BONUS HW 2
CSC 522
rsinha2
Rishabh Sinha

Q1. (15 points) Data mining short answer questions
a) (3 points) Label each of the following similarity measures as good or bad for finding similarity in document data (often represented as “word, frequency”). Provide a one-line justification for each label you provide.
(i) Correlation
Correlation is a bad measure for text analysis, as correlation give best value when a linear relationship exists, But relationship between text documents cant be described as linear.
(ii) Cosine
This is good for documents as it, takes care of the high dimensionality problem, which occurs in euclidean distance. And just takes care of words which are in the text document rather than not.
(iii) Euclidean
This is bad for documents as euclidean distance is not very meaningful for high dimensional data.

(b) (2 points) Distinguish between data normalization and data standardization; give one example of each.
Normalization of data reduces the data element and rescales it in form of [0, 1], an example of normalization is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization is when you transform the data such that mean lies at 0, and unit variance, An Example is:

$$x_{new} = \frac{x - \mu}{\sigma}$$

(c) (4 points) For given two vectors, x = (0,1,0,0,0) and y = (0,1,0,0,1), compute the following:?
(i) Cosine
    value of cosine similairty is: x.y/|x||y| = 1/(1*sqrt(2)) = 0.707
(ii) Jaccard
    value of jaccard is: F11/(F11+F10+F01) = ½ = 0.5
(iii) Euclidean
    value of euclidean distance = 1
(iv) Correlation
    Correlation is: cov(x,y)/var(x)*var(y) = 0.612

(d) (2 points) Choose the best classifier form kNN and multilayer perceptron (MLP) for each of the following scenarios and explain why?

(i) When training dataset is large, test dataset is small

> When training dataset is large and test dataset is small, the knn would be a better classifier as, in knn, as it would be a faster as well as easier to classify the data using knn, and it is also less prone to overfitting.

(ii) When training dataset and test dataset are large

> If we have a large dataset, MLP would be more helpful as we can divide this test set to a validation set to prevent overfiitng and achieve a better accuracy of the MLP.

(e) (1 point) The number of maximal frequent itemsets decreases monotonically with increase in support threshold (True/False):

True, as when we increase the size of support threshold, we will have lower and lower number of frequent dataset thus also lower maximal frequent dataset.

(f) (3 points) Based on the attribute types in the dataset, choose one best classifier from (decision trees, neural networks) for each attribute type

(i) Continuous attributes:

> A neural network would perform better in case of a continuous attribute, as there is no need to divide the data into categories, which often leads to loss of information

(ii) Categorical attributes:

> Decision tree is better in case of categorical attributes, As, each neuron in a neural net has activation functions which take in input as a numeric value, so categories need to be converted to numeric value for Neural Net, while decision tree easily handles it.

(iii) A combination of both continuous and categorical attributes:

> When we have a combination of both, still decision tree can be used, as we can easily divide the continuous values into 2 sets, where as it may not always be semantically meaningful to convert category to numeric values.

Q2. (6 points) You have two similarity measures:
Dice coefficient and Roger and Tanimoto coefficient,
which are defined as below. Dice coefficient : S_DICE = 2a/(2a+b+c)
Roger and Tanimoto coefficient: S_RT = (a+d)/(a+2b+2c+d)
Where, a is the number of features that equal to 1 in both objects;
b is the number of features that are equal to 1 in the first object but are equal to 0 in the second object;
c is the number of features that are equal to 0 in the first object but are equal to 1 in the second object;
and d is the number of features that are equal to 0 in both objects.
Answer the following questions:
(a) (2 points) Is the Dice coefficient appropriate for asymmetric variables? Why?

Yes, I think Dice coefficient is appropriate for asymmetric variable as it only includes where the feature exists, that is a (1,1) and doesn't include the case of (0,0).

