CSCI 5523 **Homework 3** **Due: Oct 29, 2019**

Please write your name, email and student ID #.

Question 1: GRADED by IOANNA

You are trying to evaluate two different blood tests, **T1 and T2**, that have been developed to detect a particular type of cancer. T1 had been evaluated on a population of 200 subjects, out of which 100 were known to be suffering from cancer, while the remaining 100 were healthy. T2 had been evaluated on a different population of 1100 subjects, out of which 100 were known to be suffering from cancer, while the remaining 1000 were healthy. The results of these tests are shown in the following confusion matrices, along with the values of the following evaluation measures: TPR, FPR, Precision, the F-measure, and TPR/FPR.

Test T1

Dataset: (100 patients)	**Predicted by Blood Test**	
Actual	**Cancer (+ class)**	**No Cancer (- class)**
Cancer (+ class)	40	60
No Cancer (- class)	10	90

Test T2:

Dataset: (1100 patients)	**Predicted by Blood Test**	
Actual	**Cancer (+ class)**	**No Cancer (- class)**
Cancer (+ class)	40	60
No Cancer (- class)	55	945

a. Calculate the TPR, FPR, Precision, F1 score and TPR/FPR for the two tests.

Test 1:TPR: 0.4, FPR: 0.1, Precision: 0.8, F1-Score: 0.533, TPR/FPr = 4

Test 2:TPR: 0.4; FPR: 0.055, Precision: 0.42, F1-Score: 0.41, TPR/FPR: 7.27

10 points: 1 point each

b. According to the F-measure, which test is better?

On the basis of F-measure, T1 is better

2 points

c. According to TPR/FPR, which test is better?

On the basis of TPR/ FPR, T2 is better

2 points

d. For this situation, which evaluation measure (between F1-Score and TPR/FPR) should you use to make your selection between the two tests, T1 or T2? Why?

Solution:TPR/FPR is the right choice since we don't know the class imbalance (also referred to as skew) in the population in which we will be using the test, and TPR/FPR is invariant to class imbalance. Also, in this case, T2 is strictly better than T1 since both have the same TPR and T2 has lower FPR. Hence for any value of class imbalance, T2 will outperform T1 on many reasonable measures.

6 points

- **Incorrect answer, no explanation: -5**
- **Incorrect answer, incorrect explanation: -4**
- **Correct answer, no explanation: -3**
- **Incorrect answer, with somewhat meaningful explanation: -3**
- **Correct answer, mistakes in explanation: -2, -1**

Question 2.

You are asked to evaluate the performance of two classification models, M1 and M2. The test set you have chosen contains 26 binary attributes, labeled as A through Z. Table 1 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(−) = 1 − P(+) and P(−|A, . . ., Z) = 1 − P(+|A, . . ., Z). Assume that we are only interested in detecting instances from the positive class.

Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.78	0.61
2	+	0.62	0.08
3	−	0.44	0.62
4	−	0.55	0.39
5	+	0.61	0.48
6	+	0.47	0.09
7	−	0.07	0.38
8	−	0.14	0.09
9	+	0.48	0.06
10	−	0.32	0.01

Table 1: Posterior probabilities

(a) Plot the ROC curve for both M1 and M2. (You should plot them on the same graph. Clearly hand-drawn plots are acceptable.) Which model do you think is better? Explain your reasons.

Answer:

The ROC curve for M1 and M2 are shown in Figure 1.

