Name _____

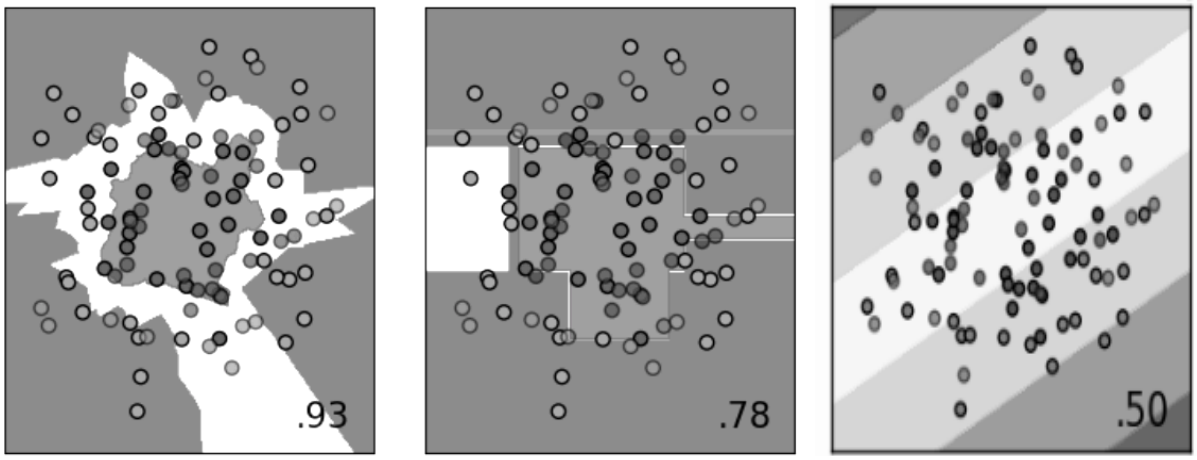
COMPSCI 348 **Practice Exam 2 Answer Sheet** (v.1) Spring 2019

Instructions (for the actual exam)

- *Do not open the exam until directed to by the instructor or TA*
- *Do not use books, notes, electronic devices, or other aids.*
- *Please avoid wrinkling the exam because that makes it difficult to scan.*
- *Your answers must be your own, so keep your eyes on your exam. Do not look at other students' exams.*
- *Answer each question. Note the point values and allocate your time accordingly.*
- *Be clear in marking your answers, and please place your answers in the designated spaces.*
- *Only the final answer in the designated space will be graded. However, other markings and calculations will be reviewed in support of regrade requests.*

Name _____

1. **Decision boundaries** — In the graphs below, the x and y axes correspond to predictor variables and the colors of the points correspond to classes. Given the example decision boundaries below, label each with the classifier family most likely to have created that boundary: Neural network (NN), classification tree (CT), or Linear discriminant analysis (LDA).



NN _____ DT _____ LDA _____

Name _____

2. **Neural network components** — Consider a specific artificial neuron that sits inside a neural network. It takes three inputs x_1=3, x_2=5, and x_3=10. The bias b is equal to 4.32. The weights are w_1=1.0, w_2=0.4, and w_3=0.5. What is the input to the non-linear transformation σ?

14.32

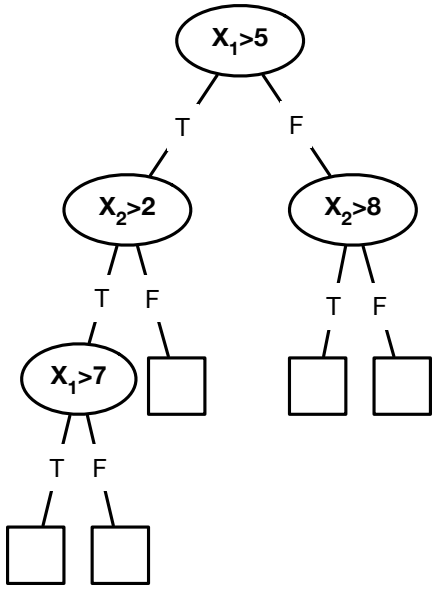
$$input = b + \sum_{x=1}^{n} x_i w_i$$

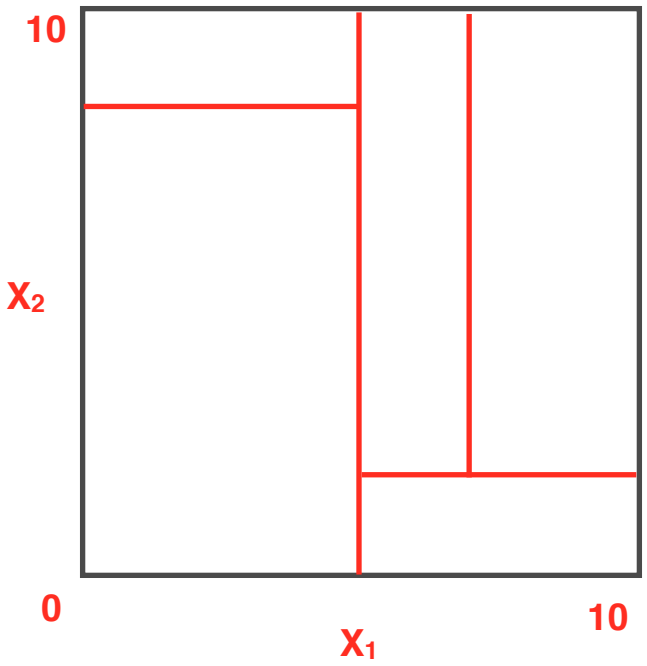
3. **Neural network structure** — Both *recurrent layers* and *long short-term memory* (LSTM) are network structures that are intended to deal with a particular type of statistical dependence present in data. Briefly describe that type of dependence and given an example of data with this type of dependence.