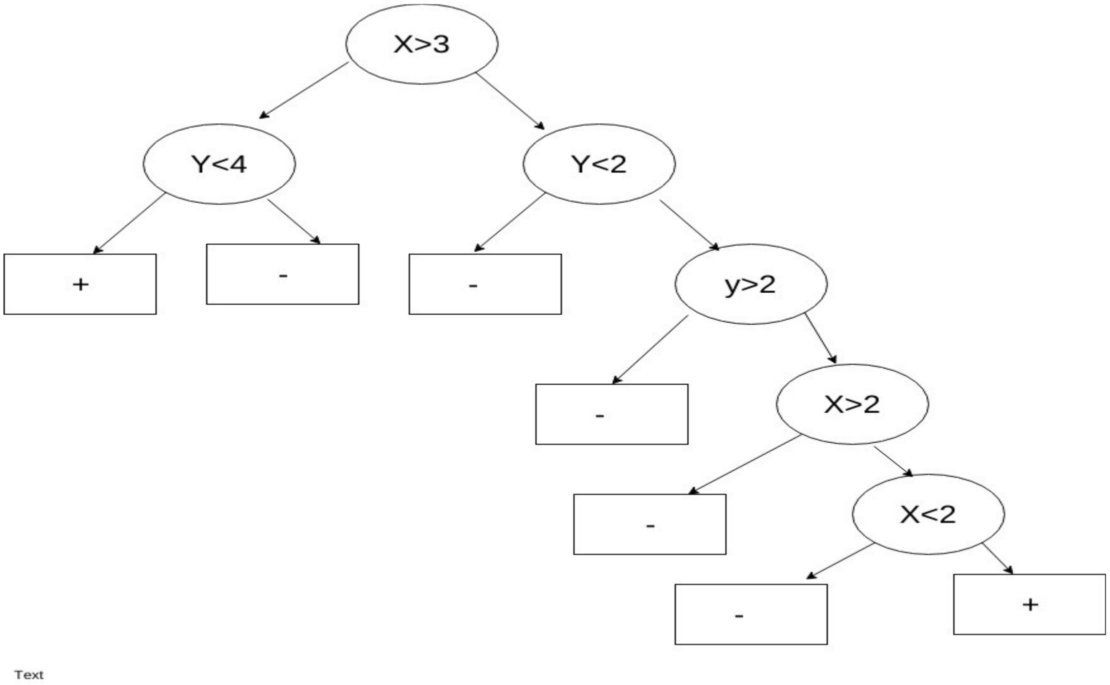
(b). (2 points) Is the Roger and Tanimoto coefficient appropriate for asymmetric variables? Why?
No, I don't think Roger and Tanimoto coefficient is appropriate for asymmetric variable as it includes both, where the feature exists, that is a (1,1) and also includes the case of (0,0) in the numerator. Thus gives equal importance to both cases, feature feature exists and doesn't exist.

(c). (2 points) Here, the task is to compute the similarity between two documents. Each document is represented as a vector of binary features, where each feature is a word in a dictionary. In this vector, a 1 indicates the presence of the word and a 0 indicates its absence. In this scenario, which measure (Dice coefficient or Roger and Tanimoto coefficient) will you choose? Why?

In this scenario, I would choose, the Dice Coefficient, as for comparison of text documents, it is important to consider, the words that exist in both documents rather than does that do not exist is both documents, as there can be millions of word in dictionary, but very few maybe in documents.

Q3. (20 points) Dataset for this problem is shown in Figure 3. This figure represents 2D scatter plot of 25 data points (two attributes X and Y), where each data point is placed at the interaction of uniform grid lines. This is binary class problem, where the two classes are represented by "-" and "+" symbols. Assume that the class labels are centered at grid intersection.