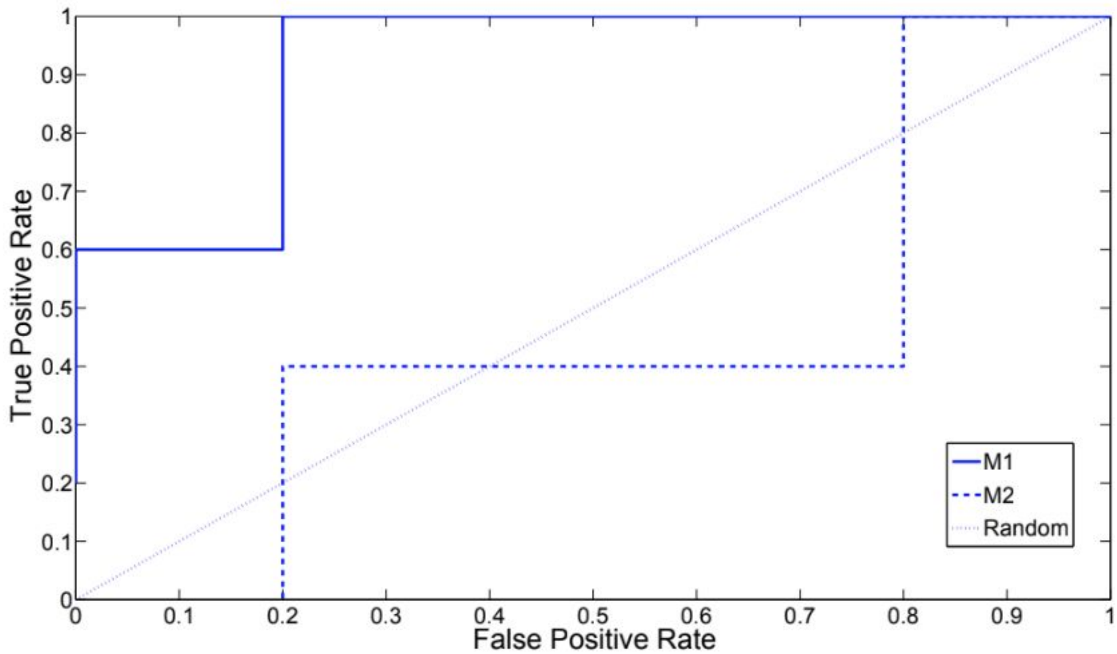


Figure 1: ROC curves.

M1 is better, since its area under the ROC curve is larger than the area under ROC curve for M2.

(b) For model M1, suppose you choose the cutoff threshold to be t = 0.5. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure obtained from the model at this threshold value.

Answer:

When t = 0.5, the confusion matrix for M1 is shown below.

		+	-
Actual	+	3	2
	-	1	4

Precision = 3/4 = 75%.

Recall = 3/5 = 60%.

F-measure = (2 × 0.75 × .6)/(0.75 + .6) = 0.667.

(c) Repeat the analysis for part (b) using the same cutoff threshold on model M2. Compare the F-measure results obtained from both the models. Which model is better? Are the results consistent with what you expect from the ROC curve?

Answer:

When t = 0.5, the confusion matrix for M2 is shown below.

		+	-
Actual	+	1	4
	-	1	4

Precision = 1/2 = 50%.

Recall = 1/5 = 20%.

F-measure = (2 × .5 × .2)/(0.5 + .2) = 0.2857.

Based on F-measure, M1 is still better than M2. This result is consistent with the ROC Plot.

(d) Repeat part (b) for model M1 using the threshold t = 0.1. Based on the F-measure, which threshold do you prefer, t = 0.5 or t = 0.1?

Answer:

When t = 0.1, the confusion matrix for M1 is shown below.

		+	-
Actual	+	5	0
	-	4	1

Precision = 5/9 = 55.55%.

Recall = 5/5 = 100%.

F-measure = (2 × .5555 × 1)/(.5555 + 1) = 0.715.

According to F-measure, t = 0.1 is better than t = 0.5.

(e) Which threshold, t = 0.5 or t = 0.1, would you use as per the results of the ROC analysis?

Answer: Using the contingency tables above, it can be seen that when t = 0.1, FPR = 1 and TPR = 0.8. On the other hand, when t = 0.5, FPR = 0.2 and TPR = 0.6. Since (0.2, 0.6) is closer to the point (0, 1), we favor t = 0.5.

(f) Are the choices of thresholds consistent based on the F-measure and ROC analyses? Briefly comment on why or why not?

Answer: The choices of thresholds based on the two analyses are inconsistent, primarily since F-measure and ROC are different ways of evaluating the performance of a classifier.
We can also show this by computing the area under the ROC curve
For $t = 0.5$, area $= 0.6 \times (1 - 0.2) = 0.6 \times 0.8 = 0.48$.
For $t = 0.1$, area $= 1 \times (1 - 0.8) = 1 \times 0.2 = 0.2$.
Since the area for $t = 0.5$ is larger than the area for $t = 0.1$, we prefer $t = 0.5$.

Question 3.

Answer the following questions:

a. Suppose you are given a data set consisting of nominal attributes, such as color, which takes values such as red, blue, green etc. Can you use this data set directly to train an SVM? If not, how will you transform a nominal attribute into a representation that can be used to train an SVM?

b. List a key similarity and a key difference between bagging and boosting.

