Name ________________________

COMPSCI 348 **Exam 2 Answer Key** 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.*
- *Answer questions on both sides of each page.*
- *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. **Neural network** (10 points) — Mark each statement below as *True* or *False* by filling in the corresponding circle.

True	False	Statement
○	●	The output of an artificial neuron in a neural network is a linear combination of inputs and a bias term.
○	●	The bias term is a hyper-parameter of a neural network.
●	○	Neural networks with recurrent layers are one way of modeling sequential data.
●	○	Multilayer feedforward neural networks are universal function approximators.
●	○	Backpropagation is a common method for learning the weights of a neural network.

Name ________________________

2. Classification trees (10 points) — Mark each statement below as *True* or *False* by filling in the corresponding circle.

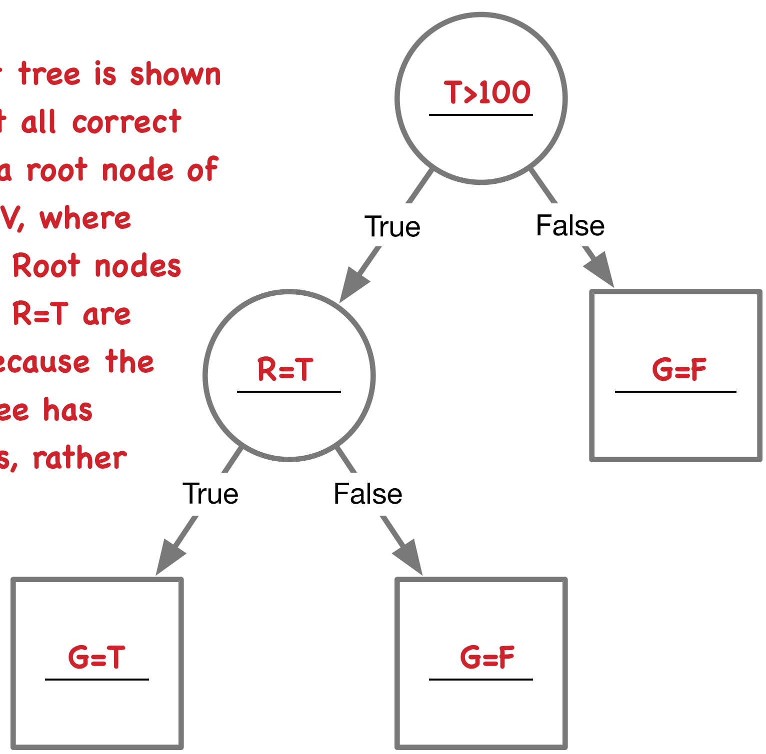
True	False	Statement
●	○	Classification trees are frequently learned by *recursive partitioning* of the training set.
○	●	The only learned parameters of a classification tree are the variables and split points for decision nodes (for example, $X_3>5$).
●	○	The max-depth of a classification tree is an example of a hyper-parameter.
●	○	Classification trees are a non-parametric model.
●	○	A regression tree is similar to a classification tree, except that the output variable is continuous rather than categorical.

Name ________________________

3. Classification tree construction (10 points) — Below is a small data set in which rows describe patients and columns are variables describing the symptoms and diagnosis of the disease ghoulisitis. Based only on the data set below, complete the classification tree that a standard learning algorithm would construct to predict the occurrence of ghoulisitis. Assume an algorithm that continues partitioning until all data at a leaf node has a single class. Assume that the algorithm evaluates possible decision nodes using the percentage of class labels (G) that are predicted correctly. Complete the tree below by filling in a test (of the form "W=F" or "T>103") for each decision node and the predicted class label (of the form "G=T") for each leaf node.

Ghoulisitis? (G)	White Complexion (W)	Red eyes (R)	Temperature (T)
T	F	T	104
T	T	T	101
T	T	T	103
T	T	T	102
T	F	T	103
F	F	T	97
F	T	T	98
F	F	T	98
F	T	F	102
F	F	F	103

One correct tree is shown at right, but all correct trees have a root node of the form T>V, where V=[98,101]. Root nodes of the form R=T are incorrect because the resulting tree has three errors, rather than two.



