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:

False. They also include the examples classified correctly, but lie on the margin boundary or inside the margin area.

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

☐ True

☐ False

Explanation if False:

False. A valid loss function must be non-negative. This can be negative if $y_i < 3f^2\left(x_i\right)$.

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

☐ True

☐ False

Explanation if False:

True

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:

False. While a closed form exists, we can solve least squares linear regression using a gradient based method, and thus use stochastic gradient ascent/descent.

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

☐ True

☐ False

Explanation if False:

False. Regularization helps prevent over-fitting.

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:

False. The size of the hypothesis class doesn't matter. What matters is our ability to find the best hypothesis within the class.

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

☐ True

☐ False

Explanation if False:

True

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:

False. While this is likely, it's not a guarantee.

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:

False. Optimal decision tree learning is NP-hard.

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:

True

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

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

Increases variance.

(b) The regularization parameter λ in

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

where h is the logistic function.

Increases bias.

12) **SVM**

We learned about an algorithm with the following objective function.

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

We will consider several modifications to this algorithm.

(a) Consider the following modification:

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

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

w will be sparse

(b) Consider the following modification:

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

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

No change. We can now decrease the margin on certain examples, but there is no advantage to doing that.

(c) Consider the following modification:

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

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

The constraint may be satisfied even when we have poor w, so the choice of w may get worse in terms of accuracy.

13) **Loss functions**

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

(a) Support Vector Machines Hinge loss

(b) Logistic Regression Logistic loss

(c) Linear Regression Squared loss

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?

Information gain

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

Increases variance. Deeper trees overfit the data.