STAT 318/462: Data Mining
Assignment 3
Due Date: 4pm, 14th October, 2020

Please submit your assignment as a single pdf on Learn.

You may do the assignment by yourself or with one other person from the same cohort (300-level students cannot work with 400-level students). If you hand in a joint assignment, you will each be given the same mark. Marks will be lost for unexplained, poorly presented and incomplete answers. Whenever you are asked to do computations with data, feel free to do them any way that is convenient. If you use *R* (recommended), please provide your code. **All figures and plots must be clearly labelled.**

1. **(6 marks)** In this question you will grow a depth two (two levels of splitting) CART tree for a two-class classification problem using the $n = 100$ training observations given in Table 1. This table summarizes a data set with three binary-valued (0 or 1) features, X_1, X_2, and X_3, with two class labels *Yes* or *No*. For example, there are 20 observations of the form (1,0,1) with class label *Yes* in the data set.

X_1	X_2	X_3	# Yes observations	# No observations
1	1	1	5	0
0	1	1	0	20
1	0	1	20	0
0	0	1	0	5
0	1	0	25	0
0	0	0	0	25

Table 1: Training data for Question 1.

   (a) Using Gini Index as the measure of impurity and splits of the form $X_i < 0.5$, find the best split for the training observations. Show the reduction in impurity for each split you consider.

   (b) Repeat part (a) for the two daughter nodes found in part (a). Then, stop splitting and sketch the resulting tree.

   (c) How many training observations are misclassified in your tree?

   (d) Now, grow a depth two tree using the same method as above, **but use X_3 to split the root node**. Sketch the resulting tree.

   (e) Use parts (c) and (d) to conclude about the greedy nature of CART.

2. **(10 marks)** In this question, you will fit regression trees to predict *sales* using the Carseats data. This dataset has been divided into training and testing sets: `carseatsTrain.csv` and `carseatsTest.csv` (download these sets from Learn). Use the `tree(), randomForest()` and `gbm()` *R* functions to answer this question (see Section 8.3 of the course textbook).

   (a) Fit a regression tree to the training set (do not prune the tree). Plot the tree and interpret the results. What are the test and training MSEs for your tree?

Transaction ID	Items Bought	Transaction ID	Items Bought
1	$\{a, b, d, e\}$	6	$\{b, d, e\}$
2	$\{b, c, d\}$	7	$\{c, d\}$
3	$\{a, b, d, e\}$	8	$\{a, b, c\}$
4	$\{a, c, d, e\}$	9	$\{a, d, e\}$
5	$\{b, c, d, e\}$	10	$\{b, d\}$

Table 2: Market basket transactions for Question 3.

(b) Use the `cv.tree()` *R* function to prune your tree (use your judgement here). Does the pruned tree perform better?

(c) Fit a bagged regression tree and a random forest to the training set. What are the test and training MSEs for each model? Was decorrelating trees an effective strategy for this problem?

(d) Fit a boosted regression tree to the training set. Experiment with different tree depths, shrinkage parameters and the number of trees. What are the test and training MSEs for your best tree? Comment on your results.

(e) Which model performed best and which predictors were the most important in this model?

3. **(4 marks)** Using the itemset lattice in Figure 1 (on page 3) and the transactions given in Table 2, answer the following questions. Assume *minsup* = 30%.

(a) Label each node in the itemset lattice with the following letter(s):

M: if the node is a maximal frequent itemset;
C: if the node is a closed frequent itemset;
F: if the node is frequent, but not maximal nor closed;
I: if the node is infrequent.

(b) Compute the confidence and lift of $\{d, e\} \to \{a\}$. Comment on what you find.

Question 4 is for students taking STAT462. STAT318 students will NOT receive additional credit if they choose to answer this question. This is an independent research question (you will not be taught this material in class), but you will find Section 9.6 of the course textbook very useful.