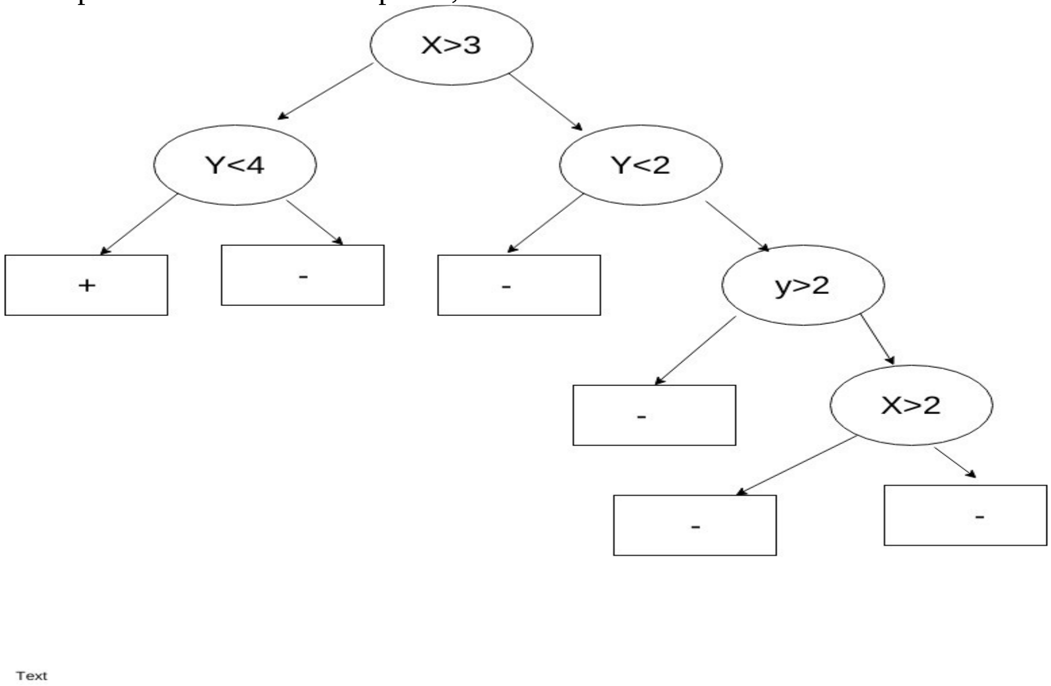
(a) (5 points) Construct a binary decision tree, which uses minimum number of splits to perfectly classify (meaning 100% accurate) each training data instance given in the figure. [Hint: Note that you are not required to compute any impurity measure at each split for constructing the tree, you can obtain solution through visual inspection of the data]. Show the resulting tree with attributes (at internal nodes), split conditions, and classes (leaf nodes) clearly marked. Also mark resulting boundaries on Figure 3.
The tree ensures to take care of equality cases, as all points lie at points where X&Y at integers.



(b) (13 points) Use pessimistic estimate of the generalization error to prune the tree constructed in (a) using the sub-tree replacement post-pruning method. Use Omega = 2 as the cost of adding a leaf node for calculating pessimistic estimate. In case of tie when determining majority class at leaf node, use “-“ as the default class. Show all the calculations and the resulting trees at each step of pruning. [Hint Stop recursive pruning when the pessimistic estimate on pruned tree start exceeding the pessimistic estimate on the original tree.]

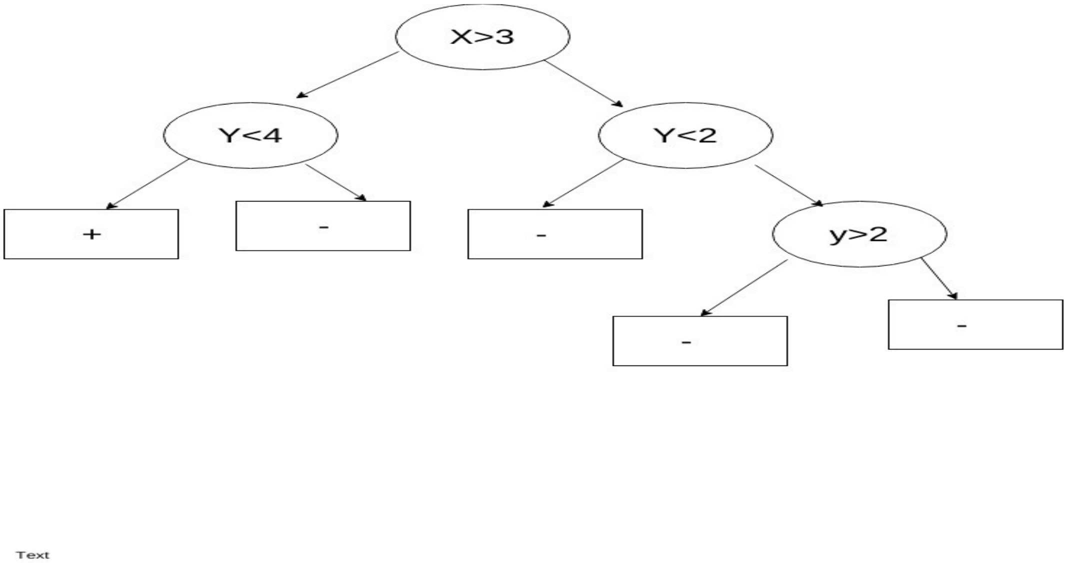
e'(pruned T) = ((6*2) + 1)/25 = 0.52
e'(un pruned T) = ((7*2))/25 = 0.56

