CS 188 — Machine Learning: Midterm practice

Winter 2017

Instructions:

1. This exam is closed book and closed notes. You may use scratch paper.
2. The time limit for the exam is 1hour, 45 minutes.
3. Mark your answers ON THE EXAM ITSELF. If you make a mess, clearly indicate your final answer (box it).
4. For true/false questions, CIRCLE True OR False **and** provide a brief justification for full credit.
5. Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one).
6. If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.

1. Consider the following training data

x_1	x_2	class
1	1	$+$
2	2	$+$
2	0	$+$
0	0	-
1	0	-
0	1	-

Are the classes $\{+, -\}$ linearly separable?

2. Design a decision tree that takes two attributes x_1 and $x_2, x_i \in \{0, 1\}$ as input and computes the boolean function XOR over these inputs.

3. Which of these functions is convex ?

   (a) e^{-x}

   (b) $\log(x)$

   (c) $\cos(x)$

   (d) $-ax$ where $a > 0$

4. (1 pts) (True/False) Past a certain point, increasing the training set size (by drawing more instances from the underlying distribution) will cause the training error to begin to exceed the test error.

5. (1 pts) (True/False) Expanding the hypothesis space (the number of functions $\mathcal{H}$ that the learning algorithm searches over) can decrease the training error.

6. If your model is underfitting, increasing the number of attributes (features) will tend to (circle all that are correct)

   (a) increase the training error

   (b) decrease the training error

(c) increase the test error

(d) decrease the test error

7. (1 pts) (True/False) A function $f(x, y, z)$ is convex if and only if the Hessian of f is positive semi-definite.

8. (1 pts) (True/False) The perceptron training algorithm is guaranteed to converge with zero training error if the two classes are linearly separable.

9. (1 pts) (True/False) The training error of k-NN is smallest for $k = 1$.

10. Consider the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$. The derivative $f'(x)$ is

    (a) $f(x)log(1 - f(x))$

    (b) $f(x)(1 - f(x))$

    (c) $\frac{1}{f(x)}$

    (d) $f(x)(1 + f(x))$

11. (1 pts) (True/False) Two competitors are trying to solve the same logistic regression problem for a dataset using gradient descent with carefully chosen step sizes. One group claims that their initialization point will lead to a much better optimum.

12. You trained a decision tree for digit recognition and notice an extremely low training error, but an abnormally large test error. What could be the cause(s) of his problem?

    (a) Decision tree is too deep

    (b) Decision tree is overfitting

    (c) There is too much training data

    (d) All of the above.

13. IF N is the number of instances in the training dataset, nearest neighbors has a classification run time of

    (a) O(1)

(b) O(N)

(c) O($\log N$)

(d) O(N^2)

14. You just finished training a decision tree for spam classification, and it is getting abnormally bad performance on both your training and test sets. You know that your implementation has no bugs, so what could be causing the problem?

    (a) Your decision trees are too shallow.

    (b) You need to increase the learning rate.

    (c) You are overfitting.

    (d) All of the above.

15. Decision Trees

The following training instances with attributes Size, Sweetness, and HasChocolate and classification Popular are given:

Size	Sweetness	HasChocolate	classification
L	Very	Y	Popular
L	Very	N	Unpopular
M	Very	N	Unpopular
L	Mild	Y	Popular
S	Medium	Y	Unpopular
L	Very	Y	Popular
S	Mild	Y	Popular
S	Mild	N	Unpopular

(a) Consider learning a decision tree. Using information gain, which attribute will the decision tree learning algorithm choose as the root?

For reference, for a random variable X that takes on two values with probability p and $1-p$, here are some values of the entropy function:

$$p \in \{\tfrac{2}{5}, \tfrac{3}{5}\} : H(X) \approx .97 \qquad p \in \{\tfrac{1}{5}, \tfrac{4}{5}\} : H(X) \approx .72$$

$$p \in \{\tfrac{1}{4}, \tfrac{3}{4}\} : H(X) \approx .81 \qquad p \in \{\tfrac{1}{3}, \tfrac{2}{3}\} : H(X) \approx .92$$

(b) Now consider learning a binary decision tree. Construct (however you want) a decision tree with zero training error.