ALDA Fall 2018
HW 2
9/19/2018

HW2 contains 6 questions. Please read and follow the instructions.

- **DUE DATE FOR SUBMISSION: 10/3/2018 11: 45 PM**
- **TOTAL NUMBER OF POINTS: 140**
- Make sure you clearly list each team member's **name and Unity ID** at the top of your submission. One submission per group.
- Your submission should be a **single zip file** containing a PDF of your answers, your codes, and a readme file with running instructions. Please follow the naming convention for your zip file: H(homework group number)_HW(homework number), e.g. H1_HW2.

---

1. (40 points) [**PCA**] [**Xi Yang**] In this problem, you will perform a PCA on the provided training dataset ("hw2q1_train.csv") and the testing dataset ("hw2q1_test.csv"), which come from the Connectionist Bench Dataset (`http://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks)`). In both datasets, each row represents a data point or sample. The first 60 columns are input features, and the last column "Class" is the output label, with the letters "R" and "M" indicating if a sample is a Rock or a Mine, respectively.

   Write code in Matlab, R or Python to perform the following tasks. Please *report your outputs and key codes* in the document file, and also include your code file (end with .m, .r or .py) in the .zip file.

   (a) (2 points) Load the data. Report the size of the training and testing sets. How many Rock (R) and Mine (M) samples are in the training set and the testing set, respectively?

   (b) (18 points) **Preprocessing Data-Normalization**: Please run normalization on all input features in both the training and testing datasets to obtain the *normalized* training and the *normalized* testing datasets. (**Hint:** you will need to use the *min/max* of the training dataset to normalize the testing dataset and do NOT normalize the output "Class" of data.)

   Use the **NEW** normalized datasets for the following tasks :

   i. (2 points) Calculate the covariance matrix of the *NEW* training dataset.

   ii. (2 points) Calculate the eigenvalues and the eigenvectors based on the covariance matrix in (i) above. Report the size of the covariance matrix and the 5 largest eigenvalues.

iii. (1 point) Display the eigenvalues using a bar graph or a plot, and choose a reasonable number(s) of eigenvectors. Justify your answer.

iv. (13 points) Next, you will combine PCA with a *K-nearest neighbor (KNN)* classifier. More specifically, PCA will be applied to reduce the dimensionality of data by transforming the original data into p ($p \leq 60$) principle components; and then KNN ($K = 3$, euclidean distance as distance metric) will be employed to the p principle components for classification (third-party packages are allowed to use for KNN).

- (5 points) Report the accuracy of the *NEW* testing dataset when using PCA ($p = 10$) and the 3NN classifier. To show your work, please submit the corresponding csv file (including the name of csv file in your answer below). Your csv file should have 12 columns: columns 1-10 are the 10 principle components, column 11 is the original ground truth output "Class", and the last column is the *predicted* output "Class".
- (6 points) Plot your results by varying p: 2, 4, 8, 10, 20, 40, and 60 respectively. In your plot, the x-axis represents the number of principle components and the y-axis refers to the accuracy of the *NEW* testing dataset using the corresponding number of principle components and 3NN.
- (2 point) Based upon the PCA +3NN's results above, what is the **most "reasonable" number** of principle components among all the choices? Justify your answer.

(c) (18 points) **Preprocess Data-Standardization**: Similarly, please run standardization on all input features to obtain the *standardized* training and the *standardized* testing datasets. Then repeat the four steps i-iv in (b) above on the two **NEW** *standardized* datasets.

(d) (2 points) Comparing the results from (b) and (c), which of the two data-processing procedures, normalization or standardization, would you prefer for the given datasets? And why? (Answer without any justification will get zero point.)

SOLUTIONS:

(a) (2 points) Training samples: 156
Testing samples: 52

Rock in training samples: 73
Mine in training samples: 83
Rock in testing samples: 24
Mine in testing samples: 28

(b) (13 points) i. (2 points) See the attached code, 'hw2q1.py'.

ii. (2 points) See the attached code, 'hw2q1.py.
Size of covariance matrix: 60 * 60

5 largest eigenvalues for normalized data:

[0.69477425, 0.45719507, 0.22325971, 0.19134231, 0.11789008]

5 largest eigenvalues for standardized data:
[12.4294178, 11.55798134, 4.97913851, 3.34509159, 3.22556626]

iii. (2 points) There are many ways to select the reasonable number of eigenvectors. Here we provide one of the methods as sample solution. Referring to the Figure 1 and Figure 3, which are the Eigenvalue vs. the number of principle components, we can choose the number where a big gap or elbow happens in the plot. For normalized data, the reasonable number is between 5 and 15 or around the range; For standardized data, the reasonable number is between 5 and 25 or around the range.

iv. (7 points) See the attached code, 'hw2q1.py.

- (2 points) When choosing 10 principle components:

  For normalized data, accuracy = 0.9038

  For standardized data, accuracy = 0.9038

- (3 points) For normalized data, see Figure 2.
  best accuracy = 0.9038 with 10 principle components

  For standarized data, see Figure 4.
  best accuracy = 0.9423 with 20 principle components

- (2 points) The best number of dimensions is the number of between 8 and 15 for normalized data, and between 10 and 30 for standardized data.

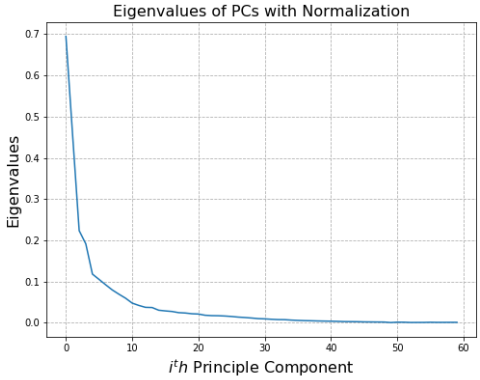


Figure 1: Eigenvalues of PCs for normalization data

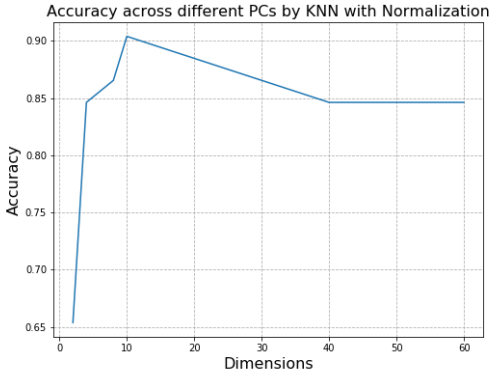


Figure 2: Accuracy of PCs by KNN for normalization data

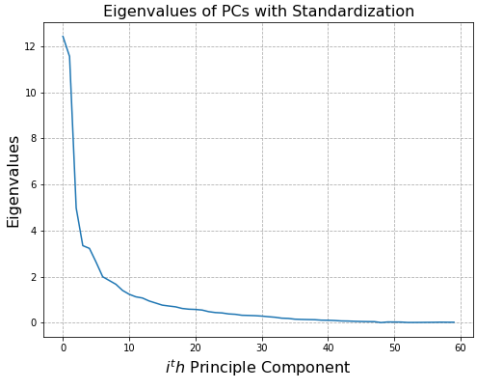


Figure 3: Eigenvalues of PCs for standardized data

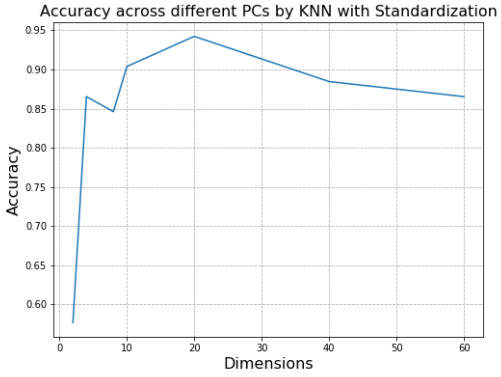


Figure 4: Accuracy of PCs by KNN for standardized data

(c) (13 points) Please refer the answers in (b).