Name ________________________

4. Feature selection (10 points) — The table at right shows each possible subset of four variables and an estimated score for a model that uses that subset of variables. For example, the entry "{1,4}" indicates that a model uses variables 1 and 4 (but not 2 and 3) and receives a score of 0.67. Higher values of the score indicate a more accurate model.

In the spaces below, fill in the subset of variables that will be selected by each method for feature selection.

a. {2,3} Best subset selection

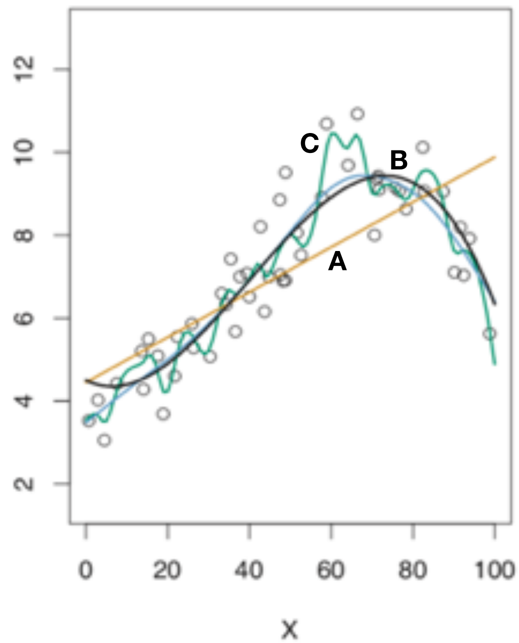
b. {1,3,4} Forward subset selection

c. {2,4} Backward subset selection

Subset	Score
{}	0.50
{1}	0.67
{2}	0.32
{3}	0.51
{4}	0.64
{1,2}	0.69
{1,3}	0.82
{1,4}	0.67
{2,3}	0.99
{2,4}	0.91
{3,4}	0.95
{1,2,3}	0.81
{1,2,4}	0.87
{1,3,4}	0.85
{2,3,4}	0.80
{1,2,3,4}	0.75

Name ________________________

5. **Model capacity** (10 points) — The figure at right shows three models, each of which uses *x* to predict *y*. Each model has a different capacity. Model A is as a straight line, model B a smooth curve, and model C as a wavy curve. Answer each question below.



a. ___C___ Which model has the highest capacity? A, B, or C?

b. ___B___ For model A, which type of errors are likely to be large, compared to the other models? Bias (B), variance (V), or noise (N)?

c. ___V___ For model C, which sort of errors are likely to be large, compared to the other models? Bias (B), variance (V), or noise (N)?

d. ___C___ Which model is likely to have the largest difference between training and test accuracy? A, B, or C?

e. ___D___ If model capacity was held constant and sample size was increased, what would happen to the variance component of error for the models? Would it increase (I), decrease (D) or stay constant (C)?

Name ______________________

6. **Multiple comparison procedures** (10 points) — Below are descriptions of methods for training and estimating the accuracy of two models (A and B). For the estimator described in the last sentence of each description, indicate whether Model A can be expected to have higher estimated accuracy (A), Model B can be expected to have higher estimated accuracy (B), or you can't tell (C). Note that the question is about *estimated* accuracy (for the designated sample), not true accuracy (for the population of all possible data instances). Note also that the question asks about *expected* accuracy (on average, over many repetitions).

   a. ___A___ Both models are trained on the same training set and have equal true accuracy on the population. The accuracy of Model A is estimated using the training set and the accuracy of Model B is estimated using a new test set drawn from the same population as the training set.

   b. ___B___ Both models are classification trees and both are trained using the same training set. Model A has lower capacity than model B. The accuracy of both models is estimated using the training set.

   c. ___A___ Two training sets and one test set, each containing 1000 instances, are drawn from a very large population of instances. All models are classification trees. Model A is created by training one model on each of 10 bootstrap samples from the first training set and selecting the model that performs best on the test set. Model B is created by training a model on each of 5 bootstrap samples from the second training set and selecting the model that performs best on the test set. The accuracy of Model A and Model B are estimated using the test set.

   d. ___C___ Model A is a classification tree and Model B is a simple Bayesian classifier. Both are trained using the same training set. Model A has higher estimated accuracy than Model B when tested on the training data. The accuracy of both models is estimated using a new test set drawn from the same population as the training set.

   e. ___B___ The situation is exactly the same as described in (d), except in this case Model A has *lower* estimated accuracy than Model B when tested on the training data. As before, the accuracy of both models is estimated using a new test set drawn from the same population as the training set.