Solution:

a. No, since SVM works naturally only with numerical valued data. So, the above data set can be used with an SVM by creating a binary attribute for each nominal attribute-value pair.
Alternative Solution: One can define a kernel for nominal attributes and then use it within an SVM.

b. Similarity: Both are ensemble methods, i.e., they combine the predictions from multiple classification models.

Differences: Boosting assigns a weight to each training example depending on the difficulty faced by the current classifier in classifying it, while bagging combines the predictions from classifiers trained on different samples of the training set in an unweighted fashion. Also, boosting is iterative, while bagging is not.

Question 4.

Answer the following questions for the data set in Table 2.

Table 2: Data set of market-basket transactions.

Transaction ID	Items Bought
1	$\{A, B, D, E\}$
2	$\{B, C, D\}$
3	$\{A, B, D, E\}$
4	$\{A, C, D, E\}$
5	$\{B, C, D, E\}$
6	$\{B, D, E\}$
7	$\{C, D\}$
8	$\{A, B, C\}$
9	$\{A, D, E\}$
10	$\{B, D\}$

(a) What is the maximum number of association rules that can be extracted from this data set (including rules that have zero support)?

Answer: There are five items in the data set. Therefore the total number of rules is

$3^5 - 2^6 + 1 = 180$

(b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

Answer: Because the longest transaction contains 4 items, the maximum possible size of a frequent itemset is 4.

(c) Calculate the maximum number of size-3 itemsets that can be derived from this data set.

Answer:

$\binom{5}{3} = 10$.

(d) Find an itemset (of size 2 or larger) that has the largest support.

Answer: {A, B}, {B, D}.

(e) Find a pair of items, say x and y, such that the rules $\{x\} \longrightarrow \{y\}$ and $\{y\} \longrightarrow \{x\}$ have the same confidence.

Answer:

Essentially, we want to find two items whose support in the data set is the same. Support of both E and C is 5.