(d) (2 points) Comparing the results in (b) and (c), standardization is preferred for the given datasets. Comparing to the normalization, it generally has better performance when fixing the classifier as 3NN.

2. (20 points) [**Decision Tree**][**Song Ju**] In the given "hw2q2.csv", all of the input features are nominal except for the first column, which is a ratio and continuous. The output label has two class values: T or F. Complete the following tasks using the decision tree algorithm discussed in the lecture. In the case of ties, break ties in favor of the leftmost feature. (You can hand-draw all of your trees on paper and scan your results into the final pdf.)

   (a) (10 points) Construct the tree *manually* using ID3/entropy computations, write down the computation process and show your tree step by step. (No partial credit)

   (b) (10 points) Construct the tree *manually* using the Gini index, write down the computation process and show your tree step by step. (No partial credit)

   SOLUTIONS:
   For (a) and (b), the answer is in the file hw2p2_solution.pdf in Solution folder.

3. (30 points) [**Evaluate Classifier**][**Song Ju**] Sepsis is the leading cause of mortality in the United States. Septic shock, the most advanced complication of sepsis due to severe abnormalities of circulation and/or cellular metabolism, reaches a mortality rate as high as 50%. It is estimated that as many as 80% of sepsis deaths could be prevented with early diagnosis and intervention. To predict whether or not a patient has septic shock (Yes/No), consider using the decision tree shown in Figure 5 which involves Systolic Blood Pressure (SBP), Mean Arterial Pressure (MAP), and vasopressor (Vaso). We will focus on the sub-tree which splits on the attribute "SBP" as shown in the red dashed region of Figure 5. Answer the following questions and show your work.

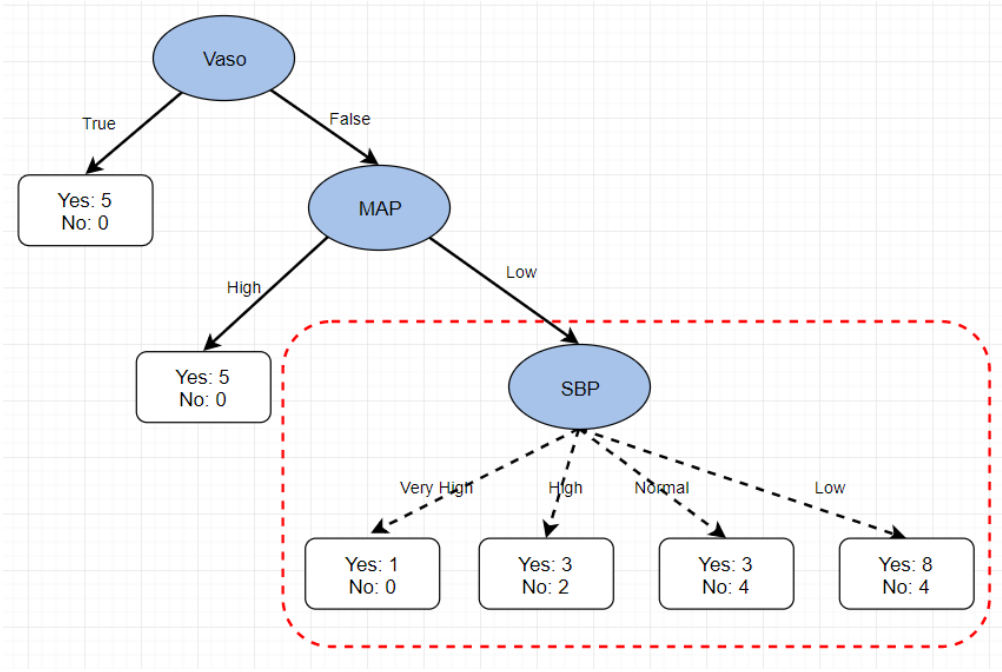


Figure 5: Decision Tree

(a) (13 points) Post-pruning based on **optimistic errors**.

i. (4 points) Calculate the optimistic errors before splitting and after splitting using SBP respectively.

ii. (3 points) Based upon the optimistic errors, would the subtree be pruned or retained? If it is pruned, draw the resulting decision tree and use it for the next question; otherwise, use the original decision tree shown in Figure 5 for the next question.

iii. (6 points) Use the decision tree from (a)-(ii) above to classify the provided testing dataset ("hw2q3_test.csv"). Report the Accuracy, Recall, Precision, Specificity, Sensitivity, and F1 Measure.

(b) (13 points) Post-pruning based on **pessimistic errors**. When calculating pessimistic errors, each leaf node will add a factor of 0.5 to the error.

i. (4 points) Calculate the pessimistic errors before splitting and after splitting using SBP respectively.

ii. (3 points) Based on the pessimistic errors, would the subtree be pruned or retained? If it is pruned, draw the resulting decision tree and use it for the next question; otherwise, use the original decision tree shown in Figure 5 for the next question.

iii. (6 points) Use the decision tree from (b)-(ii) above to classify the provided testing dataset ("hw2q3_test.csv"). Report the Accuracy, Recall, Precision, Specificity, Sensitivity, and F1 Measure.

(c) (4 points) We will compare the performance of the decision trees from (a)-(ii) and (b)-(ii) on the testing dataset ("hw2q3_test.csv"). If we only consider Accuracy,

Recall, and Precision, which decision tree would be a better model for the task of septic shock prediction. Justify your answer.

SOLUTIONS:

(a) i. Before splitting:
optimistic error = 10/25
After splitting:
optimistic error = 9/25

ii. The optimistic error decreases after splitting, so we should not prune this branch. The resulting tree is shown below:

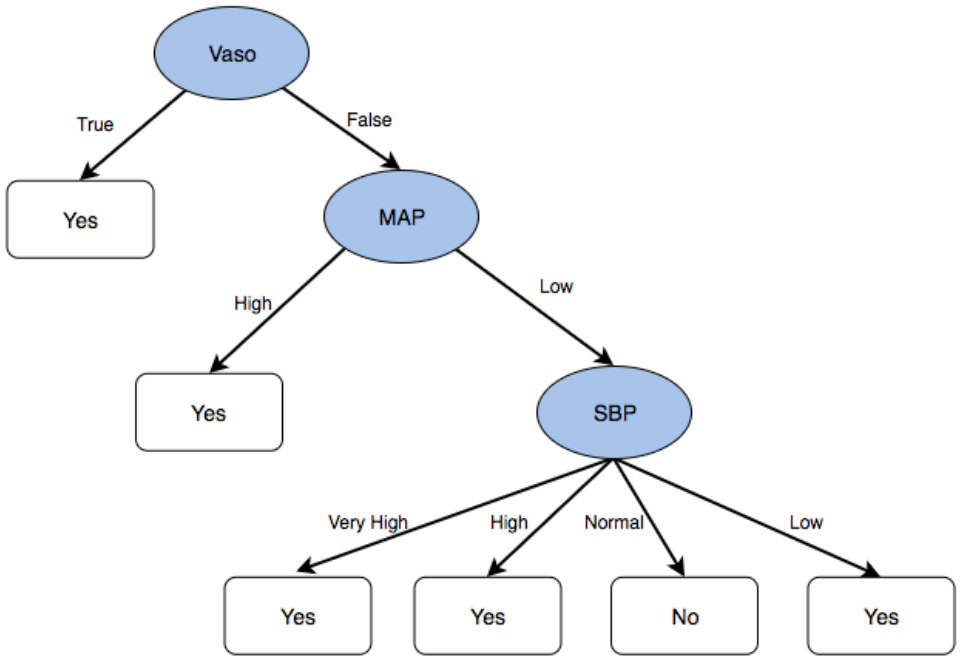


Figure 6: DT pruned based on optimistic

iii. For the tree pruned based on optimistic error:
Accuracy: = 17/20 = 0.85
Recall: = 13/15 = 0.867
Precision: = 13/14 = 0.929
Specificity: = 4/5 = 0.8
Sensitivity: = 13/15 = 0.867
F1 Measure: 26/29 = 0.897