Both structures are intended to represent cases in which data instances are not independent. Examples of such structures include 2-D images (neighboring pixels are not independent) and sequence data such as text, genetic sequences, and speech (neighboring words, letters, base pairs, and phonemes are not independent).

Name _____

4. **Classification trees** — Given the classification tree shown below, draw the decision boundary implied by the tree in the 2-D space provided. Clearly label the axes of the space and provide a numeric scale.



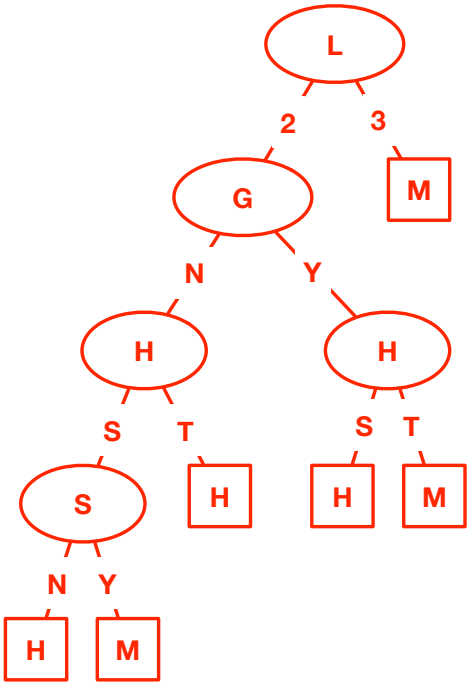


Answers are expected to: (1) Label the axes of the 2-D space as X_1 and X_2; (2) Provide some sort of scale for both axes; and (3) show the partitions of that space corresponding to the classification tree at left.

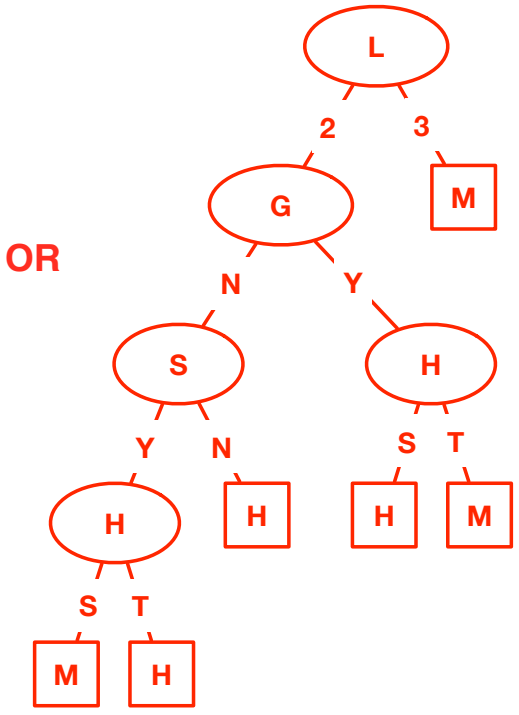
Name _____

5. Classification trees — Given the training set below, construct a classification tree to predict *Species* assuming that we use *misclassification rate* as the local loss function. Assume an algorithm that constructs a classification tree with only binary splits using greedy recursive partitioning. Also assume that the algorithm continues partitioning until all data instances at a leaf node have a single class. Draw the tree, including decision nodes, leaf nodes, and predicted class label for each leaf node.

Species (Y)	Green (G)	Legs (L)	Height (H)	Scales(S)
M	N	3	S	Y
M	Y	2	T	N
M	Y	3	T	N
M	N	2	S	Y
M	Y	3	T	N
H	N	2	T	Y
H	N	2	S	N
H	N	2	T	N
H	Y	2	S	N
H	N	2	T	Y



OR



Name _____

6. **Feature selection** — Briefly describe the three methods for features selection that we discussed in class. Describe the differences in the two methods that are actually practical and why the third isn't practical.

   Three methods are best subset selection, forward subset (or stepwise) selection, and backward subset (or stepwise) selection. The first is not computationally tractable for even small numbers of variables. Forward and backward subset selection differ in their starting subsets (the empty set and the total set, respectively) and the action taken at each iteration (adding and subtracting elements, respectively).

Name _____

7. Multiple comparison procedures — For each procedure described below, indicate whether the final statement about the outcome of the procedure is true or false. Answer the questions assuming that any one student's exam scores are unbiased estimators of the student's knowledge about the topics covered in the exam.

