

## CISC 484/684 — Intro to Machine Learning: Mid-term Practice

**True/False**

For each question, check either True or False. If False, explain why.

7 points for True, 3 points for False, 4 points for an explanation for False.

1) The support vectors in soft-margin SVM are only the examples misclassified.

True

False

Explanation if False:

2)  $\mathcal{L}(y_i, f(x_i)) = y_i - 3f^2(x_i)$  is a valid loss function for linear regression.

True

False

Explanation if False:

3) Let  $K_1$  and  $K_2$  be any two kernel functions.  $K(x_i, x_{i'}) = K_1(x_i, x_{i'}) + K_2(x_i, x_{i'})$  is also a kernel function.

True

False

Explanation if False:

4) Batch least squares linear regression has a closed form solution for the model parameters given the data matrix  $\mathbf{X}$ . Therefore, we cannot use stochastic gradient in a setting where we are given one regression example at a time.

True

False

Explanation if False:

5) Adding regularization to an objective function can help to prevent model under-fitting.

True

False

Explanation if False:

6) Suppose we have a learning algorithm  $A$  with hypothesis class  $H$  and  $A'$  with hypothesis class  $H'$ . If  $H \subset H'$ , meaning that  $H$  is a subset of  $H'$ , then we should prefer  $A$  over  $A'$ .

True

False

Explanation if False:

7) A decision tree is capable of learning any boolean function.

True

False

Explanation if False:

8) If logistic regression obtains 100% training accuracy, the algorithm has overfitted the training data and is guaranteed to do poorly on the test data.

True

False

Explanation if False:

9) We learned in class an algorithm A which produces an optimal decision tree (the smallest tree with the highest accuracy) in polynomial time.

True

False

Explanation if False:

10) The kernel trick applies to any algorithm in which we can write the relationship between  $x$  and  $x'$  as an inner-product.

True

False

Explanation if False:

## Short Answer

### 11) Regularization

A common technique to controlling the bias versus variance tradeoff of a learning algorithm is to add a regularization term to the objective function. For each of the following objectives, state whether increasing the regularization parameter increases bias or variance.

(a) The regularization parameter  $C$  in

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & (wx_i) y_i + \xi_i \geq 1, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

(b) The regularization parameter  $\lambda$  in

$$\max_w \prod_{i=1}^N \{h(w \cdot x)^y (1 - h(w \cdot x))^{1-y}\} + \lambda \sum_{j=1}^M w_j^2,$$

where  $h$  is the logistic function.

### 12) SVM

We learned about an algorithm with the following objective function.

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & (w \cdot x_i) y_i + \xi_i \geq 1, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

We will consider several modifications to this algorithm.

(a) Consider the following modification:

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|_1^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & (w \cdot x_i) y_i + \xi_i \geq 1, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

where we now take the L-1 norm of  $w$ . How will this change solutions for  $w$  ?

(b) Consider the following modification:

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^N |\xi_i| \\ \text{s.t.} \quad & (w \cdot x_i) y_i + \xi_i \geq 1, \forall i \end{aligned}$$

where we have removed the restriction that  $\xi_i$  be non-negative, and now use the absolute value in the objective. How will this change solutions for  $w$  ?

(c) Consider the following modification:

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & ((w \cdot x_i) y_i)^2 + \xi_i \geq 1, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

where we now square part of the term in the constraints. The prediction rule is unchanged. How will this change solutions for  $w$  ?

13) **Loss functions**

For each of the following algorithms name the corresponding loss function:

(a) Support Vector Machines

(b) Logistic Regression

(c) Linear Regression

14) **Decision Trees**

Suppose we wanted to learn a decision tree.

(a) What information theoretic metric would you use to select features for each node of the tree?

(b) Will increasing the depth of the tree increase bias or variance? Why?