EECS 349 Machine Learning

Exam 1

April 21, 2017

Name: ____

1. (0.5 points) Your friend Steve says ”Hey, I have a new machine learning algorithm that will always work better than any other algorithm, for all functions you're trying learn.” Do you believe Steve? Explain why or why not in one sentence.

2. (0.25 points) Carl experiments with a binary classifier over ten attributes. He achieves ten-fold cross validation accuracy of 96%. ZeroR, which ignores the attributes and just returns the most common class in the training data, achieves a 10-fold CV accuracy of 97%. Carl's accuracy is high, but should he be satisfied that his learner is good?

   (a) Probably not, if Carl's classifier is good it should probably beat ZeroR's accuracy.

   (b) Probably so, nobody expects a classifier to be able to beat ZeroR.

3. (0.25 points) Zheng splits his data into a training set of 700 examples, a validation set of 300 examples, and a test set of 500 examples. He evaluates 99 different k nearest neighbor models, trying each $k \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21\}$ and each Lp distance for $p \in \{1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5\}$. He trains on the training set and evaluates on the validation set. His best accuracy on the validation set across the 99 runs is 64.0%, achieved by the classifier using $L3$ distance and $k = 5$. If Zheng evaluates that same nearest neighbor classifier (still trained on the same 700 examples) on the *test* set, if you had to bet on what the accuracy would be, how would you bet?

   (a) The test set accuracy will probably be *lower* than 64.0%.

   (b) The test set accuracy will be *exactly* 64.0%.

   (c) The test set accuracy will probably be *higher* than 64.0%.

4. (0.25 points) Candice applies 10-fold CV to a data set of 1,000 examples. During the CV, what's the expected number of times Candice *tests* a model on the first example e_1 from the data set?

   (a) 1

   (b) 10

   (c) 100

   (d) 1000

5. (0.25 points) Claire experiments with decision trees, trying both 2-fold CV and 10-fold CV on the same data set using the same classifier settings in both cases. Which is likely to be higher?

   (a) The average accuracy from the 2-fold CV

   (b) The average accuracy from the 10-fold CV

6. (0.25 points) In Claire's experiments, which CV is likely to take *more time* to complete?

   (a) The 2-fold CV

   (b) The 10-fold CV

7. (1.25 points) For the following learning tasks, choose the best approach from Decision Trees (D), Linear Regression (L), Nearest Neighbor (N), and Perceptrons (P). Each letter should be used once.

   (a) _____ The attributes and the output are categorical, and it's a high priority that the trained model outputs a classification quickly for a given test example.

   (b) _____ The attributes are numeric, the output is categorical, you believe the target function is complex, and fast training times are desired.

   (c) _____ The attributes are numeric, the output is binary, and you believe the target function is linearly separable.

   (d) _____ The attributes are numeric, the output is numeric, and you believe the target function is a weighted average of the attributes.

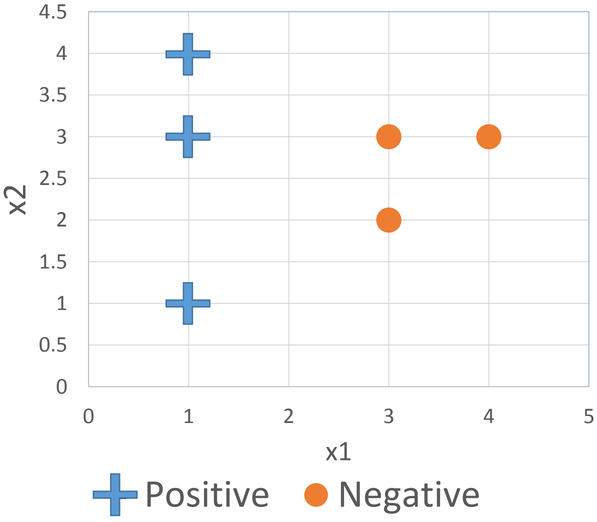


Figure 1: Training Set.

8. (2.5 points) Consider performing Leave-One-Out cross validation (LOOCV) on the data set shown in Figure 1. This is a binary classification task with two continuous attributes x1 and x2, and the data set contains six examples (three positive, three negative). What is the LOOCV accuracy of each of the following five methods? Express your answer for each as a fraction (e.g. $5/6$).