(b) i. Before splitting:
pessimistic error = $\frac{10+(0.5*1)}{25}$ = 10.5/25
After splitting:

pessimistic error $= \frac{9+(0.5*4)}{25} = 11/25$

ii. The pessimistic error increases after splitting, so we should prune this branch. The resulting tree is shown below:

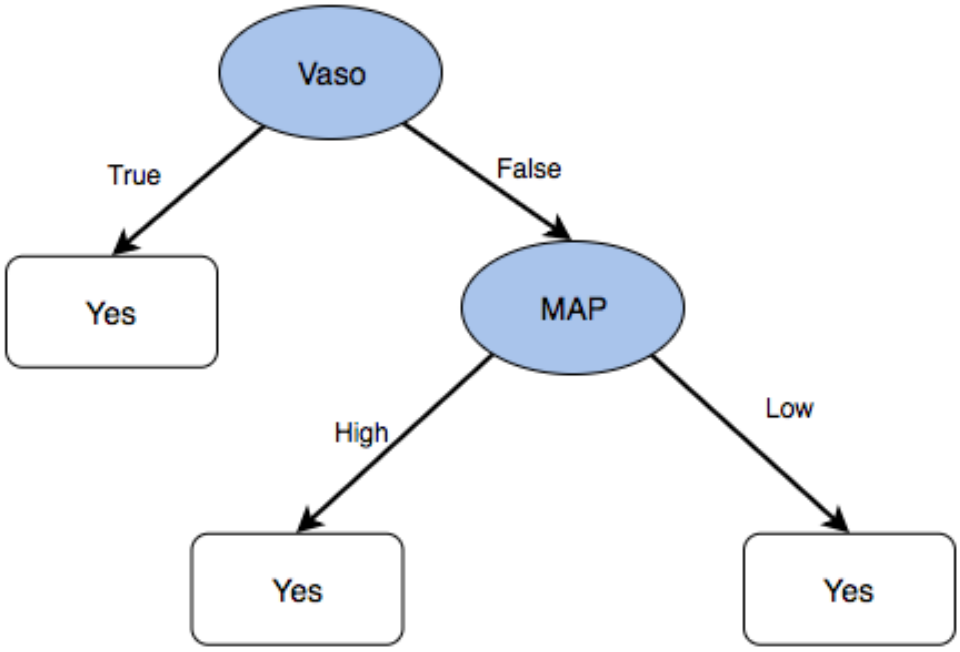


Figure 7: DT pruned based on pessimistic

iii. For the tree pruned based on pessimistic error:
Accuracy: $= 15/20 = 0.75$
Recall: $= 15/15 = 1$
Precision: $= 15/20 = 0.75$
Specificity: $0/5 = 0$
Sensitivity: $15/15 = 1$
F1 Measure: $= 30/35 = 0.857$

(c) Recall is more important. Because when we were predicting whether a patient has sepsis shock, a false negative could be considered much worse than a false positive.

4. (15 points) [**Adaboost**][**Xi Yang**] Consider the labeled data points in Figure 8, where '✕' and '●' indicate class labels. We will use AdaBoost with decision stumps to train a classifier for the '✕' and '●' labels. Each boosting iteration will select the stump that minimizes the weighted training error. Breaking ties by choosing '●'. All of the data points start with uniform weights.

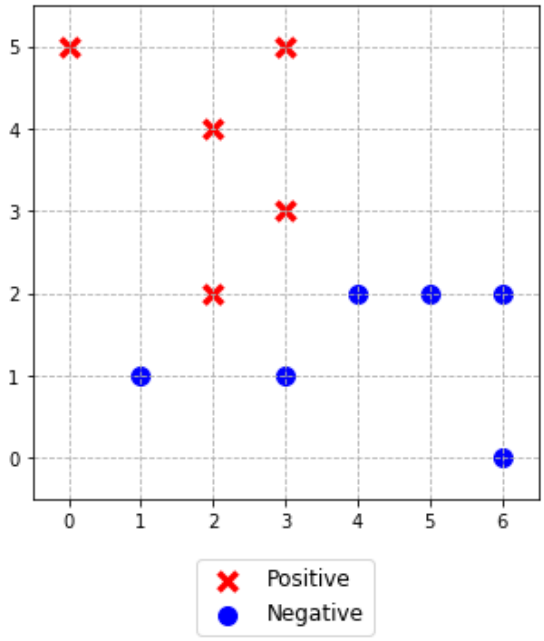


Figure 8: Adaboost

(a) (4 points) In Figure 8, draw a decision boundary corresponding to the first decision stump that the algorithm would choose (the decision boundary should be either a vertical or horizontal straight line). Label the decision boundary as (1), also indicate the ✕ / ● sides of this boundary.

(b) (2 points) Circle the point(s) that have the highest weight after the first boosting iteration.

(c) (5 points) After the labels have been reweighted in the first boosting iteration, what is the weighted error of the decision boundary (1)?

(d) (4 points) Draw the decision boundary corresponding to the second decision stump that the algorithm would choose (the decision boundary should be either a vertical or horizontal straight line). Label the decision boundary as (2), also indicate the ✕ / ● sides of this boundary.

(Please display your answers for (a), (b) and (d) in a single figure.)

SOLUTIONS:

(a) (4 points) Refer to boundary (1) and '✕' and '●' signs indicated in Figure 9.

(b) (2 points) Refer to circles indicated in Figure 9.

(c) (5 points) Initialize the data weighting coefficients as:
$\omega_n^{(1)} = \frac{1}{11}, n = 1, ...11$

The weighted error before boosting iterations:

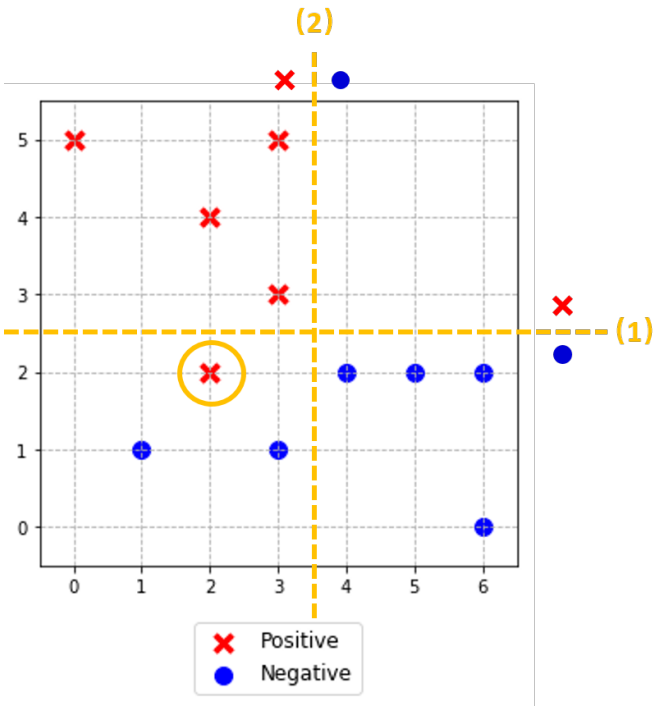


Figure 9: Adaboost Answer

$J_1 = \sum_{n=1}^{11} \omega_n^{(1)} I(y_1(x_n) \neq t_n) = \frac{1}{11} * 1 = \frac{1}{11}$