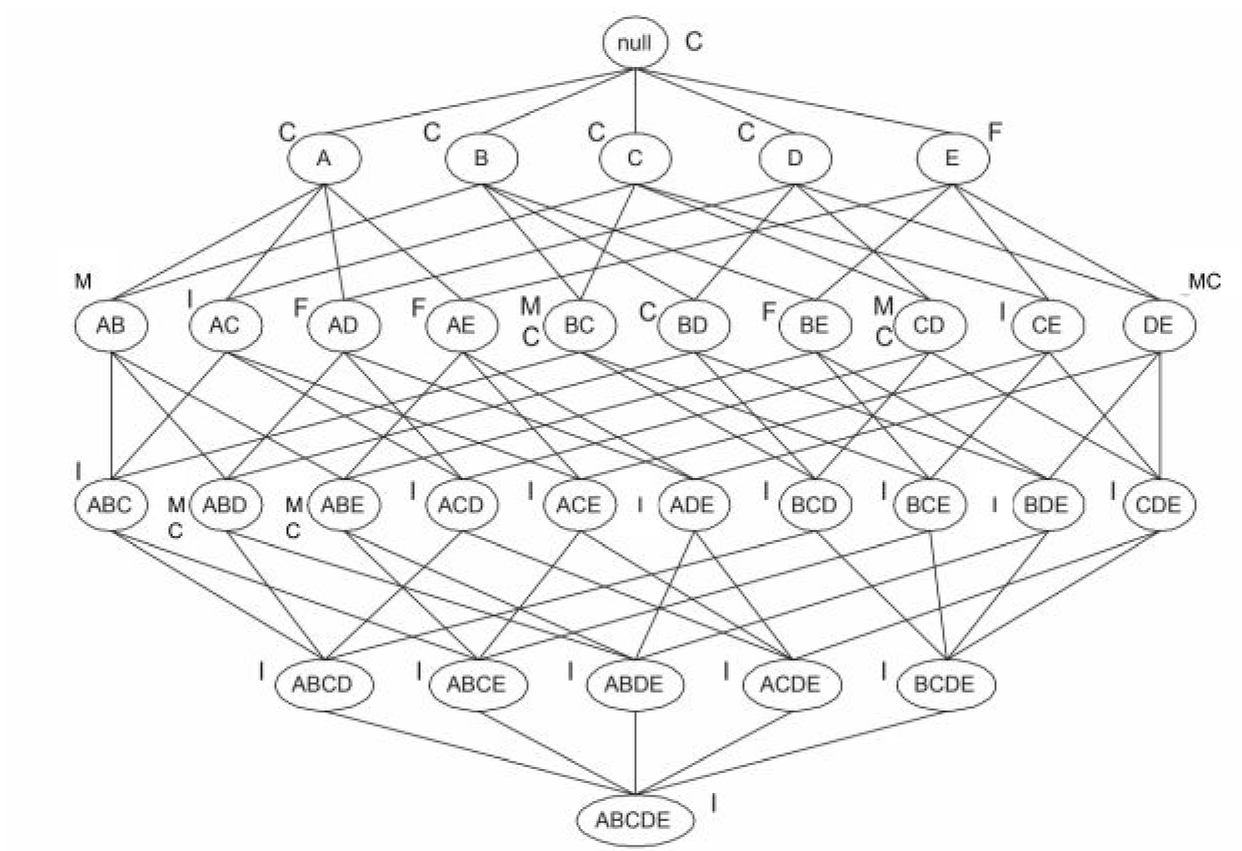
Question 5. GRADED by IOANNA

Draw the itemset lattice for the data set in Table 2. *Dataset of Market-basket transactions*. Now, given this lattice structure and the transactions given in Table 2, label each node in the lattice with the following letter(s):

- M if it is a maximal frequent itemset ,
- C if it is a closed frequent itemset,
- N if it is frequent but neither maximal nor closed, and
- I if it is infrequent.

Assume that the minimum support threshold is equal to 30%.

Answer: The lattice structure is shown below.



Generally:

- Correct lattice: 4 points
  - Partially correct lattice: -2
  - Incorrect lattice: -3
- Correct labels: 16 points
  - ∀ incorrect label: -0.5 point

Question 6.

Consider the following frequent 3-itemsets:
{a, b, c}, {p, b, c}, {p, a, b}, {p, a, c}, {p, a, w}
The book presents two algorithms for generating candidate 4-itemsets, the Fk−1 x F1 method (Page 369) and Fk−1 x Fk−1 method (Page 371).
As a pre-processing step, we first sort the items in all the itemsets following the lexicographical ordering.

a. List all the 4-itemsets that will be generated by Fk−1 x F1 candidate generation method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm.

{a,b,c,p}, {a,b,c,w}, {a,b,p,w},{a,c,p,w}, {b,c,p,w} After pruning{a,b,c,p} will survive.

b. List all the 4-itemsets that will be generated by Fk−1 x Fk−1 candidate generation method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm. Make sure you exploit lexicographical ordering.
{a,b,c,p}. After pruning{a,b,c,p} will survive.
Note: Most of the students did not exploit lexicographical ordering and ended up generating {p,a,b,w} and {p,a,c,w}.

c. Based on the list of candidate 4-itemsets generated above, is it possible to generate a frequent 5-itemset? State your reason clearly.

No. To generate a frequent 5 itemset that can survive after pruning step needs at least 5 frequent 4-itemsets. However, we only have one potential frequent 4-itemsets here.

Question 7

Consider the interestingness measure, M = [P(B|A)−P(B)] / [1−P(B)] , for an association rule: A → B

(a) What is the range of this measure? When does the measure attain its maximum and minimum values?

Range of the measure: $-\infty < M \leq 1$

Since M = [P(B|A)−P(B)] / [1−P(B)] maximum value occurs when P(A,B) = P(A):

At this value, M = [1−P(B)]/[1−P(B)] = 1

Minimum Value occurs when P(A,B)/P(A) = 0:

At this value M = −P(B)/[1−P(B)]

This value decreases with increasing P(B).

(b) How does M behave when P(A,B) is increased while P(A) and P(B) remain unchanged?

M increases.

(c) How does M behave when P(A) is increased while P(A,B) and P(B) remain unchanged?

M decreases.

(d) How does M behave when P(B) is increased while P(A,B) and P(A) remain unchanged?

Let a = P(A,B)/P(A) and y = P(B). Since P(A,B) and P(A) remain unchanged then a is a constant.

However, the measure M can be shown as a function

$f(y) = (a-y)/(1-y)$.