True	False	Procedure
○	●	Prof. Smith gives an exam to her students, and she presents an award to the student who gets the highest score on the exam. The score of that student is an unbiased estimator of how that same student would score on another exam about the same topics (but with different questions).
●	○	After an exam, Prof. Kim examines the numeric scores on the exam and then randomly selects one student. That student's exam score is an unbiased estimator of that student's knowledge about the topics covered in the exam.
○	●	Prof. Kao's class has two discussion sections that meet at the same time. Students were assigned randomly to the sections, but one section has twice as many students as the other section. After giving every exam, he selects the best exam score from each section and compares them. The section with the best of these two scores is an unbiased estimator of which section has the student who is most knowledgeable about the topics covered in the exam.

Name _____

8. **Cross-validation** — Below are four statements about cross-validation. For each statement, indicate whether it is true (T) or false (F).

a. ____T____ Given a finite sample of data, cross-validated estimates of misclassification error will generally have lower variance than estimates based on a single hold-out test set.

b. ____F____ In a data sample used for cross-validation, every data point is only used once for training and is used multiple times for testing.

c. ____F____ Randomly dividing a data set into a training set with 50% of the data and a test set with 50% of the data is equivalent to two-fold cross-validation.

d. ____T____ Cross-validation can be used to set hyper-parameters, but it can also be used to provide unbiased estimates of misclassification rate.

Name _____

9. **Bias and variance** — For each situation described below, indicate whether it most closely corresponds to bias (B) or variance (V) in the estimator of the italicized estimand.

a. ___V___ When individual, teams or countries fall short of the expected *number of medals* at the Winter Olympics, some sports commentators say: “It just wasn’t their day.” Conversely, when they do better than expected, sports commentators say “this was their day."

b. ___B___ In 2015, Volkswagen was found to have installed software that systematically lowered their estimated *emissions levels*.

c. ___B___ Some members of the US intelligence community are concerned that agents of the Russian government “meddled” in the 2016 US presidential election and affected the *vote count*.

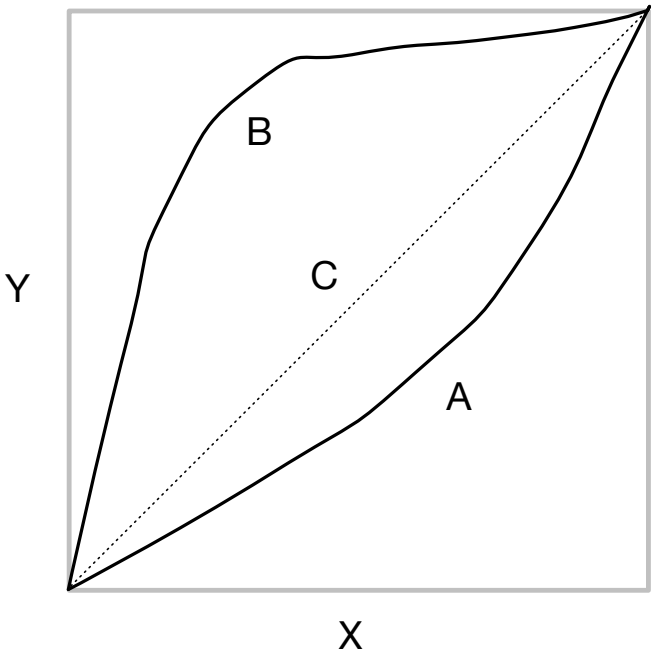
d. ___V___ Some scientists worry that their estimates of *the effect of certain drugs* are not reliable because they are based on very small sets of patients.

e. ___B___ These same scientists worry that their estimates of the *effect of certain drugs* are not realistic because they have only been tested on healthy 18-22 year olds.

Name _____

10. ROC curves — For the ROC space and ROC curves shown below, use one of the labels from the diagram {A, B, C, X, Y} to fill in the blank space corresponding to the described axis or curve. If no axis or curve in the figure corresponds to the description, leave it blank.

Label	Description of Axis or Curve
X	The false positive rate
_____	The false negative rate
Y	The true positive rate
_____	The true negative rate
C	The performance of a random classification model
B	The performance of a typical classification model
A	The performance of a classification model with very high error
_____	The performance of a classification model with zero error



Name _____

11. Ensembles — Below are five statements about ensembles. For each statement, indicate whether it is true (T) or false (F).

a. _____T_____ *Ensembles* are classifiers that combine the predictions of multiple models in an effort to improve overall accuracy.

b. _____F_____ *Bagging* is an ensemble technique that aggregates the predictions of multiple classifiers each of which has been learned on a different cross-validated sample.

c. _____T_____ The technique of *random forests* incorporates bagging, and also introduces additional independence among trees by constructing trees on randomly selected subsets of variables.

d. _____T_____ *Boosting* uses the model output by one application of a learning algorithm to weight the data instances that serve as the inputs to the next application of the learning algorithm.

e. _____T_____ An ensemble nearly always equals or reduces error when compared to the error of a single component model.