Since error of unpruned tree is more we prune, new tree is:



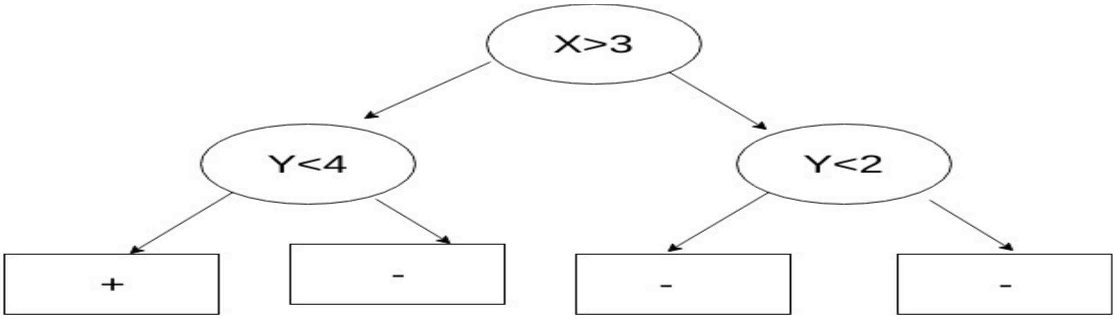
e'(pruned T) = ((5*2) + 1)/25 = 0.44
e'(un pruned T) = ((6*2)+1)/25 = 0.52

Since error of unpruned tree is more we prune, new tree is:



e'(pruned T) = ((4*2) + 1)/25 = 0.36
e'(un pruned T) = ((6*2)+1)/25 = 0.44

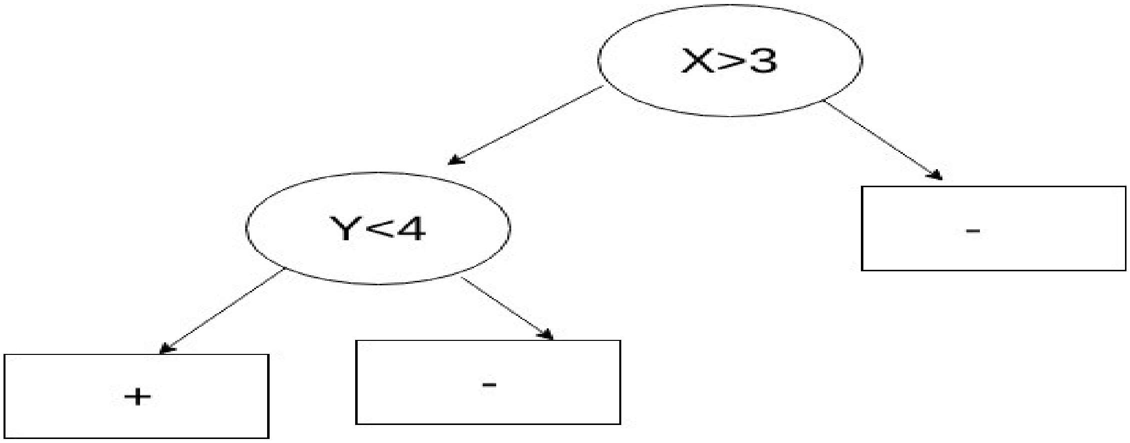
Since error of unpruned tree is more we prune, new tree is:



Text

e'(pruned T) = ((3*2) + 1)/25 = 0.28
e'(un pruned T) = ((4*2)+1)/25 = 0.36

Since error of unpruned tree is more we prune, new tree is:



e'(pruned T) = ((2*2) + 7)/25 = 0.44
e'(un pruned T) = ((3*2)+1)/25 = 0.28

Since error of pruned tree is more we dont prune, The final tree is thus above.