Evaluate the quantities:
$\varepsilon_1 = \frac{J_1}{\sum_{n=1}^{11} \omega_n^{(1)}} = \frac{1}{11}$
$\alpha_1 = \frac{1}{2} \ln \frac{1-\varepsilon_1}{\varepsilon_1} = 1.1513$

For correctly classified data: $\omega_n^{(1)} \exp^{-\alpha_1} = \frac{1}{11} * \exp^{-1.1513} = 0.0287$
For incorrectly classified data: $\omega_n^{(1)} \exp^{\alpha_1} = \frac{1}{11} * \exp^{1.1513} = 0.2875$

Normalization factor: $Z_1 = 10 * 0.0287 + 1 * 0.2875 = 0.5745$
Updated data weighting coefficients:
For correctly classified data: $\omega_n^{(2)} = 0.0287/0.5745 = 0.05$

For incorrectly classified data: $\omega_n^{(2)} = 0.2875/0.5745 = 0.5$

After the first iteration, the weighted error of the first decision boundary is:
$0.5 * 1 = 0.5$

(d) (4 points) Refer to boundary (2) and '✖' and '●' signs indicated in Figure 9.

5. (20 points) [**Naïve Bayes + Decision Tree**] [**Ruth Okoilu**] For this exercise, use the provided 'hw2q5.csv' which contains 24 data points. It has six attributes: each data point will be referred to using the first column "Id" and we will use columns 2-5 to

predict the final column "Class" (whether or not a patient should have contact lens).

(a) (15 points) Compare the performance of two classifiers: Naïve Bayes (NB) vs. Decision Tree (DT) using 5-fold cross-validation (CV) and **report their 5-fold CV accuracy**. For the ith fold, the testing dataset is composed of all the data points whose (Id mod $5 = i - 1$). Follow the lecture's code to build your decision trees except that multiple-way splitting is allowed and use Information Gain (IG) to select the best attribute. In the case of ties, break ties in favor of the leftmost feature. For each fold, show the induced Naïve Bayes and DT models.

(b) (5 points) Based on the **5-fold CV accuracy** from (a), which classifier, NB or DT, would you choose? Report your final model for the selected classifier.

Show your work. No Partial Credit.

SOLUTIONS:

(a) (15 points) In **Fold_1**:

1) For Naïve Bayes:

Class	Probability
Yes	0.35
No	0.65

patient age	Class	Probability
young	Yes	0.5714
young	No	0.2308
pre-presbyopic	Yes	0.2857
pre-presbyopic	No	0.3077
presbyopic	Yes	0.1429
presbyopic	No	0.4615

spectacle prescription	Class	Probability
myope	Yes	0.4286
myope	No	0.5385
hypermetrope	Yes	0.5714
hypermetrope	No	0.4615

astigmatic	Class	Probability
yes	Yes	0.4286
yes	No	0.5385
no	Yes	0.5714
no	No	0.4615

tear production rate	Class	Probability
reduced	Yes	0.0
reduced	No	0.7692
normal	Yes	1.0
normal	No	0.2308

When doing classification for testing data:

Id	Class	Probability	Predict	Actual
5	Yes	(0.5714*0.5714*0.5714*0.0)*0.35 = 0	No	No
	No	(0.2308*0.4615*0.4615*0.7692)*0.65 = 0.0246		
10	Yes	(0.2857*0.4286*0.5714*1.0)*0.35 = 0.0245	Yes	Yes
	No	(0.3077*0.5385*0.4615*0.2308)*0.65 = 0.0115		
15	Yes	(0.2857*0.5714*0.4286*0.0)*0.35 = 0	No	No
	No	(0.3077*0.4615*0.5385*0.7692)*0.65 = 0.0382		
20	Yes	(0.1429*0.4286*0.4286*1.0)*0.35 = 0.0092	No	Yes
	No	(0.4615*0.5385*0.5385*0.2308)*0.65 = 0.0201		

2) For Decision Tree:

Layer 1:

The entropy of the Class is: $H(Class) = -\frac{13}{20}log_2(\frac{13}{20}) - \frac{7}{20}log_2(\frac{7}{20}) = 0.9341$

Class	Yes	No	Probability	Entropy
young	$\frac{4}{7}$	$\frac{3}{7}$	$\frac{7}{20}$	$-\frac{4}{7}log_2(\frac{4}{7}) - \frac{3}{7}log_2(\frac{3}{7}) = 0.9852$
pre-presbyopic	$\frac{2}{6}$	$\frac{4}{6}$	$\frac{6}{20}$	$-\frac{2}{6}log_2(\frac{2}{6}) - \frac{4}{6}log_2(\frac{4}{6}) = 0.9183$
presbyopic	$\frac{1}{7}$	$\frac{6}{7}$	$\frac{7}{20}$	$-\frac{1}{7}log_2(\frac{1}{7}) - \frac{6}{7}log_2(\frac{6}{7}) = 0.5917$
$H(Class\|PatientAge) = \frac{7}{20} * 0.9852 + \frac{6}{20} * 0.9183 + \frac{7}{20} * 0.5917 = 0.8274$				
$IG(Class) = H(Class) - (Class\|PatientAge) = 0.9341 - 0.8274 = 0.1067$				

Class	Yes	No	Probability	Entropy
myope	$\frac{3}{10}$	$\frac{7}{10}$	$\frac{10}{20}$	$-\frac{3}{10}log_2(\frac{3}{10}) - \frac{7}{10}log_2(\frac{7}{10}) = 0.8813$
hypermetrope	$\frac{4}{10}$	$\frac{6}{10}$	$\frac{10}{20}$	$-\frac{4}{10}log_2(\frac{4}{10}) - \frac{6}{10}log_2(\frac{6}{10}) = 0.9710$
$H(Class\|SpectaclePrescription) = \frac{10}{20} * 0.8813 + \frac{10}{20} * 0.9710 = 0.9261$				
$IG(Class) = H(Class) - (Class\|SpectaclePrescription) = 0.9341 - 0.9261 = 0.0080$				

Class	Yes	No	Probability	Entropy
yes	$\frac{3}{10}$	$\frac{7}{10}$	$\frac{10}{20}$	$-\frac{3}{10}log_2(\frac{3}{10}) - \frac{7}{10}log_2(\frac{7}{10}) = 0.8813$
no	$\frac{4}{10}$	$\frac{6}{10}$	$\frac{10}{20}$	$-\frac{4}{10}log_2(\frac{4}{10}) - \frac{6}{10}log_2(\frac{6}{10}) = 0.9710$
$H(Class\|Astigmatic) = \frac{10}{20} * 0.8813 + \frac{10}{20} * 0.9710 = 0.9261$				
$IG(Class) = H(Class) - (Class\|Astigmatic) = 0.9341 - 0.9261 = 0.0080$				

Class	Yes	No	Probability	Entropy
reduced	$\frac{0}{10}$	$\frac{10}{10}$	$\frac{10}{20}$	$-\frac{0}{10}log_2(\frac{0}{10}) - \frac{10}{10}log_2(\frac{10}{10}) = 0$
normal	$\frac{7}{10}$	$\frac{3}{10}$	$\frac{10}{20}$	$-\frac{7}{10}log_2(\frac{7}{10}) - \frac{3}{10}log_2(\frac{3}{10}) = 0.8813$
$H(Class\|TearProductionRate) = \frac{10}{20} * 0 + \frac{10}{20} * 0.8813 = 0.4406$				
$IG(Class) = H(Class) - (Class\|TearProductionRate) = 0.9341 - 0.4406 = 0.4934$				

The root node is: **tear production rate**.

