**Homework 2 (Walker Basham, Cole Gannaway, & Jack Hawblitzel)**

**Question 1 (14 points)**

Consider the dataset shown in Table 1.1 for a binary classification problem.

Customer ID Housing Type Gender Marital Status Class

1 Apartment Male Married C0

2 House Male Single C1

3 House Female Married C1

4 Apartment Female Single C0

5 Apartment Male Married C0

6 Hostel Male Single C1

7 House Female Married C1

8 Apartment Female Single C0

9 Apartment Male Married C0

10 House Male Single C1

11 Hostel Female Married C1

12 House Female Single C0

13 House Male Married C0

14 Hostel Male Single C1

15 Hostel Female Married C1

16 Apartment Female Single C0

Table 1.1

1. Compute the entropy for the overall data.

Pc1 = 8/16

Pc2 = 8/16

Entropy = .3

1. Compute the entropy obtained for each of the four attributes (consider a multi-way split using each unique value of an attribute).

Both split evenly: .3

Gender split: .3

Apartment 6/16: .28

House 6/16: .28

Hostel 4/16: .24

Weighted housing: .27

1. Compute the Information Gain (IG) obtained by splitting the overall data using each of the four attributes. Which attribute provides the highest IG, and which attribute provides the lowest IG.

Marital status = 0

Gender = 0

Housing = .03

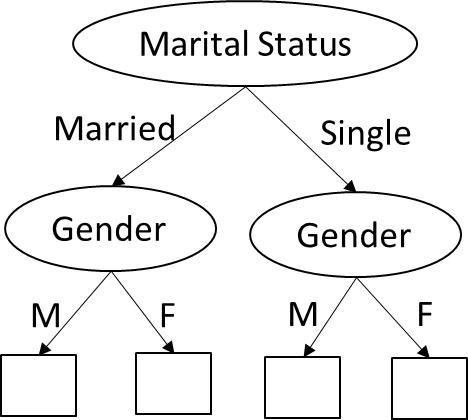
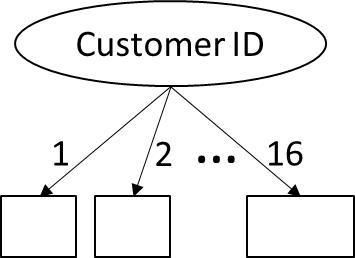
Housing provides more information than anything

1. For splitting at the root node, would you choose the attribute that provides the maximum IG? Briefly explain your choice.

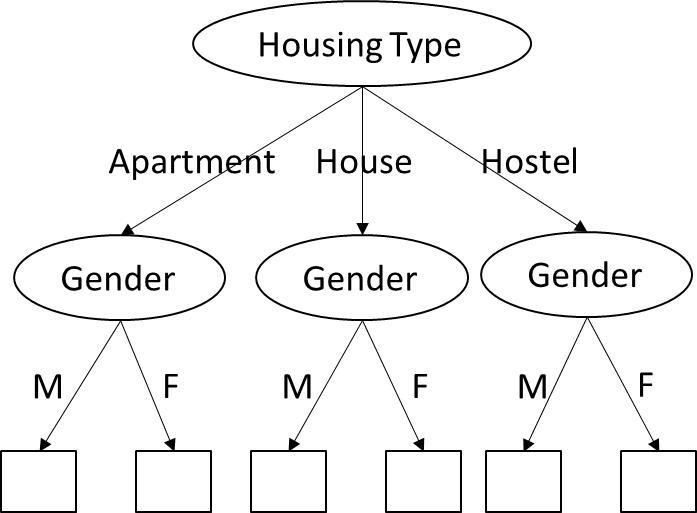
The most IG is housing, but housing also could result in a lot of splitting. To avoid overfitting we would not choose housing.

f. Consider the following 3 decision trees:

Tree 1



Tree 2



Tree3

Compute the difference between the entropy of overall data with the weighted entropy of the leaves for each of the three trees. Based on these differences, which tree would you choose for performing classification? Is the attribute chosen at the root of this tree same as the attribute chosen for splitting in (e)? Briefly comment on the nature of your results, and the properties of the impurity measure used while constructing decision trees.

**Question 2 (16 points)**

Consider the dataset shown in Table 4.1.