(c) (2 points)Compare resulting pruned tree in (b) with original tree in (a) through visual inspection. Would you use the original tree or the pruned tree for classifying new data? Explain?

I would use the pruned tree, as the pruned tree is much simpler, and the original tree was a result of overfitting the model, due to the presence of an outlier. This is a direct result of Occam's razor, which states always one must choose the simpler model, when results are similar.

Q4. (15 points) (Anomaly Detection)
(a) (5 points) You are given a data set containing the height, weight, age, and blood pressure of a representative sample of people from a major metropolitan area. Comment (justify with proper arguments) on the suitability of using a statistically based versus a cluster-based outlier detection scheme to identify people with anomalous characteristics for this data set.

If I am given a dataset containing, height, weight, age and blood pressure of people in major metropolitan area, If we use as Statistical based approach to identify anomalous characteristics, It would be simpler using Statistical based approach as computationally it is simple, and is backed by strong mathematics, however, Since, the data is multi dimensional, it would be difficult to asses the dataset, as each of the attributes may follow a different distribution, Further features such as height and weight, are not same for everyone and is affected by things such as gender as well as race of people, and since it is a metropolitan, it would be a diverse group of people. Thus it would be difficult to find anomalies due to this reason,
However if we are using clustering based approach, we can, divide data into different clusters based on gender etc, and in that it would be very simple to identify, anomalous characteristics in people, However, Clustering based approach has its own, problems, that is selecting an appropriate value of k as well as, what kind of clustering algorithm can be used, Would it be based on euclidean distance, such as K Means, or based on density or other characteristics or maybe hierarchical. Further computationally, Clustering would be expensive as compared to Statistic based approach.

(b). (10 points) Write down Grubbs outlier test procedure (including null and alternate hypothesis). For the following data (assume normally distributed), verify if the point (value = 57.0) is an outlier. Number of observations, n = 93; Mean = 52.2; Standard deviation (s) = 1.38. From the “G” table at n=93 and α = 0.05, the critical value = 3.18.

The procedure for Grubb's test is:

Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat

– H0: There is no outlier in data
– HA: There is at least one outlier
Grubb's test Statistic is $G = Max(|X – \overline{x}|)/s$
- Reject H0 if $G>((N-1)/sqrt(N))*sqrt(t^2_{(\alpha/(2 N),N-2)}/(N-2+t^2))$

In the above Algorithm, The test statistic that is used for the process of anomaly detection is G, Which is the Grubbs test Statistic,
X, is the value in dataset,
$\overline{x}$, is the mean
and s is the standard deviation.

The value of grubbs test statistic for the given value is:
(57-52.2)/1.38 = 3.478
Now
((N-1)/sqrt(N))*sqrt($t^2_{(\alpha/(2 N),N-2)}$/(N-2+t^2)) =
(92/9.643)*sqrt((3.18)^2/(91+(3.18)^2))
= 9.541*sqrt(10.1124/101.1124)
= 9.541 * 0.3162
= 3.016.

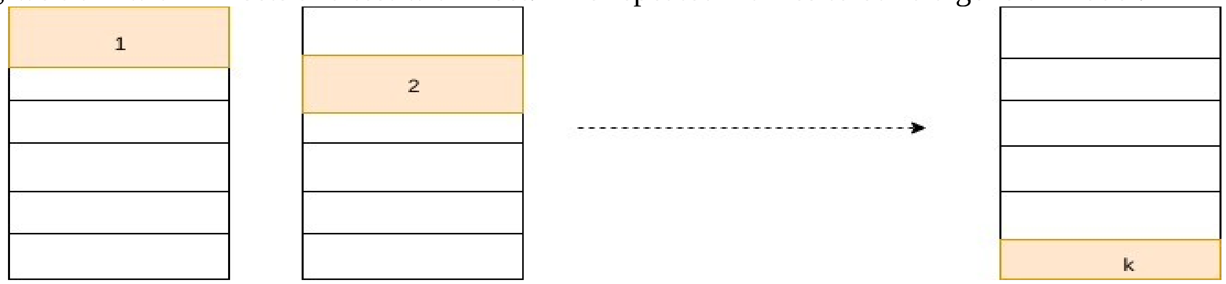
Since this value is less that Grubbs test Statistic, We reject the null hypothesis and conclude, That the given point is an outlier.