Layer 2:

- [**Layer 2**] For the left node (tear production rate = reduced):
All labels belong to the No Class.

- [**Layer 2**] For the right node (tear production rate = normal):

Class	Yes	No	Probability	Entropy
young	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{4}{10}$	$-\frac{4}{4}log_2(\frac{4}{4}) - \frac{0}{4}log_2(\frac{0}{4}) = 0$
pre-presbyopic	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{3}{10}$	$-\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.9183$
presbyopic	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{3}{10}$	$-\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3}) = 0.9183$
$H(Class\|PatientAge) = \frac{4}{10} * 0 + \frac{3}{10} * 0.9183 + \frac{3}{10} * 0.9183 = 0.5510$				
$IG(Class) = H(Class) - (Class\|PatientAge) = 0.8813 - 0.5510 = 0.3303$				

Class	Yes	No	Probability	Entropy
myope	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{4}{10}$	$-\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4}) = 0.8113$
hypermetrope	$\frac{4}{6}$	$\frac{2}{6}$	$\frac{6}{10}$	$-\frac{4}{6}log_2(\frac{4}{6}) - \frac{2}{6}log_2(\frac{2}{6}) = 0.9183$
$H(Class\|SpectaclePrescription) = \frac{4}{10} * 0.8113 + \frac{6}{10} * 0.9183 = 0.8755$				
$IG(Class) = H(Class) - (Class\|SpectaclePrescription) = 0.8813 - 0.8755 = 0.0058$				

Class	Yes	No	Probability	Entropy
yes	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{5}{10}$	$-\frac{3}{5}log_2(\frac{3}{5}) - \frac{2}{5}log_2(\frac{2}{5}) = 0.9710$
no	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{5}{10}$	$-\frac{4}{5}log_2(\frac{4}{5}) - \frac{1}{5}log_2(\frac{1}{5}) = 0.7219$
$H(Class\|Astigmatic) = \frac{5}{10} * 0.9710 + \frac{5}{10} * 0.7219 = 0.8464$				
$IG(Class) = H(Class) - (Class\|Astigmatic) = 0.8813 - 0.8464 = 0.0349$				

The right node will be split by **patient age**.

Layer 3:

- [**Layer 3**] For the left node (patient age = young):
All labels belong to the Yes Class.
- [**Layer 3**] For the middle node (patient age = pre-presbyopic):

Class	Yes	No	Probability	Entropy
myope	$\frac{1}{1}$	$\frac{0}{1}$	$\frac{1}{3}$	$-\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$
hypermetrope	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$
$H(Class\|SpectaclePrescription) = \frac{1}{3} * 0 + \frac{2}{3} * 1 = 0.6667$				
$IG(Class) = H(Class) - (Class\|SpectaclePrescription) = 0.9183 - 0.6667 = 0.2516$				

Class	Yes	No	Probability	Entropy
yes	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$
no	$\frac{1}{1}$	$\frac{0}{1}$	$\frac{1}{3}$	$-\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$
$H(Class\|Astigmatic) = \frac{2}{3} * 1 + \frac{1}{3} * 0 = 0.6667$				
$IG(Class) = H(Class) - (Class\|Astigmatic) = 0.9183 - 0.6667 = 0.2516$				

Generate a tie, select the left-most feature **spectacle prescription**.

Layer 3 → Layer 4:

- [**Layer 4**] For the left node (spectacle prescription = myope):
All labels belong to the Yes Class.

- [**Layer 4**] For the right node (spectacle prescription = hypermetrope):

Class	Yes	No	Probability	Entropy
yes	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{1}{2}$	$-\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$
no	$\frac{1}{1}$	$\frac{0}{1}$	$\frac{1}{2}$	$-\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$
$H(Class\|Astigmatic) = \frac{1}{2} * 0 + \frac{1}{2} * 0 = 0$				
$IG(Class) = H(Class) - (Class\|Astigmatic) = 1 - 0 = 1$				

- [**Layer 3**] For the right node (patient age = presbyopic):

Class	Yes	No	Probability	Entropy
myope	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{1}{3}$	$-\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$
hypermetrope	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$
$H(Class\|SpectaclePrescription) = \frac{1}{3} * 0 + \frac{2}{3} * 1 = 0.6667$				
$IG(Class) = H(Class) - (Class\|SpectaclePrescription) = 0.9183 - 0.6667 = 0.2516$				

Class	Yes	No	Probability	Entropy
yes	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{1}{3}$	$-\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$
no	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$
$H(Class\|Astigmatic) = \frac{1}{3} * 0 + \frac{2}{3} * 1 = 0.6667$				
$IG(Class) = H(Class) - (Class\|Astigmatic) = 0.9183 - 0.6667 = 0.2516$				

Generate a tie, select the left-most feature **spectacle prescription**.

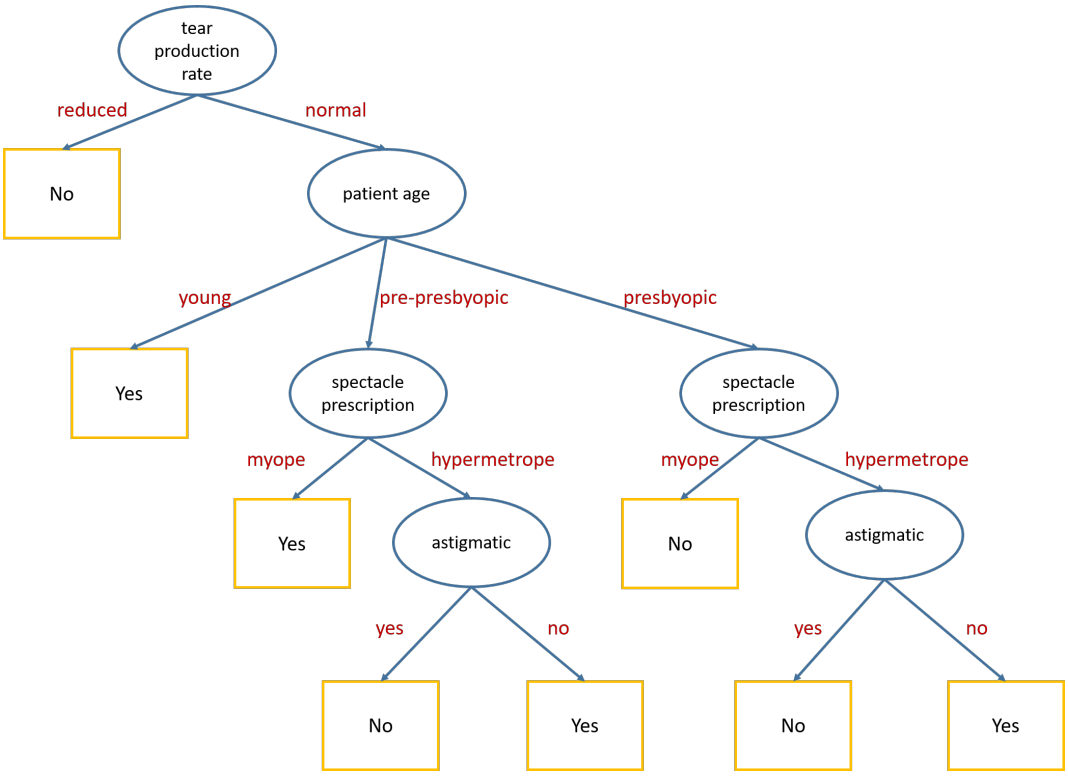
Layer 3 → Layer 4:

- [**Layer 4**] For the left node (spectacle prescription = myope):
All labels belong to the No Class.

