Due: 2018/05/2

Machine Learning: Homework 4

Student Number:
Name:

Problem 1. The VC dimension, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest number of points (in some configuration) that can be shattered by H. Suppose with probability $(1 - \delta)$, a PAC learner outputs a hypothesis within error ϵ of the best possible hypothesis in H. It can be shown that the lower bound on the number of training examples m sufficient for successful learning, stated in terms of $VC(H)$ is

$$m \geq \frac{1}{\epsilon}(4 \log_2 \frac{2}{\delta} + 8VC(H) \log_2 \frac{13}{\epsilon})$$

Consider a learning problem in which $X = \mathcal{R}$ is the set of real numbers, and the hypothesis space is the set of intervals $H = \{(a < x < b) | a, b \in \mathcal{R}\}$. Note that the hypothesis labels points inside the interval as positive, and negative otherwise.

(a) What is the VC dimension of H?

(b) What is the probability that a hypothesis consistent with m examples will have error at least ϵ?

Problem 2. The concept space $\mathcal{C}$ is the region between two parallel lines, either $(x = a, x = b)$ or $(y = a, y = b)$ for $a < b$. That is, each concept $f \in \mathcal{C}$ is defined by two numbers, a and b and another boolean indicator that determines whether the lines are parallel to the x axis or the y axis. An example (x, y) is positive for the concept (X, a, b) if and only if $a \leq x \leq b$. An example (x, y) is positive for the concept (Y, a, b) if and only if $a \leq y \leq b$.
For the above concept space, give the VC dimension and prove that your answer is correct.

Problem 3. We consider here a discriminative approach for solving the classification problem illustrated in Figure1. We attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y = 1 | \overrightarrow{x}, \overrightarrow{\omega}) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2) = \frac{1}{1 + exp(-\omega_0 - \omega_1 x_1 - \omega_2 x_2)}.$$

Notice that the training data can be separated with zero training error with a linear separator. Consider training *regularized* linear logistic regression models where we try to maximize

$$\sum_{i=1}^{n} log(P(y_i | x_i, \omega_0, \omega_1, \omega_2)) - C\omega_j^2$$

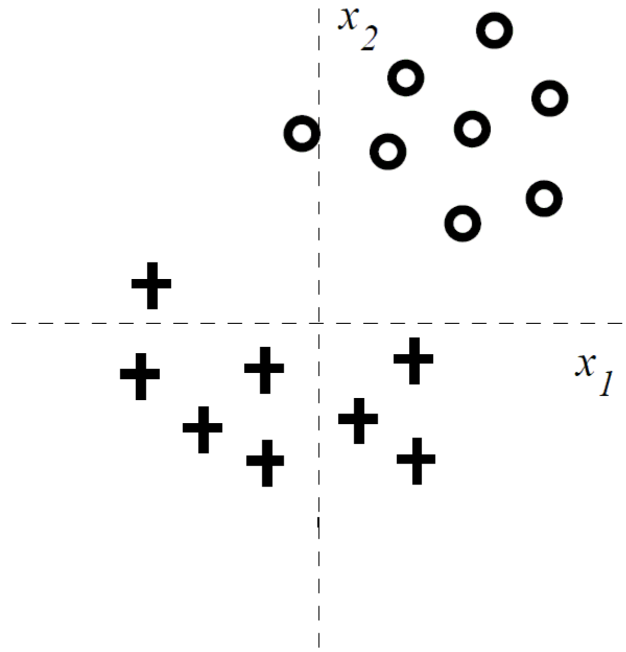

Figure 1: The 2-dimensional labeled training set, where '+' corresponds to class y=1 and 'O' corresponds to class y = 0.

for very large C. The regularization penalties used in penalized conditional log-likelihood estimation are $-C\omega_j^2$, where $j = \{0, 1, 2\}$. In other words, only one of the parameters is regularized in each case. Given the training data in Figure 1, how does the training error change with regularization of each parameter ω_j? State whether the training error increases or stays the same (zero) for each ω_j for very large C. Provide a brief justification for each of your answers.

(a) By regularizing ω_2

(b) By regularizing ω_1

(c) By regularizing ω_0

Problem 4. If we change the form of regularization in the problem 3 to L1-norm (absolute value) and regularize ω_1 and ω_2 only (but not ω_0),we get the following penalized log-likelihood

$$\sum_{i=1}^{n} log(P(y_i|x_i, \omega_0, \omega_1, \omega_2)) - C(|\omega_1| + |\omega_2|).$$

Consider again the problem in Figure 1 and the same linear logistic regression model

$$P(y = 1|\overrightarrow{x}, \overrightarrow{\omega}) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2).$$

(a) As we increase the regularization parameter C which of the following scenarios do you expect to observe? (Choose only one) Briefly explain your choice:
( ) First ω_1 will become 0, then ω_2 .
( ) First ω_2 will become 0, then ω_1 .
( ) ω_1 and ω_2 will become zero simultaneously.
( ) None of the weights will become exactly zero, only smaller as C increases.

(b) For very large C, with the same L1-norm regularization for ω_1 and ω_2 as above, which value(s) do you expect ω_0 to take? Explain briefly. (Note that the number of points from each class is the same.) (You can give a range of values for ω_0 if you deem necessary).

(c) Assume that we obtain more data points from the '+' class that corresponds to y=1 so that the class labels become unbalanced. Again for very large C, with the same L1-norm regularization for ω_1 and ω_2 as above, which value(s) do you expect ω_0 to take? Explain briefly. (You can give a range of values for ω_0 if you deem necessary).