CS6220 — Fall 2015

Final exam

Tuesday December 15, 2015

Time: 1 hour 40min

Name (please print): _____

Show all your work and calculations. Partial credit will be given for work that is partially correct. Points will be deducted for false statements, even if the final answer is correct. Please circle your final answer where appropriate.

This exam is closed-book. You may consult one page with your personal notes. Calculators are permitted.

Honor code: I promise not to cheat on this exam. I will neither give nor receive any unauthorized assistance. I will not to share information about the exam with anyone who may be taking it at a different time. I have not been told anything about the exam by someone who has taken it earlier.

Signature: _____ Date: _____

Question	Possible Points	Actual Points
1	21	
2	7	
3	21	
4	20	
5	10	
6	7	
7	14	

1. For the following questions, circle **TRUE** or **FALSE**, and provide an explanation.

   (a) **(7 pts)** In agglomerative hierarchical clustering, single-linkage means that the distance between two clusters is taken to be the distance between their cluster centers.

   TRUE **FALSE**

   Answer: FALSE

   The single-linkage means that the distance between two clusters is taken to be the distance between two elements (one in each cluster) that are the closest to each other.

   (b) **(7 pts)** In k-means, the cluster representative (cluster center) does not need to be an observed data point.

   TRUE **FALSE**

   Answer: TRUE

   The coordinates of the centroid are the (arithmetic) mean of the feature values over all the observations in the cluster. It does not need to be one of the data points.

   (c) **(7 pts)** Suppose that the data are stored in an array with elements x_{ij}, where rows $i = 1, \ldots, I$ indicate observations and columns $j = 1, \ldots, J$ indicate variables. The objective of k-means is to minimize

   $$\sum_{cluster\ k=1}^{K} \sum_{\substack{i: \\ i \in cluster\ k}} \sum_{j=1}^{P} (x_{ij} - \bar{x}_{\cdot j})^2 \text{ but not } \sum_{cluster\ k=1}^{K} \sum_{\substack{i, i': \\ i \neq i' \\ i, i' \in cluster\ k}} \sum_{j=1}^{P} (x_{ij} - x_{i'j})^2$$

   TRUE **FALSE**

   Answer: FALSE *The two criteria are equivalent*

2. **(7 pts)** Explain the meaning of "the first principle component explains 80% of variation".

   Answer: *If the data are an array X with I rows (i.e., observations), and P columns (i.e., variables), then the total variation is defined as*

$$\sum_{dimensions} \sum_{j=1}^{P} Var(x_{ij})$$

   i.e. the sum of the variances of the observations over all the original dimensions.

   The PCA centers and rotates the system of coordinates. "The first principle component explains 80% of variation" means that the variance of the observations in the direction of the first transformed coordinate explains 80% of the variance summed over all dimensions. In other words, the variation along the first principle component explains 80% of variation along all the original dimensions combined.

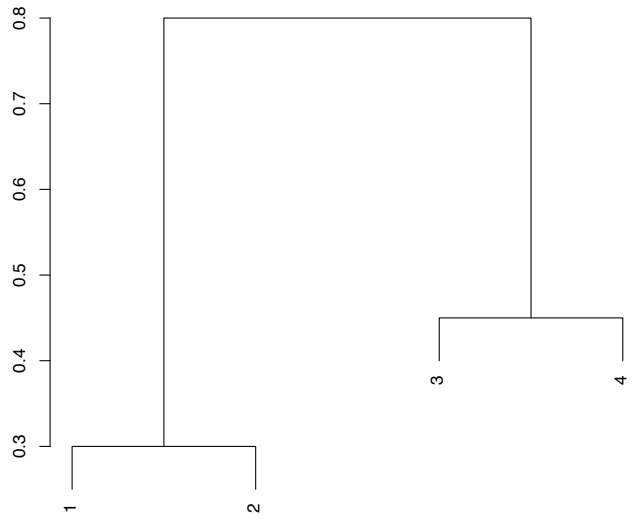
3. Suppose that we have four observations, for which we are given the dissimilarity matrix

$$\begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0.0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0.0 \end{bmatrix}$$