- [**Layer 4**] For the right node (spectacle prescription = hypermetrope):

Class	Yes	No	Probability	Entropy
yes	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{1}{2}$	$-\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$
no	$\frac{1}{1}$	$\frac{0}{1}$	$\frac{1}{2}$	$-\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$
$H(Class\|Astigmatic) = \frac{1}{2} * 0 + \frac{1}{2} * 0 = 0$				
$IG(Class) = H(Class) - (Class\|Astigmatic) = 1 - 0 = 1$				

The decision tree is:



The predicted results for testing data are:

Id	Predict	Actual
5	No	No
10	Yes	Yes
15	No	No
20	No	Yes

In **Fold_2**:

For Naïve Bayes:

Class	Probability
Yes	0.4211
No	0.5789

patient age	Class	Probability
young	Yes	0.375
young	No	0.2727
pre-presbyopic	Yes	0.375
pre-presbyopic	No	0.2727
presbyopic	Yes	0.25
presbyopic	No	0.4545

spectacle prescription	Class	Probability
myope	Yes	0.625
myope	No	0.4545
hypermetrope	Yes	0.375
hypermetrope	No	0.5455

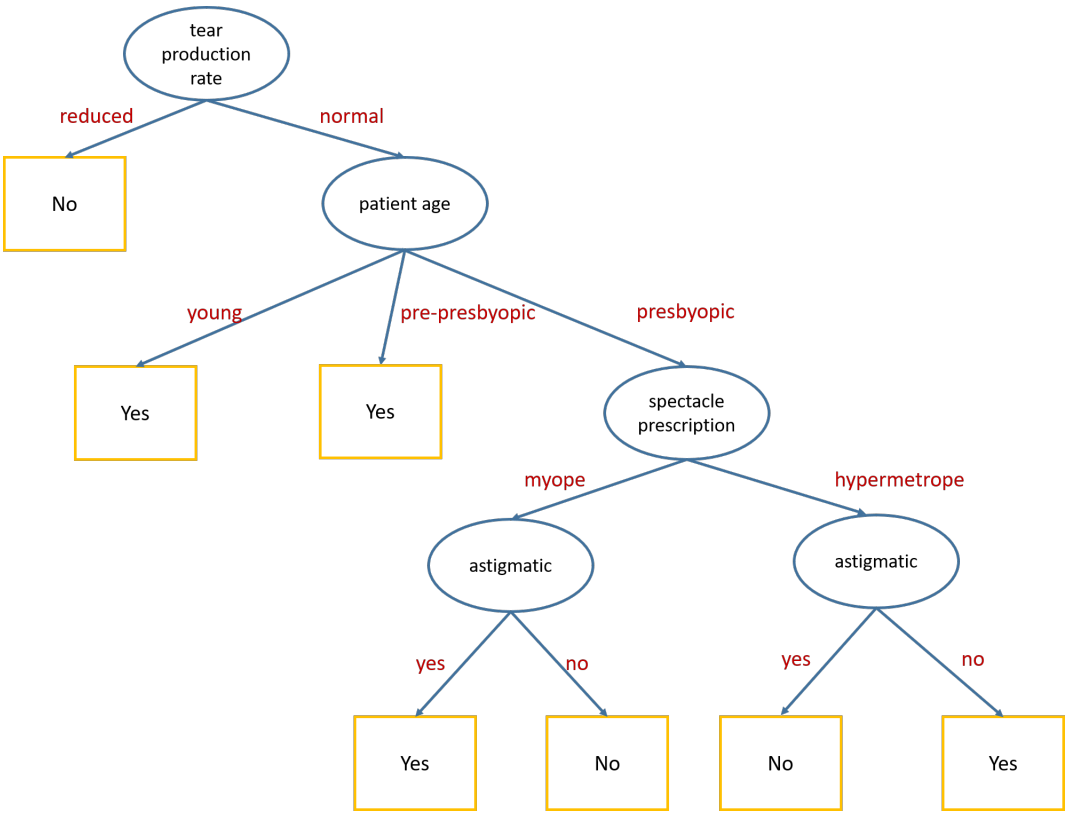
astigmatic	Class	Probability
yes	Yes	0.5
yes	No	0.5455
no	Yes	0.5
no	No	0.4545

tear production rate	Class	Probability
reduced	Yes	0.0
reduced	No	0.8182
normal	Yes	1.0
normal	No	0.1818

When doing classification for testing data:

Id	Class	Probability	Predict	Actual
1	Yes No	(0.375*0.625*0.5*0.0)*0.4211 = 0 (0.2727*0.4545*0.45454*0.8182)*0.5789 = 0.0267	No	No
6	Yes No	(0.375*0.375*0.5*1.0)*0.4211 = 0.0296 (0.2727*0.5455*0.4545*0.1818)*0.5789 = 0.0071	Yes	Yes
11	Yes No	(0.375*0.625*0.5*0.0)*0.4211 = 0 (0.2727*0.4545*0.5455*0.8182)*0.5789 = 0.0320	No	No
16	Yes No	(0.375*0.375*0.5*1.0)*0.4211 = 0.0296 (0.2727*0.5455*0.5455*0.1818)*0.5789 = 0.0085	Yes	No
21	Yes No	(0.25*0.375*0.5*0.0)*0.4211 = 0 (0.4545*0.5455*0.4545*0.8182)*0.5789 = 0.0534	No	No

2) For Decision Tree:
The calculations are omitted. The decision tree is:



The predicted results for testing data are:

Id	Predict	Actual
1	No	No
6	Yes	Yes
11	No	No
16	Yes	No
21	No	No

In **Fold_3**:
For Naïve Bayes:

Class	Probability
Yes	0.3158
No	0.6842

patient age	Class	Probability
young	Yes	0.5
young	No	0.2308
pre-presbyopic	Yes	0.3333
pre-presbyopic	No	0.3846
presbyopic	Yes	0.1667
presbyopic	No	0.3846

spectacle prescription	Class	Probability
myope	Yes	0.5
myope	No	0.4615
hypermetrope	Yes	0.5
hypermetrope	No	0.5385

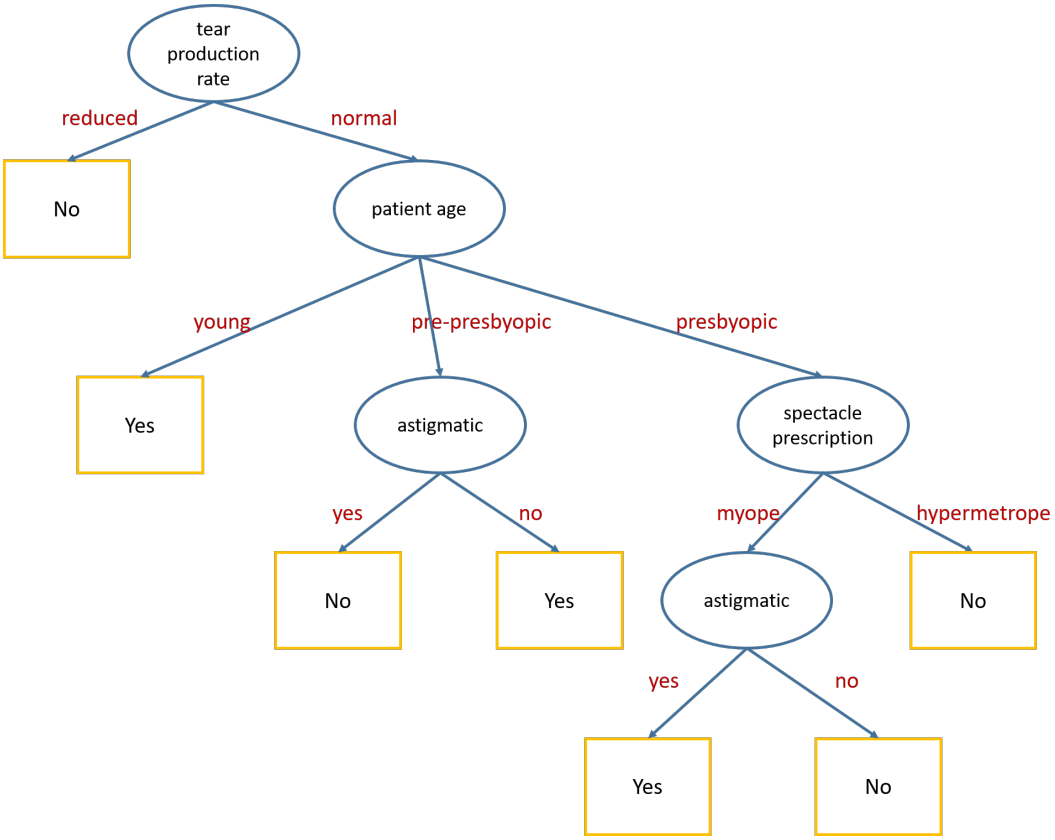
astigmatic	Class	Probability
yes	Yes	0.5
yes	No	0.5385
no	Yes	0.5
no	No	0.4615

