 -0.4<0.5

所以输出的是1110

1. 代入就行了

第1层是 满足条件

第k-1层是 =

第k层是=

只要k-1层是线性的，k层就是线性的。

Q5

Hold out节省算力，缺点是训练没有覆盖全部数据集，而且如果模型方差较大，结果评估变动会比较大。

10 cv 优点是利用了全部数据集在训练（训练多次），降低了过拟合对评估的影响，评估波动更稳定。缺点是算力消耗大而且数据集切分十份可能会使得训练样本太少。

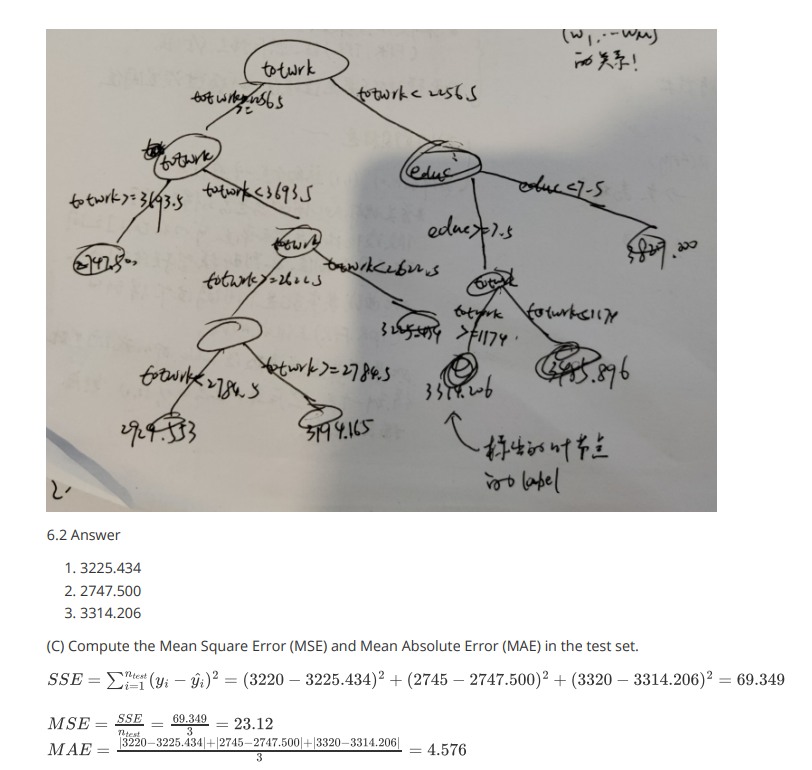
2. 过拟合就是模型对训练数据拟合得太specialized（训练时评估很好） 以至于对unseen的数据表现性能差很多的情况。

3.随机森林就是利用bootrap技术有放回抽样，生成许多的训练集，在每个训练集上生成一棵树。最终综合所有树的结果来判定样本归属。

决策树容易过拟合，随机森林对不同的训练集训练了很多树来帮助抗过拟合。

1. M的作用类似于SVM的margin，优化目标就是使得margin尽可能大。所有的变量和超平面的距离必须比M\*（放松系数）大。
2. epsilon 是slack variable，它越大，限制越松，虽然对非线性可分的拟合效果变好了，但是会导致过拟合问题。参数C是限制epsilon的值，类似于L1正则化。而且注意的是如果没有C值那么该函数没有办法收敛。这是因为对于任何M的值来说，都可以通过适配一个epsilon的值（比如接近1）来满足几个subject的条件。所以beta理论上任何值都可以（当然要满足第一个条件），也就是说M和beta脱钩了，参数训练没有意义。函数也无法收敛。

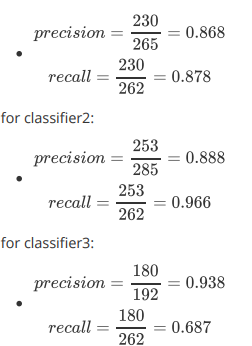
Q6：

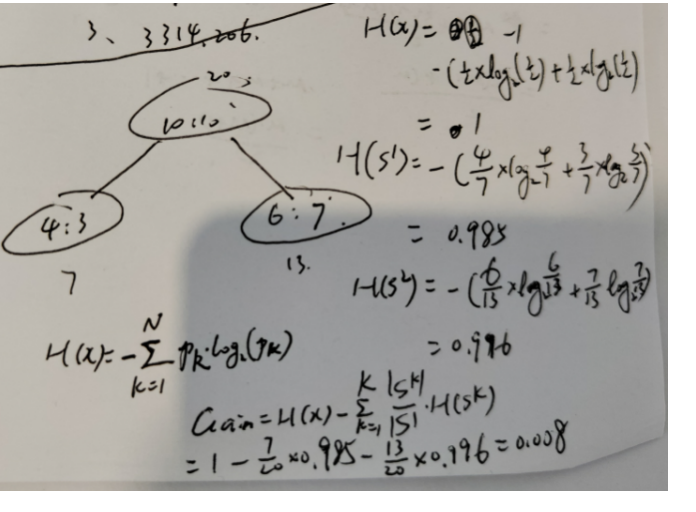


D 支持向量是离分割超平面最近的点。 而usual observation离分割超平面的距离都大于支持向量到超平面的距离。支持向量会支撑最后的模型，usual observation可以删掉都没影响。

Q7:

1.

算F1就可以比较了。

2. 

3.有

4.有

5.