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.)

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)

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 1 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 1. Answer the following questions and show your work.

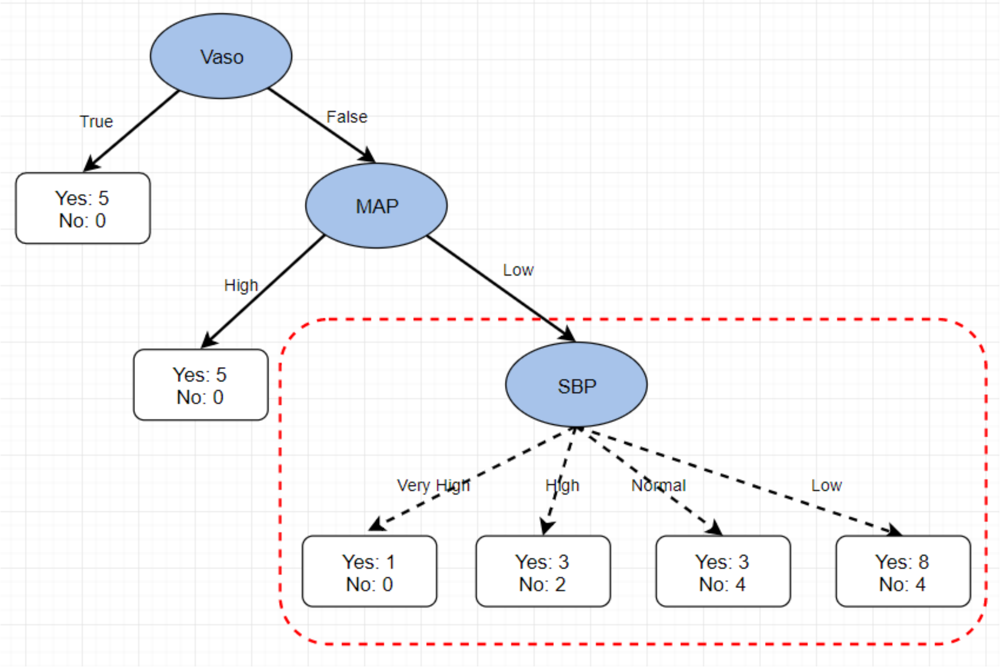

Figure 1: Decision Tree

(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 1 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 1 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.

4. (15 points) [**Adaboost**][**Xi Yang**] Consider the labeled data points in Figure 2, 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.

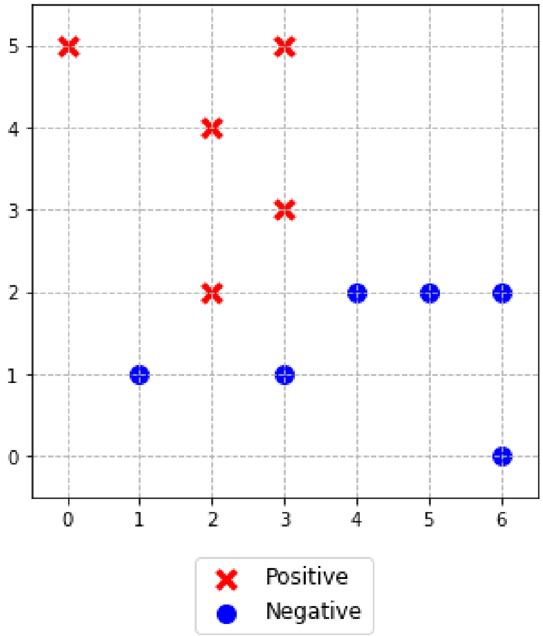

Figure 2: Adaboost

(a) (4 points) In Figure 2, 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.)

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.

6. (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.

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

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

   (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.)