The derivative of f is: $\mathbf{df}(\mathbf{y})/\mathbf{dy} = (\mathbf{a}-1)/(1-\mathbf{y})^2$

Since a **<=** 1, M decreases while y = P(B) increases.

(e) Is the measure symmetric under the variable permutation?

No.

Question 8. Extra points - will be graded: (5 points)

Give an example scenario where you would reverse your choice of evaluation measure that you made in part (d) of Question 1? (That is, if you chose {TPR/FPR} in part (d), give an example of a scenario where you would prefer {the F measure} over {TPR/FPR} and vice-versa)

Solution: In a scenario where you know the skew in the target population, the F measure can be preferred over {TPR/FPR}, as it captures both precision and recall for the target population, and thus allows you to identify a test that does not compromise one of these for the other.

PRACTICE QUESTIONS

Question 1

You are given a classifier that attempts to predict whether it will rain tomorrow (+) or not (-). The confusion matrix below gives the results of this classification algorithm on a sample of 1000 consecutive days:

actual/ predicted	+	-
+	30	20
-	50	900

(a) Compute the accuracy, precision, recall, and F measure for the confusion matrix. (Compute precision, recall, and the F-measure with respect to + class only.)

Answer:

Accuracy = 930/1000 = 0.93
Precision = 30/(30+50) = 0.375
Recall = 30/(30+20) = 0.6
F-measure = $(2 \cdot 30)/(2 \cdot 30+50+20)$ = 60/130 = 0.46

(b) Which of these metrics is a poor indicator of the overall performance of your algorithm? Which of these metrics is a good indicator of the overall performance? Explain briefly why this is the case?

Answer:

Accuracy is a poor indicator of overall performance.
F-measure is a good indicator of overall performance.

There is a class imbalance problem (the + class is the minority), so accuracy is not a good indicator in this situation while the F-measure is an appropriate metric.

(c) Construct a trivial rule-based model which gives better accuracy than the classification algorithm above.

Answer:

{ } → -
Accuracy = 950/1000 = 0.95

Question 2

You are given a task to evaluate how well a new fire mapping algorithm works. The fire mapping algorithm is a Bayesian classifier which labels all the locations into two classes, burned and unburned. To evaluate the algorithm, two regions are tested. The confusion matrices of these two regions are given in Table 1 and Table 2.

Data set 1		Predicted class	
		Burned	Unburned
Actual Class	Burned	30	20
	Unburned	10	40

Data set 2		Predicted class	
		Burned	Unburned
Actual Class	Burned	30	20
	Unburned	1000	4000

a) Calculate the TPR, FPR, Precision and Recall of M for the “burned” class for both these data sets.

Answer:

TNR (DS1) = TN(DS1)/N(DS1) = 1 - FPR(DS1) = 1 - 0.2 = 0.8
TNR (DS2) = TN(DS2)/N(DS2) = ⅘ or is equal to: 1 - FPR(DS2) = 1 - 0.2 = 0.8

FPR (DS1) = 10/(10+40) = 0.2
FPR (DS2) = 1000/(1000+4000) = 0.2

Precision (DS1) = 30/(30+10) = ¾ = 0.75
Precision(DS2) = 30/(30+1000) = 3/103 = 0.2912

Recall (DS1) = 30/(30+20) = 0.6
Recall(DS2) = 30/(30+20) = 0.6

b) Is there a difference in their values for the two data sets? If so, what characteristic of the data sets (that are used to derive the above contingency tables) lead to the differences between the values of (TPR,FPR) and (Precision, Recall) that you observe above.

FPR and TPR(and also Recall) are same for the two datasets. However, Precision are different. The class skew between the two classes is significantly different. Since (TPR, FPR) are invariant to class skew, they don’t change. However, Precision is not invariant to class skew, and thus its value changes.

c) Compute Accuracy and F-measure with respect to ‘burned’ class for dataset 2.

Accuracy(DS2) = 4030/5050 = 0.798

F-measure (DS2) = 2*0.02912*0.6 /(0.02912+0.6) = 0.056

Question 3

Consider a data set with instances belonging to one of two classes - positive(+) and negative(-). A classifier was built using a training set consisting of equal number of positive and negative instances. Among the training instances, the classifier has an recall *m=50%* on the positive class and an recall of *n=90%* on the negative class.