tear production rate	Class	Probability
reduced	Yes	0.0
reduced	No	0.7692
normal	Yes	1.0
normal	No	0.2308

When doing classification for testing data:

Id	Class	Probability	Predict	Actual
2	Yes	(0.5*0.5*0.5*1.0)*0.3158 = 0.0395	Yes	Yes
	No	(0.2308*0.4615*0.4615*0.2308)*0.6842 = 0.0078		
7	Yes	(0.5*0.5*0.5*0.0)*0.3158 = 0	No	No
	No	(0.2308*0.5385*0.5385*0.7692)*0.6842 = 0.0352		
12	Yes	(0.3333*0.5*0.5*1.0)*0.3158 = 0.0263	Yes	Yes
	No	(0.3846*0.4615*0.5385*0.2308)*0.6842 = 0.0151		
17	Yes	(0.1667*0.5*0.5*0.0)*0.3158 = 0	No	No
	No	(0.3846*0.4615*0.4615*0.7692)*0.6842 = 0.0431		
22	Yes	(0.1667*0.5*0.5*1.0)*0.3158 = 0.0132	No	Yes
	No	(0.3846*0.5385*0.4615*0.2308)*0.6842 = 0.0151		

2) For Decision Tree:
The calculations are omitted. The decision tree is:



The predicted results for testing data are:

Id	Predict	Actual
2	Yes	Yes
7	No	No
12	No	Yes
17	No	No
22	No	Yes

In **Fold_4**:

Class	Probability
Yes	0.4211
No	0.5789

patient age	Class	Probability
young	Yes	0.375
young	No	0.2727
pre-presbyopic	Yes	0.375
pre-presbyopic	No	0.3636
presbyopic	Yes	0.25
presbyopic	No	0.3636

spectacle prescription	Class	Probability
myope	Yes	0.625
myope	No	0.4545
hypermetrope	Yes	0.375
hypermetrope	No	0.5455

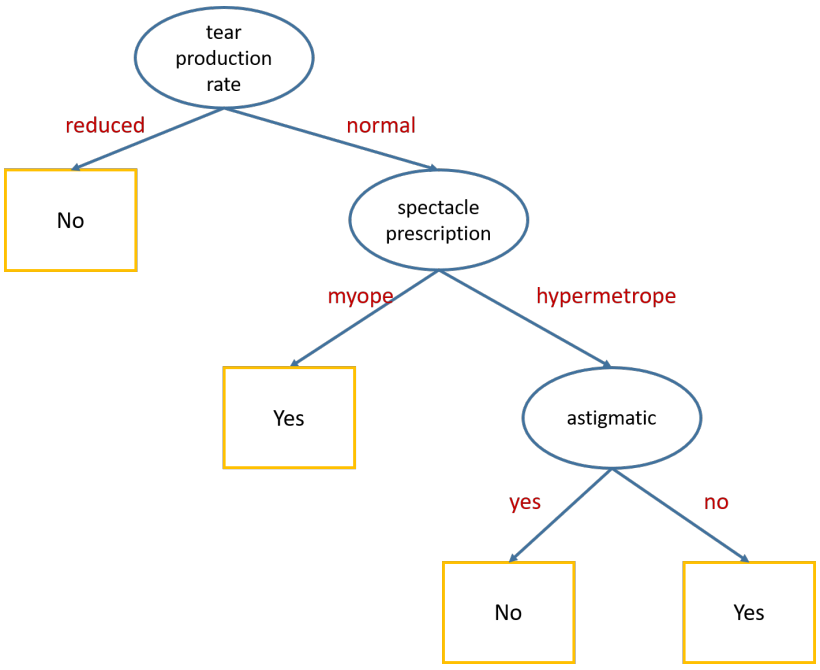
astigmatic	Class	Probability
yes	Yes	0.375
yes	No	0.5455
no	Yes	0.625
no	No	0.4545

tear production rate	Class	Probability
reduced	Yes	0.0
reduced	No	0.8182
normal	Yes	1.0
normal	No	0.1818

When doing classification for testing data:

Id	Class	Probability	Predict	Actual
3	Yes No	(0.375*0.625*0.375*0.0)*0.4211 = 0 (0.2727*0.4545*0.5455*0.8182)*0.5789 = 0.0320	No	No
8	Yes No	(0.375*0.375*0.375*1.0)*0.4211 = 0.0222 (0.2727*0.5455*0.5455*0.1818)*0.5789 = 0.0085	Yes	Yes
13	Yes No	(0.375*0.375*0.625*0.0)*0.4211 = 0 (0.3636*0.5455*0.4545*0.8182)*0.5789 = 0.0427	No	No
18	Yes No	(0.25*0.625*0.625*1.0)*0.4211 = 0.0411 (0.3636*0.4545*0.4545*0.1818)*0.5789 = 0.0079	Yes	No
23	Yes No	(0.25*0.375*0.375*0.0)*0.4211 = 0 (0.3636*0.5455*0.5455*0.8182)*0.5789 = 0.0512	No	No

2) For Decision Tree:
The calculations are omitted. The decision tree is:



The predicted results for testing data are:

Id	Predict	Actual
3	No	No
8	No	Yes
13	No	No
18	Yes	No
23	No	No

In **Fold_5**:

Class	Probability
Yes	0.3684
No	0.6316

For Naïve Bayes:

patient age	Class	Probability
young	Yes	0.4286
young	No	0.3333
pre-presbyopic	Yes	0.2857
pre-presbyopic	No	0.3333
presbyopic	Yes	0.2857
presbyopic	No	0.3333

spectacle prescription	Class	Probability
myope	Yes	0.5714
myope	No	0.4167
hypermetrope	Yes	0.4286
hypermetrope	No	0.5833