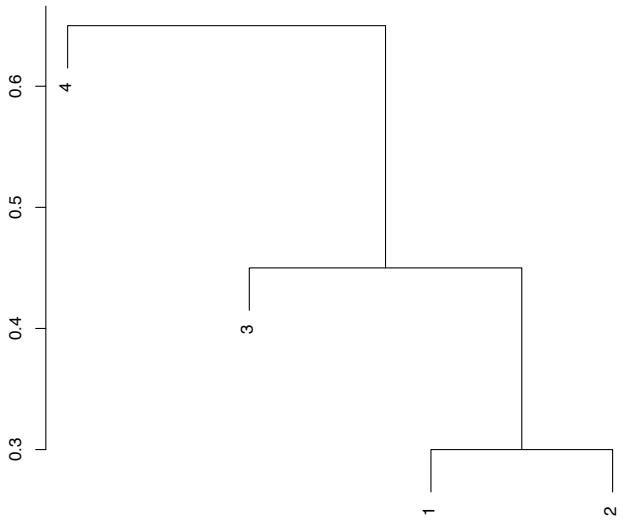
   For example, the dissimilarity between the first and the second observations is 0.3.

   (a) **(7 pts)** On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using **complete** linkage. Indicate the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

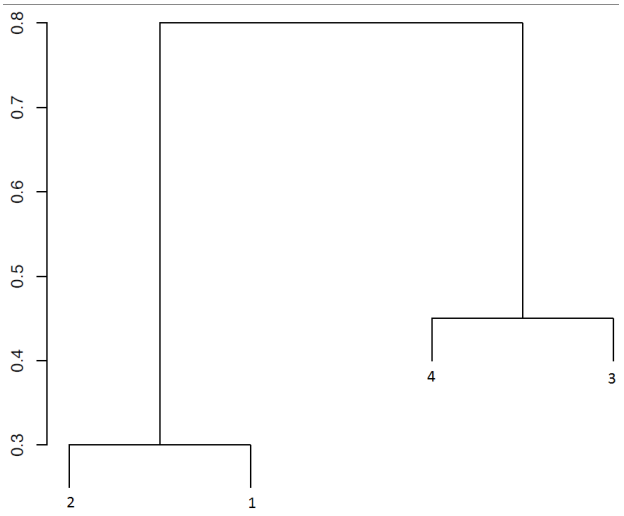
   Answer:



(b) **(7 pts)** Repeat (a), this time using **average** linkage clustering.
Answer:



(c) **(7 pts)** The order of the leaves in the dendrogram is not unique. Draw a dendrogram that is equivalent to the dendrogram in (a), but which has a different order of the leaves.
Answer:



4. Consider the following data. X is the only predictor, and Y is the response that can take on two different values, 0 and 1. We use the Gini index as the impurity measure. The optimal split is the split that maximizes the impurity reduction.

X	8	11	12	12	12	15	15	15	18	23
Y	0	0	0	0	1	0	0	1	1	1

(a) **(10 pts)** Which candidate split(s) should we evaluate to determine the optimal split?

(*Hint:* It is not necessary to enumerate all possible splits. The optimal split can only occur between two segments, where a segment is a collection of consecutive X values for which the class distribution is identical. For example, an optimal split will not separate $X = 12$ and $X = 15$, since they have an identical class distribution.)

Answer:

At the very least, we only want to consider the midpoints between the distinct values of X, i.e. $\{9.5, 11.5, 13.5, 16.5, 20.5\}$.

However, we can do better. According to the hint, it is not necessary to consider all possible splits. The optimal split can only occur between two segments, where a segment is a collection of consecutive X values for which the class distribution is identical. The splits that separate the segments are $\{11.5, 16.5\}$. *Therefore, only* $X \leq 11.5$ *and* $X \leq 16.5$ *need to be considered.*

(b) **(10 pts)** What is the optimal split on X, and what is the impurity reduction of that split?

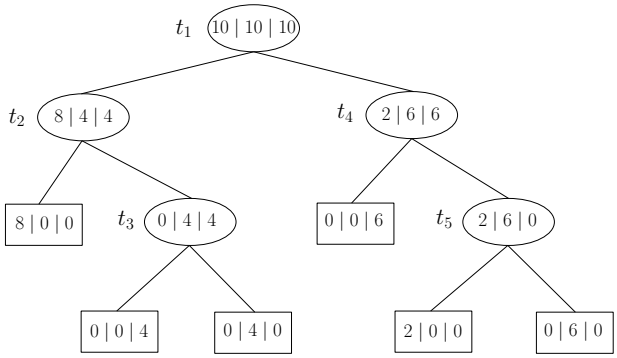
Answer:

The first split produces child nodes with class distributions $2|0$ *and* $4|4$. *The second produces child nodes with class distributions* $2|0$ *and* $6|2$. *The second split is better. The Gini index is:*

$$i(t) = \sum_j p(j|t)(1 - p(j|t))$$

The Gini Index of root:
$\frac{4}{10} \cdot \frac{6}{10} = \frac{24}{100}$
The Gini Index of the right node:
$\frac{2}{10} \cdot 1 \cdot 0 = 0$
The Gini Index of the left node:
$\frac{8}{10} \cdot (\frac{2}{8} \cdot \frac{6}{8}) = \frac{3}{20}$
The impurity reduction:
$\Delta i = \frac{24}{100} - \frac{3}{20} = 0.09$

5. **(10 pts)** The tree given below, denoted by T_{max}, has been constructed on the training sample. In each node, the number of observations with class 0 is given in the left part, the number of observations with class 1 in the middle part, and the number of observations with class 2 in the right part. The leaf nodes have been drawn as rectangles.



Find the first internal node of the tree that needs to be pruned. Specifically, use resubstitution error in the "cost' part of the cost-complexity formula. Fill in the table below the value of the penalty parameter α that would result in pruning of this node. Circle the label of the node that will be selected first for cost-complexity pruning.

	t_1	t_2	t_3	t_4	t_5
α					

Answer:

	t_1	t_2	t_3	t_4	t_5^*
α	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

The starred node corresponds to the smallest α, and should be pruned first.

$\alpha = \frac{R(t)-R(T_t)}{(|T_t|-1)}$

$\alpha(t_1) = \frac{\frac{20}{30}-0}{(6-1)} = \frac{2}{15}$

$\alpha(t_2) = \frac{\frac{16}{30}\cdot\frac{8}{16}-0}{(3-1)} = \frac{2}{15}$

$\alpha(t_3) = \frac{\frac{8}{30}\cdot\frac{4}{8}-0}{(2-1)} = \frac{2}{15}$

$\alpha(t_4) = \frac{\frac{14}{30}\cdot\frac{8}{14}-0}{(3-1)} = \frac{2}{15}$

$\alpha(t_5) = \frac{\frac{8}{30}\cdot\frac{2}{8}-0}{(2-1)} = \frac{1}{15}$

6. **(7 pts)** Provide a detailed explanation of the boosting algorithm. What are the advantages/disadvantages of boosting as compared to trees, bagging, and random forest?

Answer:

Boosting is a machine learning algorithm that build a strong classifier from a set of weak classifiers. The method aims at reducing bias and variance of predictions.

In the context of decision trees, boosting iteratively creates trees, using modified versions of the full dataset at each iteration.

1. *Set f(x) = 0 and $r_i = y_i$ for all i in the training set.*
2. *For b = 1, 2, . . .,B, repeat:*
   (a) *Fit a tree f^b with d splits (d+1 terminal nodes) to the training data (X, r).*
   (b) *Update f by adding in a shrunken version of the new tree:*
   $$f(x) = f(x) + \lambda f^b(x)$$
   (c) *Update the residuals,*
   $$r_i = r_i \lambda f^b(x_i)$$
3. *Output the boosted model,*
   $$f(x) = \sum_{b=1}^{B} p^b(x)$$

Advantages/disadvantages: bagging uses subsets of the observations, and random forest also subsets of variables, and grow maximal trees. Boosting uses the full set of observations and variables, and short trees. It's advantage is that it can improve the performance of the classifier for a small number of hard-to-predict samples. It's disadvantage is that it has more tuning parameters, and is more likely to overfit. Since it grows short trees it mainly decreases bias but does not decrease the variance.

7. Suppose that P, Q, and R are three different web pages. For each question below, give an example graph. You do not need to provide the numerical values, but you need to provide an explanation.

   (a) **(7 pts)** Give a simple example of a graph, where adding a link from P to Q can **raise** the PageRank of R.

   Answer:

   Initial graph: $P; Q \rightarrow R$

   Adding a link from P to Q raises the PageRank of Q and thus indirectly the PageRank of R.

   (b) **(7 pts)** Give a simple example of a graph, where adding a link from P to Q can **lower** the PageRank of R.

   Answer: *Initial graph:* $P \rightarrow R; Q$

   If you add a link from P to Q, then P's "contribution of importance" is divided between Q and R rather than going exclusively to R, so the PageRank of R decreases.