

2. (Perceptron)

- a. (5pt) In class we derived the perceptron algorithm as gradient descent on the hinge loss objective function given by

$$J(w) = \frac{1}{N} \sum_{i=1}^N \max(0, -y_i w \cdot x_i)$$

where x_i is the input feature vector for the i 'th training instance and y_i is the label for the i 'th instance. We also consider using an alternative loss function known as the 0/1 loss given by,

$$J(w) = \frac{1}{N} \sum_{i=1}^N L(\text{sgn}(w \cdot x_i), y_i)$$

where $L(y, y') = 1$ if $y \neq y'$ and 0 otherwise.

What was the reason for using the hinge loss instead of 0/1 loss?

- b. (6pt) Consider two different weight vectors w and w' that are consistent with the training data where w has a large margin and w' has a very small margin. Which weight vector will achieve a better hinge loss? Explain.

- c. (9pt) Derive the gradient of the hinge loss objective $\nabla_w J(w)$.

3. (Support Vector Machines)

- a. (6pt) In class we introduced two notions of margin: the *functional margin* and the *geometric margin*. When deriving the SVM optimization problem why did we choose to maximize the geometric margin rather than the functional margin? Specifically, what would go wrong if we simply tried to maximize the functional margin and how does maximizing the geometric margin correct this problem?

- b. (7pt) Consider the following constrained optimization problem for finding the maximum margin classifier:

$$\begin{aligned} & \max_{w,b,\gamma} \gamma \\ \text{subject to: } & y^i \frac{(w \cdot x^i + b)}{\|w\|} \geq \gamma, \quad i = 1, \dots, N \end{aligned}$$

Prove that this problem is equivalent to the following optimization problem solved by SVMs. Make sure that you explain each step.

$$\begin{aligned} & \min_{w,b} \|w\|^2 \\ \text{subject to: } & y^i (w \cdot x^i + b) \geq 1, \quad i = 1, \dots, N \end{aligned}$$

- c. (7pt) Consider the objective function that the softmargin SVM tries to minimize:

$$\|w\|^2 + C \sum_i \xi_i$$

Suppose that we have a linearly separable training set. Is it true that the solution weight vector w will achieve zero training error. If true then explain why. If false then show a 1-d linearly separable training set where you would expect the training error will be non-zero, explain.

4. a. (4pt) Consider a training set whose labels were randomly corrupted. For the k-nearest neighbor classifier, which of the following choices of k is more robust to the labeling noise: k=1 and k=4? Explain
- b. (4pt) Consider a training set that has a relatively large number of completely random features, that is, features that have no correlation with the class label. Which algorithm would you choose for this data set k-nearest neighbor or a decision tree learner? Explain.
- c. (4pt) Why is it a good idea to initialize the weights of a neural network to be close to zero?
- d. (4pt) Why might it be a bad idea to initialize all weights of a neural network to be zero?
- e. (4pt) Consider learning a predictor for whether someone has cancer or not. If the predictor predicts cancer then the subject will be sent to a doctor, otherwise the subject will be sent home. Write down a loss function $L(\hat{y}, y)$ that you think is appropriate for this task. Explain.

5. (PAC Learning)

- a. (10pt) In class we showed that if a consistent learning algorithm for a finite hypothesis space H is provided with

$$m \geq \frac{1}{\epsilon} \left(\ln |H| + \ln \frac{1}{\delta} \right)$$

randomly drawn training instances, then we can state a certain guarantee. What is that guarantee? Make sure to clearly indicate the roles of ϵ and δ .

- b. (10pt) Consider the hypothesis space H of decision stumps for an input space containing n binary features. That is, each hypothesis is a decision tree that contains exactly one binary test. Prove that H is PAC-learnable. You may use the inequality from part a if you would like.