Q5. (12 points) (Classification and Cross Validation)
a) (2 points) Why do we use cross validation?
We use cross validation as a technique, to ensure that model remains general, that is to check if the model, will perform as well in real word testing as it performs for the training data. For this purpose we use a cross validation data set. If adding something to the model, it improves the accuracy of the model on cross validation dataset, we use it in the model, else we reject it as, it merely overfits on the data.

b) (2 points) Read about the bias-variance trade-off in model building. Explain what cross validation attempts to do for this.
The bias-variance trade-off in model building, is a tradeoff, which states, that is is very difficult to be both precise as well as accurate in a model used for data mining. Cross validation. Helps in reducing the bias, as it basically prevents the model from overfitting, to the training data, however, If the amount of data used for cross validation is small, then we don't have enough variance in data, so we arent sure if our test results on validation set is statistically significant or not.
For this purpose, we can use something like k fold cross validation to try to overcome this problem.

c) (2 points) Assertion: Cross validation is a way to do regularization (avoiding overfitting). Provide arguments for or against this assertion.
True, Cross validation is a technique, which allows us to prevent a model from overfitting to a particular dataset. This is because, the validation set acts as a general data, which the model has not yet seen for training, and acts as data that it might see later for testing when it is run in real world. Thus it prevents the model from being over, accurate for the training dataset and loosing its generality.

d) (2 points) Describe K- Fold cross validation with a simple example and graphic.
A k fold cross validation is basically a cross validation technique where in we divide the data into k sets, we train with k-1 sets and test with 1 set. This repeated k times to build a general model.



We can see, data divided into k sets, and each time, 1 is used as validation set
e) (2 points) LOOCV == K-Fold cross validation when k = ?
When k = number of data elements.

f) (2 points) Which type of cross validation would you use on a large dataset? K-Fold or LOOCV? Justify your answer.
On a larger data set, I would use K-Fold Cross validation, as the training data would still be large enough when we remove some values for cross validation, Further, The Validation set would also be large, to give statistically significant results.

Q6. (6 points) Based on the data given for Question 3 (Figure 3), if you are given the following two splits to consider while constructing the decision tree: Split 1: If Y < 2.5 then Class = +, else Class = - Split 2: If Y < 3.5 then Class = -, else Class = - Answer the following:

(a) (2 points) If you had to choose between Split 1 and Split 2, which one would you prefer using misclassification error as your impurity measure?
If the misclassification error, Then the misclassification error for split 1 is: 7/25= 0.28
& misclassification error of split 2 is: 7/25 = 0.28

Using misclassifcation error, both cases would be identical.

(b) (2 points) If you had to choose between Split 1 and Split 2, which one would you prefer using Gini as your impurity measure? Do not compute Gini, but rather answer based on your intuition, provide brief justification.
Using Gini Coefficient, The lower value of gini would be for the second split rather than the first split, This is because, the second split has one pure class node, and since gini uses a square factor, this purity of class node would get higher precedence in case of gini than misclassification error.

(c) (2 points) Based on answering the above two parts, answer why “misclassification error” is not used as impurity measure.
Misclassification error is not used, as it may give wrong results in certain cases, by giving lower preference to the actual information gained rather than just the error, which we can see in case of gini as well as entropy. Further if error is equal in 2 splits, it would give equal values, thus we will have to use gini or entropy in such scenarios.

Q7. (20 points) Apriori algorithm: Consider the dataset given in Table 7 and answer the sub-questions (a) and (b) using apriori algorithm.

(a)(10 points) Show (compute) each step of frequent itemset generation process using apriori algorithm, for support count of 2.

