Rishabh Sinha
CSC- 522
HW5
rsinha2@

1. Anomaly detection (10 points)
a) (1 point) How does the supervised anomaly detection problem relate to previously seen class imbalance topic?

The supervised anomaly detection problem is very similar to the previously seen class imbalance problem, As just like class imbalance problem, even in anomaly detection problem, We have few classes which maybe very small in size compared to the other classes, thus they can be considered as an anomaly in the given dataset. Thus the two are very similar in process but different in what the goal is, as in one you want to identify anomalous entities, while in other the goal is classifying the data precisely.

b) (1 point) Are point, contextual and collective anomalies mutually exclusive? If yes, elaborate and if not, provide one counter example.

No, point, contextual and collective anomalies are not mutually exclusive to one another, but just are dependent on how the data is represented. This is because, If we consider, the case of an irregular heartbeat. If we consider a heart beat as a single entity, Then the irregular beats can be considered as point anomaly. While if We plot the Heart rate as an ECG, then each point on a plot would be a data point, Thus now it is not a point anomaly but a collective anomaly. Further it is also a contextual anomaly as, in context of regular heart beats, it is different.

c) (2 points) Write down the procedure for Grubb's test for outliers. Explain what each symbol used denotes.
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 $G = \max(|X - \bar{x}|)/s$
• Reject H0 if $G > ((N-1)/\sqrt{N}) * \sqrt{t^2_{(\alpha/(2N),N-2)}/(N-2+t^2)}$

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,
$\bar{x}$, is the mean
and s is the standard deviation.

H0 is the null hypothesis, H1 is the alternate Hypothesis, H0 is rejected when G is greater that the above mentioned value. The terms in it are:
N is the number of data elements

t is the students-t-test statistic with N-2 degrees of freedom and significance level of $\alpha/(2\ N)$.

d) (2 points) Of the three main measures of central tendency, which ones are skewed by outliers? What challenge does this pose for unsupervised statistical outliers detection approaches? (Hint: Think about how a measure of central tendency could be a parameter of a distribution)

Of the three main measures of central tendencies, the most affected by outliers is the mean, because if a value is an extreme, it would severely affect the value of mean. While median and mode are not that much affected as they just represent the middle and the most common element in a dataset. When the value of mean gets skewed, it severely hampers the statistical tests which are highly dependent on the value of mean. Thus, in cases where values of mean is affected due to outliers, the model would not perform affectively.

e) (4 points) What are two strengths and two weaknesses of each of these outlier detection approaches? Answer in a tabular format such as:

Approach	Strengths	Weaknesses
Statistical outlier detection	- Firm mathematical foundation - can be efficient	- many times, data distribution may not be known -Anomalies can distort the parameters of distribution
Proximity-based outlier detection	- Simple to implement - Can use many types of distance measures according to requirement	-Expensive as $O(N^2)$ in time -Distance becomes less meaningful in high dimensions
Density-based outlier detection	- Simple to implement - Immune to globular shapes, can be used for data distributed with any shape	-Expensive as $O(N^2)$ in time -Density becomes less meaningful in high dimensions
Clustering-based outlier detection	- Simple - Many Techniques can be used	- outliers can distort clusters - Difficult to decide number of clusters

2. (10 points) Apriori Algorithm Using Table 2 given below, answer(a) and (b) based on the apriori algorithm.

Transaction ID	items
1	a,b,d
2	b,c,e
3	a,b,c,e
4	b,c,e

Transaction ID

a) (5 points) Show (compute) each step of frequent itemset generation process using apriori algorithm, for support count of 2.

Item sets of size 1:

Items	Count
a	2
b	4
c	3
d	1
e	3

The highlighted values are infrequent thus we apply apriori principle:

Item sets of size 2:

Items	Count
a, b	2
a, c	1
a, e	1
b, c	3
b, e	3
c, e	3

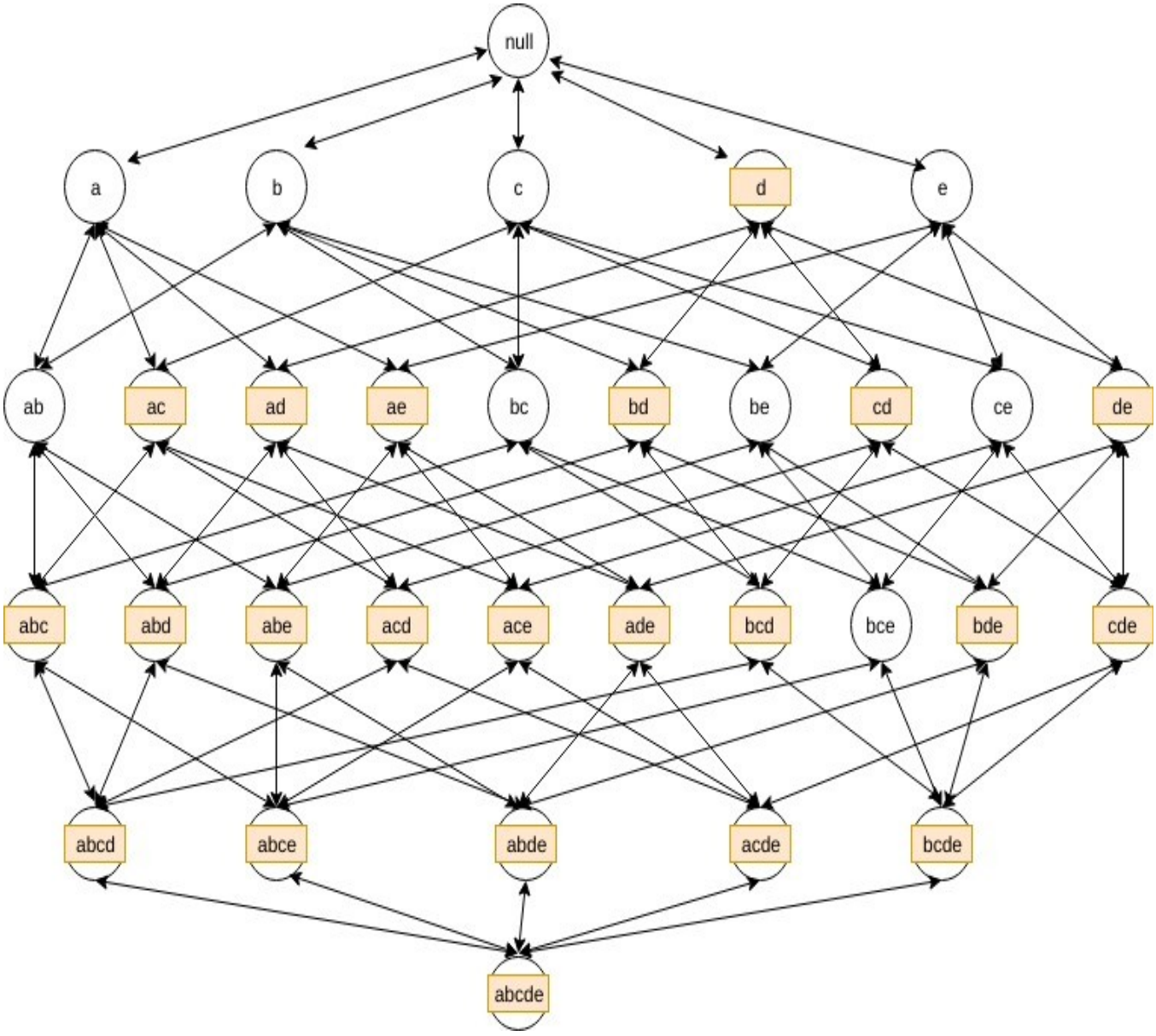
The Highlighted values are once again infrequent thus we again apply apriori principle

Item sets of size 3:

Items	Count
a, b, c	1
a, b, e	1
a, c, e	1
b, c, e	3

The Highlighted values are once again infrequent. Since we cant further expand as b,c,e is the only frequent itemset of size 3 remaining

b) (5 points) Show the lattice structure for the data given in Table 2, and mark the pruned branches if any



The Nodes which are shaded are the nodes which get pruned from the lattice structure because either they themselves or their sub sets have been found to be infrequent using apriori principle.

3. (10 points) Association Rule Mining For the dataset given in Table 3, compute the following:
a) for support threshold of 6 (by support count), list all frequent itemsets and maximal itemsets. (5 points)
The frequent itemsets are:
{E}, {F}, {EF}
and the maximal itemset is:
{EF}

b) Repeat (a) for support thresholds of 5 and 4. (5 points)
min support = 5;

frequent itemsets are
{C}, {E}, {F}, {EF}
maximal itemsets are:
{C}, {EF}

min support = 4;
frequent itemsets are
{C}, {E}, {F}, {J}, {EF}
maximal itemsets are:
{C}, {EF}, {J}

4. (10 points) SVM For the following 2-D training points:

Point ID	x1	x2	y
1	1	1	-1
2	1	0	1
3	0	1	1
4	0	0	-1