The trained classifier is now tested on two data sets. Both have similar data characteristics as the training set. The first data set has 1000 positive and 1000 negative instances. The second data set has 100 positive and 1000 negative instances.

A. Draw the expected confusion matrix summarizing the *expected* classifier performance on the two data sets.

	Algorithm Output = (+)	Algorithm Output = (-)
True Label = (+)	1000*0.5	1000*0.5
True label = (-)	1000*0.1	1000*0.9

	Algorithm Output = (+)	Algorithm Output = (-)
True Label = (+)	100*0.9	100*0.5
True label = (-)	1000*0.1	1000*0.1

B. What is the accuracy of the classifier on the training set? Compute the precision, TPR and FPR for the two test data sets using the confusion matrix from part A. Also report the accuracy of the classifier on both data sets.

Training accuracy = (0.5+*0.9)/2

	Precision	TPR	FPR	Accuracy
Data set 1	0.5/0.1	0.5	0.1	(0.5+0.9)/2
Data set 2	0.5	0.5	0.1	(0.5+10*0.9)/11

C. In the scenario where the class imbalance is pretty high, how are precision and recall better metrics in comparison to overall accuracy? What information does precision capture that recall doesn't?

When the skew gets pretty high - i.e the total number of negatives are much more than the number of positives, the overall accuracy of the classifier is mainly dominated by the accuracy of the classifier on the majority negative class. From the curve in part C, we can see that the accuracy is almost constant after the skew increases beyond a certain point. If a classifier does well on the majority negative class but does poorly on the positive class, the overall accuracy wouldn't be able to capture this.

Precision is important to consider because it keeps track of the number of false positives of the algorithm and how significant are they in comparison to the true positives. If your algorithm flags 1000 instances as positive but only 100 of them are truly positives, it's performance is not good even if it captures all the truly positive instances, thereby giving it a high recall.

Question 4

Both Minimum Description Length (MDL) and the pessimistic error estimate are techniques used for incorporating model complexity. State one similarity and one difference between them in the context of decision trees.

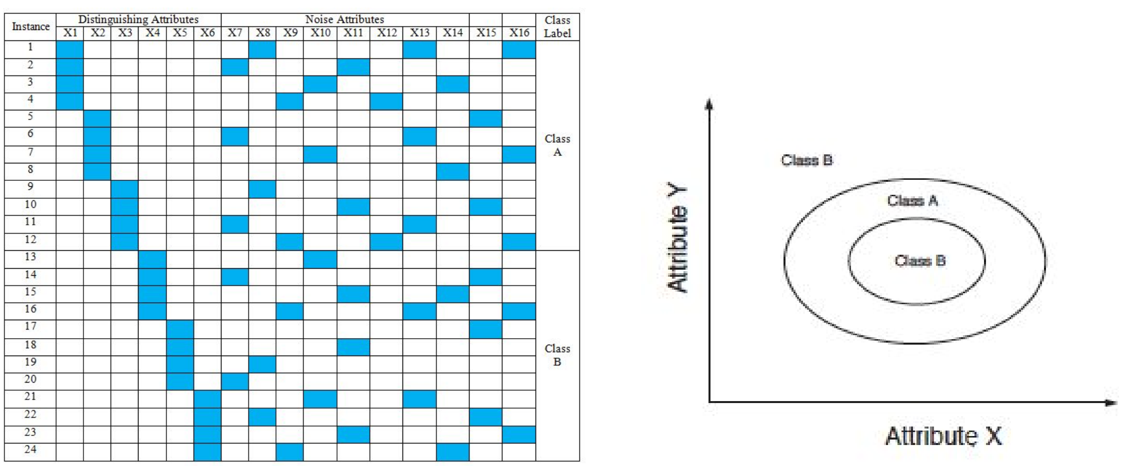
Answer:

One similarity between the two techniques is that they are both additive in nature, and add the complexity of the model to its accuracy in order to derive the overall complexity.

However, they are different, since MDL specifies a penalty for internal nodes and leaf nodes, while the pessimistic error estimate specifies a penalty for leaf nodes only. This has the implication that MDL can capture the depth of a tree while the pessimistic error estimate cannot. Another difference is that in MDL the cost of the model is a function of the number of attributes (the cost of each internal node is a function of the number of attributes), while for the pessimistic error estimate the cost of the model is independent of the number of attributes.