   (a) _____ Decision Trees[1]

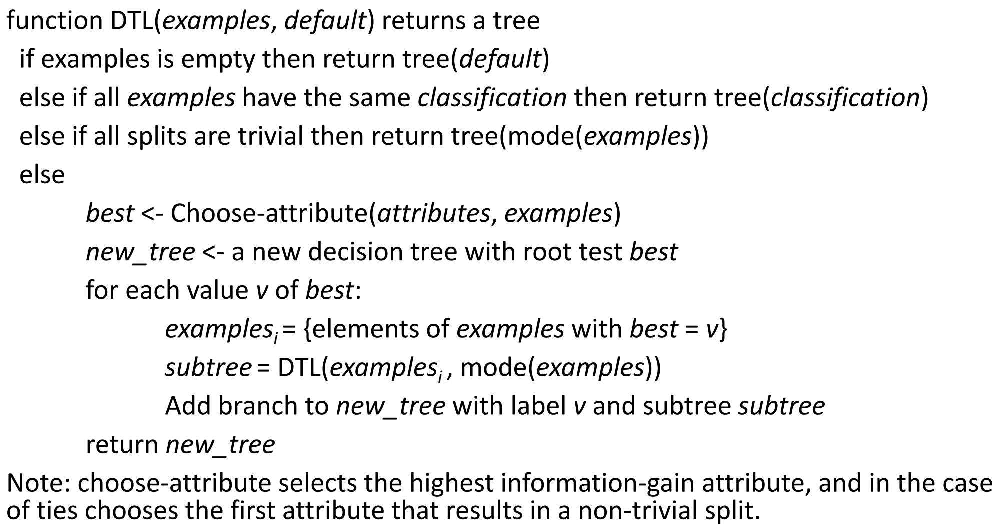
   (b) _____ 1-nearest-neighbor using L1 as the distance metric.

   (c) _____ 1-nearest-neighbor using L2 as the distance metric.

   (d) _____ 3-nearest-neighbor using L1 as the distance metric.

   (e) _____ 3-nearest-neighbor using L2 as the distance metric.

[1] The decision-tree learner is a standard DT learner like we discussed in class, and its details are not critical for answering the question. In general you can assume it follows the algorithm in Figure 2, augmented to handle continuous attributes by making binary splits on one attribute at a time, choosing an attribute and threshold that maximizes information gain.



```
function DTL(examples, default) returns a tree
  if examples is empty then return tree(default)
  else if all examples have the same classification then return tree(classification)
  else if all splits are trivial then return tree(mode(examples))
  else
          best <- Choose-attribute(attributes, examples)
          new_tree <- a new decision tree with root test best
          for each value v of best:
                    examples_i = {elements of examples with best = v}
                    subtree = DTL(examples_i , mode(examples))
                    Add branch to new_tree with label v and subtree subtree
          return new_tree
Note: choose-attribute selects the highest information-gain attribute, and in the case
of ties chooses the first attribute that results in a non-trivial split.
```

Figure 2: Decision tree algorithm.

9. (3.5 points) Consider the following training set for a binary classification task with three categorical (binary) attributes A, B, and C. Draw the decision tree learned over these examples, using the algorithm in Figure 2 (which is identical to the algorithm you were asked to implement in Problem Set 2). You should be able to intuit which split has highest information gain, but if it is helpful you can use the fact that: $-\frac{2}{3}\left(\frac{1}{4}\lg\frac{1}{4}+\frac{3}{4}\lg\frac{3}{4}\right) > -\frac{1}{3}\left(\frac{1}{2}\lg\frac{1}{2}+\frac{1}{2}\lg\frac{1}{2}\right)$

A	B	C	$f(A,B,C)$
0	1	1	1
0	0	1	0
1	0	1	0
1	1	0	0
1	1	0	0
1	1	1	0

10. (1 point) The DTL algorithm in Figure 2 includes the base case “else if all splits are trivial then return tree(mode(*examples*)).” A trivial split is one in which all examples sort to the same branch. What if we created a different algorithm DTL', in which we replaced that base case with “else if all splits have zero information gain, return tree(mode(*examples*)).” For the binary classification task over two binary attributes A and B, fill in the $f(A, B)$ values in the below training data, such that DTL' achieves 50% training accuracy on the dataset, whereas the original DTL achieves 100% training accuracy.

A	B	$f(A, B)$
0	0	
0	1	
1	0	
1	1	