Instance A B C Class

1 0 0 1 -

2 1 0 1 +

3 0 1 0 -

4 1 0 0 -

5 1 0 1 +

6 0 0 1 +

7 1 1 0 -

8 0 0 0 -

9 0 1 0 +

10 1 1 1 + Table 4.1

a. Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|−), P(B = 1|−), and P(C = 1|−),

P(A = 1|+) = 3/5 = 0.6

P(B = 1|+) = 2/5 = 0.4

P(C = 1|+) = 4/5 = 0.8

P(A = 1|-) = 2/5 = 0.4

P(B = 1|-) =  2/5 = 0.4

P(C = 1|-) =  1/5 = 0.2

b. Use the conditional probabilities in part (a) to predict the class label for a test sample (A =

1, B = 1, C = 1) using the naïve Bayes approach.

We have R: (A=1,B=1,C=1) is test record. Compute P(+|R) and P(-|R).

Bayes: P (+|R) = P (R|+)P (+)/P (R) and P (−|R) = P (R|−)P (−)/P (R)

A B C Class: 1 0 0 1 − 2 1 0 1 + 3 0 1 0 − 4 1 0 0 − 5 1 0 1 + 6 0 0 1 + 7 1 1 0 − 8 0 0 0 − 9 0 1 0 + 10 1 1 1 +

P (R|+) = P (A = 1|+) × P (B = 1|+) × P (C = 1|+) = 0.192

P (R|−) = P (A = 1|−) × P (B = 1|−) × P (C = 1|−) = 0.032

P(R|+) greater than P(R|-), we assign record to + class

c. Compare P(A = 1, B = 1|Class = +) against P(A = 1|Class = +) and P(B = 1|Class = +). Are the variables conditionally independent given the class?

Look at P (A = 1, B = 1|+) = 0.2 vs. P (A = 1|+) = 0.6 and P(B = 1|Class = +) = 0.4

Because P (A = 1|+) \* P (A = 1|−) ≠ P (A = 1, B = 1|+),

A & B not conditionally independent

d. Let us consider the data instance (A=1, B=1, C=1). Compute the probability of this instance belonging to Class = + using

i. no attributes( i.e. calculate prior probability)

ii. attribute A [ P(Class = +|A=1) ]

iii. attributes A and B [ P(Class = +|A=1, B=1) ]

iv. attributes A, B and C [ P(Class = +|A=1, B=1, C=1) ]

Comment on the change in probability values as we proceed from (i) to (iv).

Now, consider the data shown in Table 4.2.

e. Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|−), P(B = 1|−), and P(C = 1|−) using table 4.2.

P(A = 1|+) = 0/5 = 0.0

P(B = 1|+) = 2/5 = 0.4

P(C = 1|+) = 4/5 = 0.8

P(A = 1|-) = 2/5 = 0.4

P(B = 1|-) =  0/5 = 0.0

P(C = 1|-) =  1/5 = 0.2

f. For a new data instance, **x** = (A = 1, B = 1, C = 1), compute the posterior probabilities, P(+|x) and P(-|x) using the naıve Bayes approach.

g. What kind of problems will you encounter in predicting the class of **x** using the posterior probabilities computed in (e) and how can you resolve them?

Instance A B C Class

1 0 0 1 -

2 0 0 1 +

3 0 0 0 -

4 1 0 0 -

5 0 0 1 +

6 0 0 1 +

7 1 0 0 -

8 0 0 0 -

9 0 1 0 +

10 0 1 1 +

**Table 4.2**

**Question 3 (6 points)**

**Classifier Comparison** In each of the classification scenarios listed below, you are given a set of classifiers and a description of the classification scenario. For each scenario, state the choice of the classifier that is best suited for the dataset along with a brief explanation supporting your answer.

1. Scenario: Data contains some missing values for certain attributes in the training and test data. Classifiers available: ANN, Naïve Bayes, KNN

To avoid problems with the missing information, we would utilize the most basic form, KNN.

1. Scenario: Many of the attributes are irrelevant (contain no information about the class). Classifiers available: KNN, ANN, Decision Trees

Decision tree’s handle irrelevant information the best.

1. Scenario: Dataset contains attributes that are not discriminative by themselves, but are discriminative in combination. Classifiers available: KNN, Naïve Bayes, Decision Trees (taking single attribute at a time)

We want to process one attribute at a time so we classify using Naïve Bayes because operates linearly.

**Question 4 (6 points)**

**Classification**. You are given a classification dataset with 100 instances, which has been partitioned into two subsets, dataset A with 50 instances and dataset B with 50 instances. Dataset A is used for training and dataset B is used for testing. You are supposed to compare two classification models: Model 1, which is an unpruned decision tree, and Model 2, which is a pruned version of decision tree. The accuracy of the two classification models on datasets A and B are shown in Table 7.1

Classification Accuracy Dataset A Dataset B

Model 1 0.98 0.72

Model 2 0.82 0.8

Table 7.1 Classification accuracies of the two models using dataset A as the training set.