Items	Frequency
a	2
b	2
c	4
d	1

e	2
f	0

So frequent item sets for size 1 are:
{a},{b},{c},{e}

Items	Frequency
ab	0
ac	2
ae	2
bc	2
be	0
ce	2

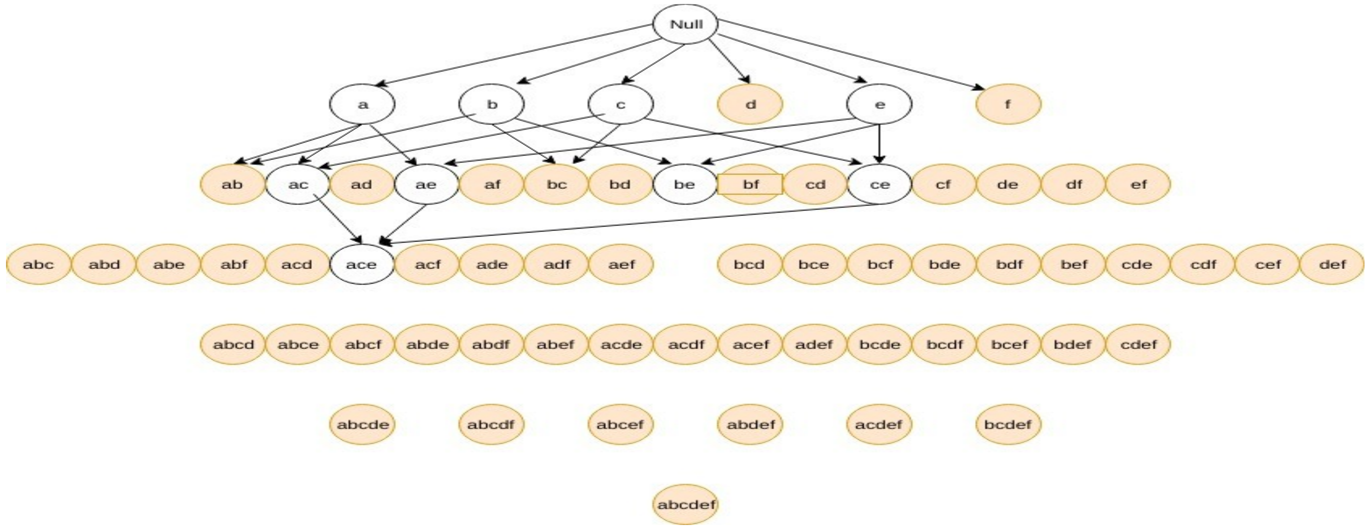
So frequent itemsets of size 2 are:
{ac},{ae},{bc},{ce}

Items	Frequency
ace	2

The only frequent itemset of size 3 is {ace}

Thus all frequent itemsets are {a},{b},{c},{e},{ac},{ae},{bc},{ce},{ace}

(b) (10 points) Show the lattice structure for the data given in Table 7, and mark the pruned branches if any



The branches that have not been drawn get pruned, Further, The nodes in white are frequent

Q8. (6 points) Accuracy Assessment You are asked to evaluate the performance of two classification models, M1 and M2, for a binary classification problem with classes '+' and '–'. For every test instance, x, each of the two models provides a posterior probability of x belonging to class '+'. Following table provides a list of 10 test instances with their true classes, and their posterior probabilities of belonging to class '+', according to M1 and M2. Assume that we are mostly interested in detecting instances from the positive class.

Suppose you choose a cutoff threshold to be t = 0.4 for both the models, M1 and M2. In other words, any test instance whose posterior probability is greater than t will be classified as a positive example. Compute the Precision, Recall, and F-Measure for M1 and M2 after using the cutoff threshold of t. Which model is better using Fmeasure as the evaluation criterion?

Model 1

Predicted

		+	-
Actual	+	3	2
	-	3	2

The precision is 3/6 = 0.5
The recall is 0.6
F measure is 2*(0.5*0.6)/1.1 = 0.5454

Model 2

Predicted

		+	-
Actual	+	3	2
	-	0	5

The precision is 1
The recall is 0.6
The F Measure is 2*0.6/1.6 = .75

Since model 2 has a higher value of F measure, Model 2 is the better model.

This is because, as + is the more important class, the recall for both the cases are same, however, the precision of model 2 better thus, we use Model 2.