Name ________________________

7. **Bias and variance** (10 points) — Suppose we wish to predict student responses to a campus survey on alcohol consumption. In each case, our *estimator* is a simple percentage based on a sample of students surveyed. In each row of the table below, fill in values for the last two columns to indicate whether using the estimator with the stated sample would produce bias and variance that is higher (+), lower (–), or equal (0) *when compared to an estimator that uses a random sample of 100 UMass students.*

Estimand	**Sample**	**Effect on bias**	**Effect on variance**
The percentage of UMass students opposed to on-campus drinking	A random sample of 500 UMass students	0	-
The percentage of UMass students in favor of on-campus drinking	A random sample of 500 UConn students	+	-
The percentage of all Massachusetts college students in favor of on-campus drinking.	A random sample of 100 Massachusetts college students	-	0 | +
The percentage of US college students in favor on-campus drinking	A random sample of 100 Massachusetts college students	-	0 | +
The percentage of CMPSCI students opposed to on-campus drinking	A random sample of 50 CMPSCI students	-	(Any)

In some cases, the answer is unclear, because the inherent variability of the elements of one sample is greater than another (e.g., UMass students vs. CMPSCI students) and this can trade off with another effect (e.g., larger sample size). Therefore, we accept both answers (e.g., "0 | +") or all possible answers ("Any").

Name ________________________

8. Cross-validation (10 points) — Given a 10-fold cross-validation procedure applied to a data set with 1000 instances, answer the questions below.

a. ___9___ How many *times* is each data instance used for *training*? (Answer with an integer)

b. ___1___ How many *times* is each data instance used for *testing*? (Answer with an integer)

c. ___N___ Is the *variance* of models constructed during cross-validation exactly the same as models constructed using the entire set of data? (Answer with Yes (Y) or No (N))

d. ___[9,11]___ Cross-validation increases the runtime of a learning algorithm. Roughly *how many times longer* would we expect the runtime to be of an algorithm that uses cross-validation as opposed to a baseline algorithm that used the same data set without cross-validation? (Answer with an integer)

e. ___N___ Is using a training set with 50% of all available data and a test set with the other 50% of all available data is equivalent to *two-fold cross-validation*? (Answer with Yes (Y) or No (N))

Name ________________________

9. ROC curves (10 points) — The table at right shows the actual class labels and the score of a ranking classifier. For the ROC curve corresponding to this classifier, fill in the missing entries in the table below.

x-axis value of the ROC curve	y-axis value of the ROC curve
0	0
0.25	0.50
0.1	0.45
0.70	0.80
1	1

ID	Label	Score
1	+	0.99
2	+	0.96
3	–	0.95
4	+	0.86
5	+	0.85
6	+	0.79
7	–	0.70
8	–	0.67
9	+	0.60
10	–	0.57
11	+	0.55
12	–	0.40
13	+	0.39
14	–	0.35
15	–	0.32
16	–	0.30
17	+	0.22
18	–	0.21
19	+	0.05
20	–	0.01

Name ________________________

10. Ensembles (10 points) — Answer each question below.

a. ___N___ Is reducing bias the primary way that bagging improves accuracy? (Answer with Yes (Y) or No (N)).

b. ___Y___ Does a bootstrap sample contains fewer unique data instances than appear in the original data sample? (Answer with Yes (Y) or No (N)).

c. ___S___ Which method learns a meta-classifier? (Answer with Bagging (G), Boosting (O), or Stacking (S)).

d. ___G,O,S___ Which ensemble method are implemented by random forests? (Answer with Bagging (G), Boosting (O), or Stacking (S)).

e. ___O___ Which ensemble method re-weights data instances? (Answer with Bagging (G), Boosting (O), or Stacking (S)).