1. Based on the accuracies shown in Table 7.1, which classification model would you expect to have better performance on unseen instances? Support your answer with a brief explanation.

Based on the consistency of the accuracy figures, Model 2.

1. Now, you tested Model 1 and Model 2 on the entire dataset (A + B) and found that the classification accuracy of Model 1 on dataset (A + B) is 0.85, whereas the classification accuracy of Model 2 on the dataset (A + B) is 0.81. Based on this new information and your observations from Table 7.1, which classification model would you finally choose for classification? Provide a brief explanation.

Because model 1’s accuracy for A was greater and in B it was somewhat close, we conclude a combined better accuracy in 1.

**Question 5 (14 points)**

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 accuracy m on the positive class and an accuracy of n 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.

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

|  |  |  |
| --- | --- | --- |
| Data set 1 | Actual + | Actual - |
| Predicted + | m\*1000 | (1-n)\*1000 |
| Predicted - | (1-m)\*1000 | N\*1000 |

|  |  |  |
| --- | --- | --- |
| Data set 2 | Actual + | Actual - |
| Predicted + | M\*100 | (1-n)\*1000 |
| Predicted - | (1-m)\*100 | N\*1000 |

1. 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.

|  |  |  |
| --- | --- | --- |
|  | Data set 1 | Data set 2 |
| Precision  Tp/(tp+fp) | M\*1000/(m\*1000 + ((1-n)\*1000)) | M\*1000/(m\*1000 + ((1-n)\*1000)) |
| TPR  TP/(TP+FN) | M\*1000/(m\*1000 + ((1-m)\*1000)) | M\*100/(m\*100 + ((1-m)\*100)) |
| FPR  FP/(FP+TN) | (1-n)\*1000/((1-n)\*1000 + n\*1000) | (1-n)\*1000/((1-n)\*1000 + n\*1000) |
| accuracy | (M\*1000+n\*1000)/( M\*1000+n\*1000+(1-m)\*1000+(1-n)\*1000) | (M\*100+n\*1000)/( M\*100+n\*1000+(1-m)\*100+(1-n)\*1000) |

1. i). If the skew in the test data - the ratio of the number of positive instances to the number of negative instances, is 1:s, what is the accuracy of the algorithm on this data set? Express your answer in terms of s, m, n.

data set 1: (m+n\*s)/( m+n\*s+(1-m)+(1-n)\*s)