astigmatic	Class	Probability
yes	Yes	0.4286
yes	No	0.5
no	Yes	0.5714
no	No	0.5

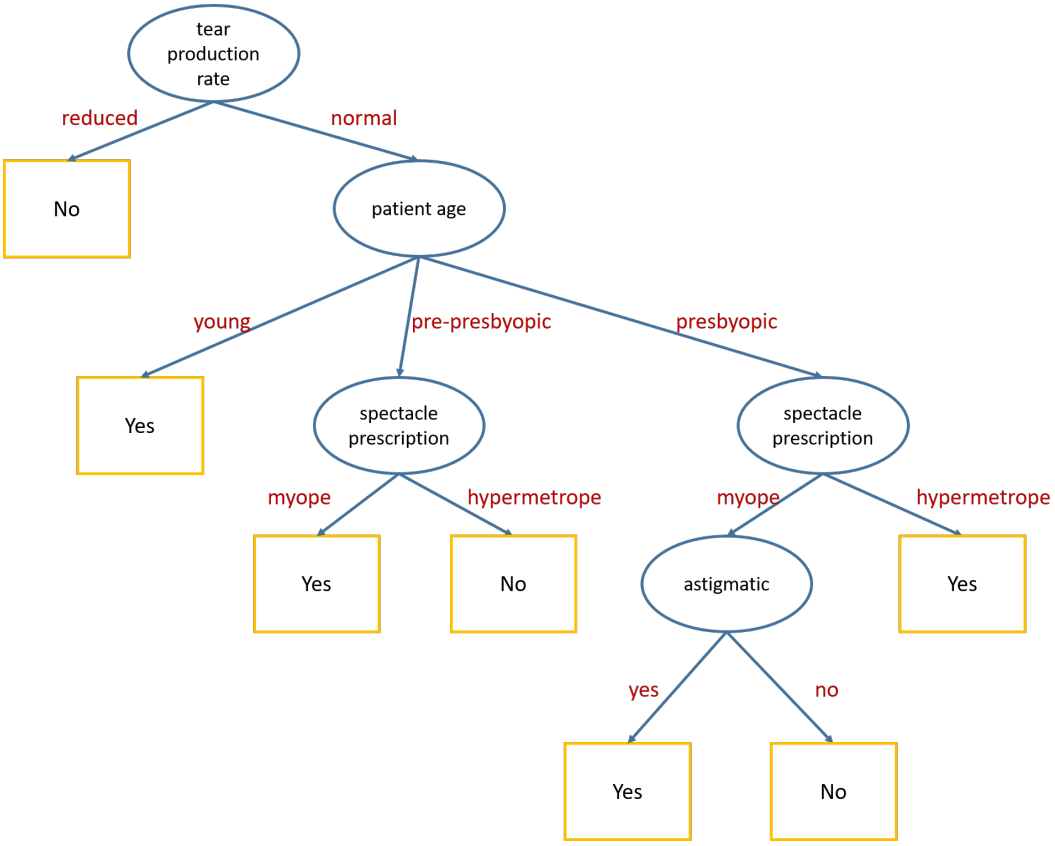
tear production rate	Class	Probability
reduced	Yes	0.0
reduced	No	0.8333
normal	Yes	1.0
normal	No	0.1667

When doing classification for testing data:

Id	Class	Probability	Predict	Actual
4	Yes	(0.4286*0.5714*0.4286*1.0)*0.3684 = 0.0387	Yes	Yes
	No	(0.3333*0.4167*0.5*0.1667)*0.6316 = 0.0073		
9	Yes	(0.2857*0.5714*0.5714*0.0)*0.3684 = 0	No	No
	No	(0.3333*0.4167*0.5*0.8333)*0.6316 = 0.0365		
14	Yes	(0.2857*0.4286*0.5714*1.0)*0.3684 = 0.0258	Yes	Yes
	No	(0.3333*0.5833*0.5*0.1667)*0.6316 = 0.0102		
19	Yes	(0.2857*0.5714*0.4286*0.0)*0.3684 = 0	No	No
	No	(0.3333*0.4167*0.5*0.8333)*0.6316 = 0.0365		
24	Yes	(0.2857*0.4286*0.4286*1.0)*0.3684 = 0.0193	Yes	No
	No	(0.3333*0.5833*0.5*0.1667)*0.6316 = 0.0102		

2) For Decision Tree:

The calculations are omitted. The decision tree is:



The predicted results for testing data are:

Id	Predict	Actual
4	Yes	Yes
9	No	No
14	No	Yes
19	No	No
24	Yes	No

Considering all of 5-folds: The accuracy of Naïve Bayes is: $19/24 = 0.792$
The accuracy of Decision Tree is: $16/24 = 0.667$

(b) (5 points) Based on the **5-fold CV accuracy** from (a), the Decision Tree is better. With all training data, the final model is:

Class	Probability
Yes	0.375
No	0.625

For Naïve Bayes:

patient age	Class	Probability
young	Yes	0.4444
young	No	0.2667
pre-presbyopic	Yes	0.3333
pre-presbyopic	No	0.3333
presbyopic	Yes	0.2222
presbyopic	No	0.4

spectacle prescription	Class	Probability
myope	Yes	0.5556
myope	No	0.4667
hypermetrope	Yes	0.4444
hypermetrope	No	0.5333

astigmatic	Class	Probability
yes	Yes	0.4444
yes	No	0.5333
no	Yes	0.5556
no	No	0.4667

tear production rate	Class	Probability
reduced	Yes	0.0
reduced	No	0.8
normal	Yes	1.0
normal	No	0.2

(c) (15 points) [**KNN + CV**] [**Ruth Okoilu**] Consider the following dataset with two real-valued inputs x_1 and x_2 and a binary output class y shown in Table 1. Each data point will be referred to using the first column "ID" in the following. Use KNN with unweighted Euclidean distance to predict the class y.

(a) (2 points) What are the 3 nearest neighbors for data points 5 and 10 respectively. (No partial credit).

ID	x_1	x_2	y(Class)
1	27	6	-
2	-6	2	-
3	2	2	+
4	36	2	+
5	-8	4	-
6	40	2	+
7	35	4	-
8	30	2	+
9	20	6	+
10	-1	4	-

Table 1: KNN + CV

(b) (5 points) What is the leave-one-out cross-validation error of 1-NN on this dataset? (No partial credit).

(c) (5 points) What is the 5-fold cross-validation error of 3-NN on this dataset? For the ith fold where $i = 1, 2, 3, 4, 5$, the testing dataset is composed of all the data points whose (ID mod $5 = i - 1$). (No partial credit).

(d) (3 points) Based on the results of (b) and (c), can we determine which is a better classifier, 1-NN or 3-NN? Why? (Answers without a correct justification will get zero points.)

SOLUTIONS:

(a) 3 nearest neighbors of the data point 5: {2,10,3}
3 nearest neighbors of the data point 10: {3,2,5}

(b) For the LOOCV in this case, we temporarily remove x_i from the dataset, and train on the remaining 9 data points. The label is categorical, so thus the error is 1 when the true label and the prediction match, otherwise 0. Table 2 shows the error after each run. Thus, Error = 7/10 = 0.7

(c) Table 3, shows the predicted labels and error in each run. Thus, Error =7/10 = 0.7

(d) No, LOOCV 1NN and 5 fold CV 3NN are not comparable.

[c].4

ID	x1	x2	y(Class)	1-NN	pred	error
1	27	6	-	8	+	1
2	-6	2	-	5	-	0
3	2	2	+	10	-	1
4	36	2	+	7	-	1
5	-8	4	-	2	-	0
6	40	2	+	4	+	0
7	35	4	-	4	+	1
8	30	2	+	1	-	1
9	20	6	+	1	-	1
10	-1	4	-	3	+	1

Table 2: Leave-one-out of cross val.

[c].4

fold	ID	x1	x2	y(Class)	3 NN	pred	error
1	5	-8	4	-	2,3,9	+	1
	10	-1	4	-	2,3,9	+	1
2	1	27	6	-	8,9,7	+	1
	6	40	2	+	4,7,8	+	0
3	2	-6	2	-	5,10,3	-	0
	7	35	4	-	4,6,8	+	1
4	3	2	2	+	10,2,5	-	1
	8	30	2	+	1,4,7	-	1
5	4	36	2	+	7,6,8	+	0
	9	20	6	+	1,8,7	-	1

Table 3: 5-fold cross validation