4. **(4 marks)** In this question, you will fit support vector machines to the Banknote data from Assignment 2 (on the Learn page). **Only use the predictors x_1 and x_2 to fit your classifiers.**

(a) Is it possible to find a separating hyperplane for the training data? **Explain.**

(b) Fit a support vector classifier to the training data using `tune()` to find the best `cost` value. Plot the best classifier and produce a confusion matrix for the testing data. **Comment on your results.**

(c) Fit a support vector machine (SVM) to the training data using the radial kernel. Use `tune()` to find the best `cost` and `gamma` values. Plot the best SVM and produce a confusion matrix for the testing data. **Compare your results with those obtained in part (b).**

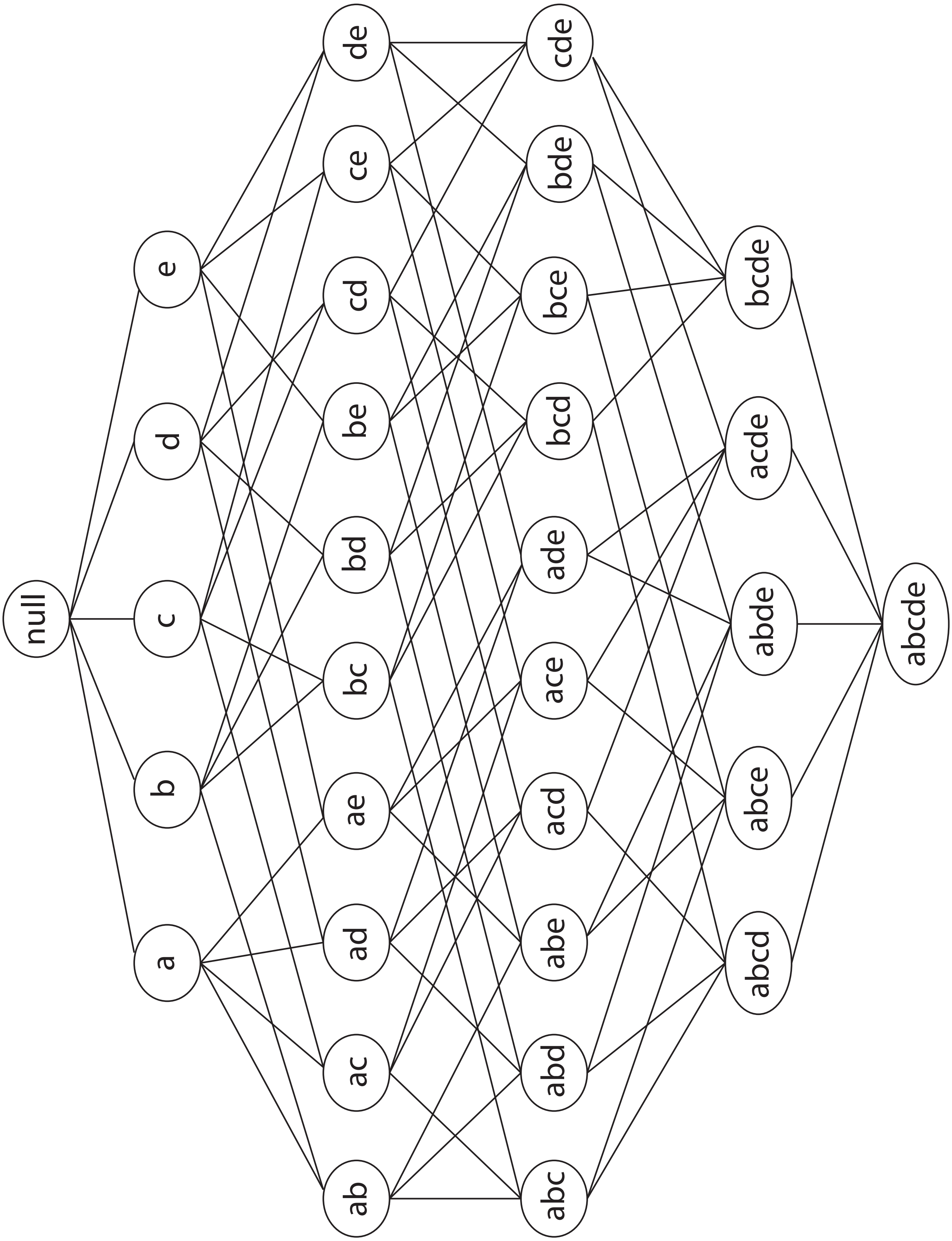

Figure 1: Itemset lattice for Question 3.