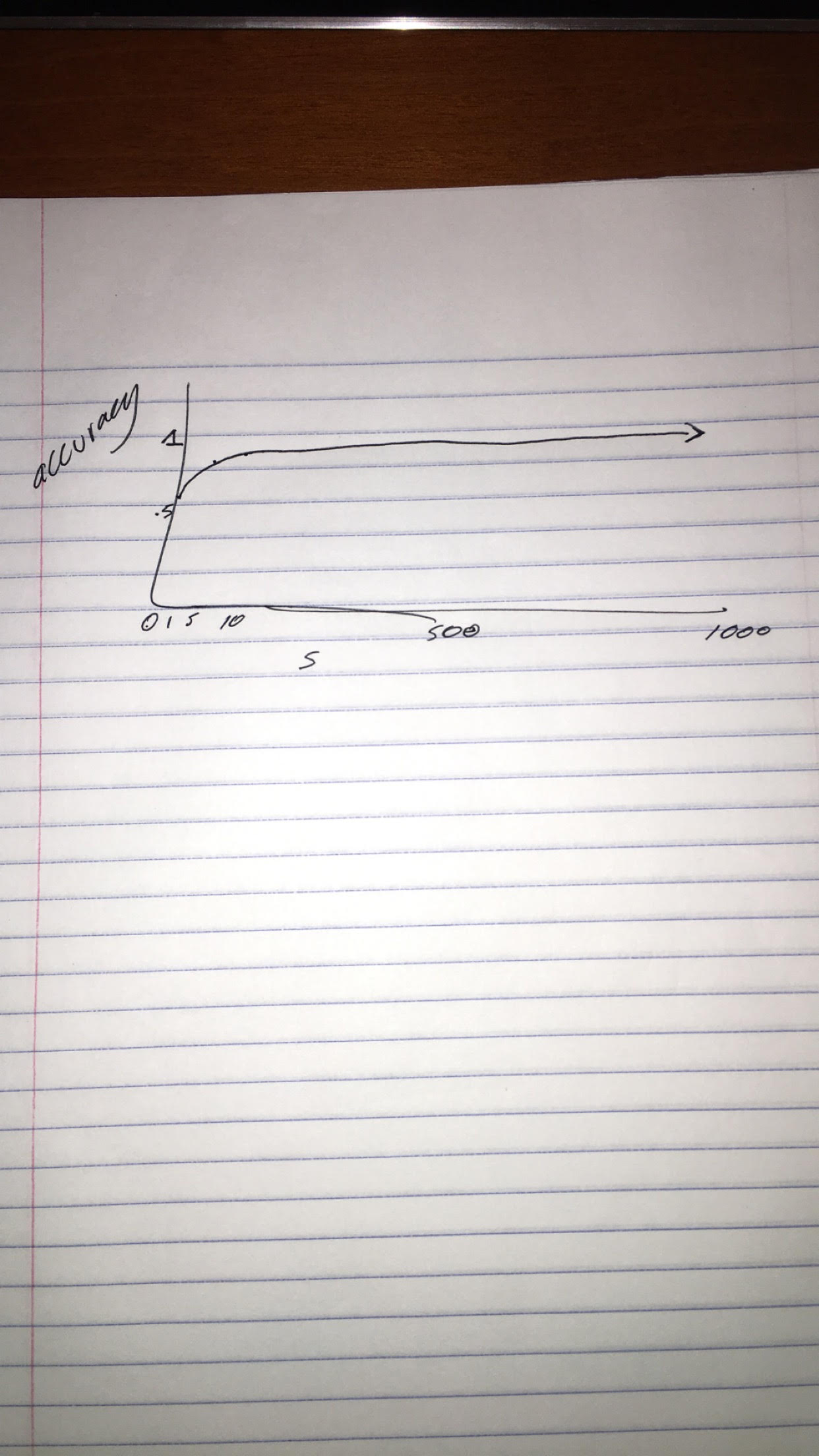
data set 2: (m+n\*s)/( m+n\*s+(1-m)+(1-n)\*s)

ii). In the expression for overall accuracy obtained from part (i), fix the values of m, n at

m = 0.6, n = 0.95. Now compute the accuracy for the following values of s: 0.0001,

0.005, 0.1, 0.5, 1, 5, 10, 500, 10000. Plot the obtained values keeping accuracy on the vertical axis and value of s on the horizontal axis. Also, plot horizontal lines corresponding to accuracy = m and accuracy = n on this graph for reference.

|  |  |  |
| --- | --- | --- |
| s | Data set 1 | Data set 2 |
| .0001 | .6 | .6 |
| .005 | .6 | .6 |
| .1 | .63 | .63 |
| .5 | .7 | .7 |
| 1 | .775 | .775 |
| 5 | .89 | .89 |
| 10 | .91 | .91 |
| 500 | .949 | .949 |
| 1000 | .949 | .949 |



iii). What value does the overall accuracy approach to if s is very large (>>1)? And when

s is very small (<<1)?

The values seem to approach a limit .95 when s is big, and a approach .6 when s is small

1. In the scenario where the class imbalance is pretty high (say, s>500 for part C), how are precision and recall better metrics in comparison to overall accuracy? What information does precision capture that recall doesn’t?

Precision is good because it calculates only how good the model is at capturing the correct value, instead of taking the negatives into account.

**Question 6 (6 points)**

Answer the following questions on classification:

1. 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 these attributes into a representation that can be used to train an SVM?

No because the data set for SVMs must be binary attributes. I would transform these attributes by creating binary attributes for each nominal attribute such as red, blue, or green.

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

In boosting, weights may change at the end of each boosting round, and this does not happen in bagging.

Both bagging and boosting are ensemble techniques where a number of weak learners combine to create a strong learner for accurate predicitions.

**Question 7 (9 points)**

Answer the following questions based on the Bayes Theorem:

1. Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 25%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

P(G|S) = P(S|G) \* P(G) / P(S) =

0.25 \* 0.2 / 0.8 = 0.0625

1. Given the information in part (a), is it more likely for a randomly chosen college student to be a graduate student or an undergraduate student?

It is more likely for a college student to be a undergraduate student because it is 80% compared to 20% of the students who are graduate students.

1. Repeat part (b) assuming that the randomly chosen student is a smoker.

Let G = graduate U = undergraduate and S = smoker.