Question 5

Given the data sets shown in Figure 2, explain how the decision tree, naïve Bayes (NB), and k-nearest neighbor (k-NN) classifiers would perform on these data sets.



(a) Synthetic data set 1. **(b) Synthetic data set 2.**

Figure 2: Data sets for Question 5

Answer:

(a) Both decision tree and NB will do well on this data set because the distinguishing attributes have better discriminating power than noisy attributes in terms of entropy gain and conditional probability. k-NN will not do as well due to relatively large number of noise attributes, which will adversely affect the computation of similarity between two examples.

(b) k-NN will work the best due to the proximity of the examples of the same class to each other. NB does not work well for this data set since the attributes that determine the class boundaries are not independent. Decision tree will have to be large in order to capture the circular decision boundaries, and thus is not the ideal solution.

Question 6

Answer True or False and briefly explain.

(a) ANN is able to handle redundant attributes.

Answer: True. ANN can handle redundant attributes by assigning attribute weights.

(b) SVM is particularly effective for categorical data compared to other techniques such as decision trees.

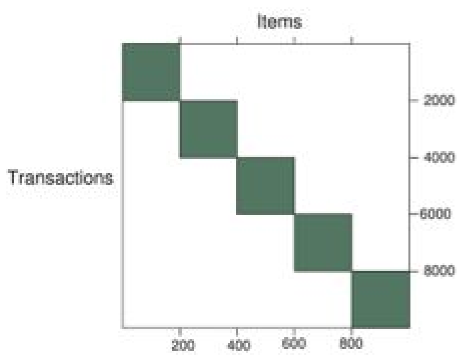
Answer: False. It is difficult to define a kernel on categorical attributes, while decision tree can handle categorical data well. Categorical data can be binarized and still used with SVMs but the way decision trees handle categorical data is more natural.

(c) SVM and Neural Network always produce the same decision boundary for a given data set with two classes.

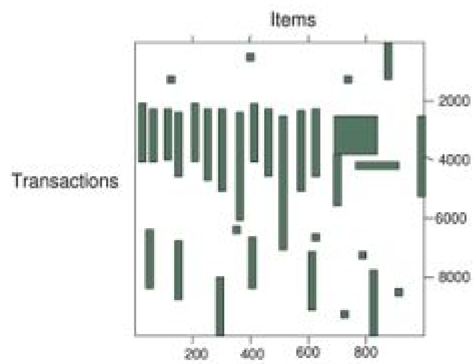
Answer: False. Although these two techniques could produce the same decision boundary, because of the different underlying approaches used, it is more likely that this would not happen.

Question 7

Answer the following questions based on the data sets shown in the figure below. Note that each data set contains 1000 items and 10000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup = 10% (i.e, itemsets must contain at least 1000 transactions). For ease of comparison, you can assume that the vertical bars in data set B are approximately of width 10 items each and the square block centered around item 800 is of the same size as each of the 5 square blocks in data set A.



Data set A



Data set B

1. Which data set will produce the greatest number of frequent item sets?

2. Assume that the minimum support threshold is equal to 10%. How many closed frequent itemsets will be discovered from data set 1?

3. Which data set will produce the longest frequent itemset?

4. Which data set will produce frequent itemset with highest support?

5. Which data set will produce the most number of closed frequent itemsets?

Answer:

1. **Graph (b)**. Graph (a) will produce 5 × - 4 frequent itemsets. On the other hand, there are at least 340 items that are supported by transactions 2000-4000, and have a support greater than 10% in the data set of Graph (b). So, in the latter case, frequent itemsets can be produced, which is much larger than those produced from Graph (a).
2. **5**, since each of the diagonal blocks will be produced as a closed itemset. Each of their constituent itemsets will have the same support as the diagonal block they are a part of.
3. **Graph (b)**, since it will produce a frequent itemset of length 340 or more that is supported by transactions 2000-4000. On the other hand, graph (a) will produce a frequent itemset of length at most 200.
4. **Graph (b)**, which has an itemset constituted by items in the range 400-600 and has a support count of over 4000 (40%). This is larger than the maximum support of 20% that an itemset in Graph (a) can have. Thus, Graph (b) will produce itemsets of higher support.

5. **Graph (b)**, since Graph (a) will produce 5 closed frequent itemsets. On the other hand, since Graph (b) will produce many more frequent itemsets of different sizes and supports, it will also produce more closed frequent itemsets.