P(S) = P(S&G) + P(S&U) =

P(S|G) \* P(G) + P(S|U) \* P(U) =

0.25 \* 0.2 + 0.15 \* 0.8 = 0.05 + 0.12 = 0.17

**Question 8 (11 points)**

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 the two tables below respectively.

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

1. Calculate the TPR, FPR, Precision and Recall of M for the + class for both these data sets.

Data Set 1:

TPR = 30 / 30 + 20 = 30 / 50 = 0.6

FPR = 10 / 10 + 40 = 10 / 50 = 0.2

Precision = 30 / 30 + 10 = 30 / 40 = 0.75

Recall = 30 / 30 + 20 = 30 / 50 = 0.6

Data Set 2:

TPR = 30 / 30 + 20 = 30 / 50 = 0.6

FPR = 1000 / 1000 + 4000 = 1000 / 5000 = 0.2

Precision = 30 / 30 + 1000 = 30 / 1030 = 0.02913

Recall = 30 / 30 + 20 = 30 / 50 = 0.6

1. 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.

The only difference between the values of the two data sets is the precision. This is because in data set 2 there are a lot more False Positive values than in data set 1, and the True Positive values are the same in both sets. Because the ratio of False positive to True negative values remained the same in both data sets, the FPR remained the same.

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

Accuracy = 30 + 4000 / 30 + 20 + 1000 + 4000 = 4030 / 5050 = 0.798

F-measure = 2 \* (Precision \* Recall)/(Precision + Recall) =

2 \* (30 / 1030 \* 30 / 50) / (30 / 1030 + 30 / 50) =

2 \* (900 / 51,500) / (32,400/51,500) = 2 \* (900/32,400) = 0.05556

(d) Comment on the performance of the given algorithm. Your answer should include 1)

which evaluation metric (F-measurement or accuracy) is chosen. 2) Why you choose it.

3) What is your conclusion (Does the new algorithm work well or not) based on the

evaluation metric.

You should choose an F-measurement because accuracy produces does not differentiate between false positives or false negatives. The new F-measurement algorithm works well on one class, but does not even consider false negatives.

(e) Is it possible to construct a trivial rule-based classifier with better i) F-measure? ii)

accuracy? If yes, construct the classifier.

No, because it typically depends on the type of information given and which data is more or less important than others. For example, a false positive where a doctor unnecessarily prescribes a common cold has massively different implications than if a doctor unnecessarily prescribes a patient heart surgery.

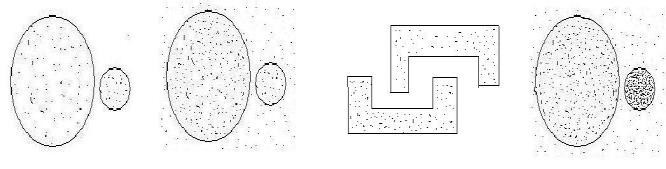
**Question 9 (12 points)**

How will single-link, complete-link, and DBSCAN will perform for following cases? The points are evenly distributed for first three cases (a-c), while the last case (d) has one dense cluster with 50000 points and one relatively sparse cluster with 50 points with noise data points in between. Assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points (case b,d).

Single-link would work well in a and c because there is no noise for it to incorrectly link clusters. However, in both b and d, single-link could possible misidentify clusters because of their proximity.

Complete link would work well in a,b, and d because it forces spherical with consistent diameters but would not work well in c because it uses the furthest distance and is not spherical.

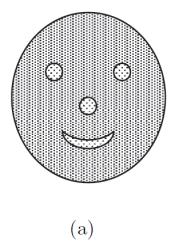
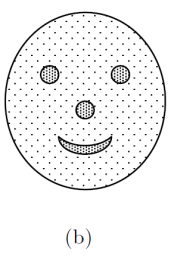
DBSCAN will work well for a-c because the points are evenly distributed but will not work well in case d because of the varying densities separated by spaces.



(a) (b) (c) (d)

**Question 10 (6 points)**

Consider the following two sets of points (faces) shown in figures (a) and (b). The darker shading indicates a denser point distribution.

1. For each figure, could you use DBSCAN to find clusters corresponding to the patterns represented by the nose, eyes, and mouth? Explain.

Yes, for figure A, because the dense points are connected. For figure B, DBSCAN would not work well because the dense points are not close in proximity to one another.

1. For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain.

Yes, you could use K-means by selecting the correct centroids in the nose, eyes, and mouth, but because these areas are small relative to the rest of the face, it would be extremely difficult to correctly guess these centroids.