Theory and Algorithm for Machine Learning 2020
National University of Singapore CS5339
Prof Lee Wee Sun

Homework 2 (Due Date Sunday 29 March 11.59pm)

Please write the following on your homework:

- Name
- Collaborators (write none if no collaborators)
- Source, if you obtained the solution through research, e.g. through the web.

While you may collaborate, you *must write up the solution yourself*. While it is okay for the solution ideas to come from discussion, it is considered as plagiarism if the solution write-up is highly similar to your collaborator's write-up or to other sources.
You solution should be submitted to IVLE workbin. Scanned handwritten solutions are acceptable but must be legible.
Late Policy: A late penalty of 20% per day will be imposed (no submission accepted after 5 late days) unless prior permission is obtained.

---

1. **MDL with multiplicative error bound**
   In deriving the error bound for MDL in class, we used Hoeffding's inequality and obtained the bound

$$L_{\mathcal{D}}(h) \leq \left[L_S(h) + \sqrt{\frac{|h| + \log(2/\delta)}{2m}} \right].$$

   Instead of using Hoeffding's inequality, we will use the following inequality due to Haussler. Consider m i.i.d. random variables $Y_1, \ldots, Y_m$ in range $[0, M]$ with expected value μ. Let $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} Y_i$. Assume $\epsilon > 0$ and $0 < \alpha < 1$. Then

$$P\left(|\hat{\mu} - \mu| > \alpha(\epsilon + \hat{\mu} + \mu)\right) \leq 2\exp\left(-\frac{\alpha^2 \epsilon m}{M}\right).$$

   (a) Rederive the MDL-type bound using Haussler's inequality instead of Hoeffding bound (i.e. obtain a bound for the expected loss in terms of the empirical loss and the hypothesis length). Use the MDL weight function $w(h) = 1/2^{|h|}$ for a hypothesis of length $|h|$.

   (b) What is the regularized cost function suggested by the new bound?

   (c) As m increases, compare the new bound with the bound obtained using Hoeffding's inequality for the case when $L_S(h)$ is small or zero, and for the case when $L_S(h)$ is large.

2. **Margin Vs Network Size**
   In this question, we will look at the error bounds provided using VC-dimension of a single hidden layer neural network when the inputs are points in the plane.

(a) Argue that the VC dimension of convex polygons with k or fewer vertices is at least k (where the points on the inside and boundary of the polygon is labeled positive).

(b) A convex polygon with k vertices can be represented as the intersection of k halfspaces. Using this, argue that the VC-dimension of a single hidden layer neural networks with k linear threshold hidden units is at least k when the inputs are points in the plane.

(c) Bound the Rademacher complexity of single hidden layer neural networks with arbitrary number of linear threshold hidden units when the ℓ_l norm of the output weights is 1 and the inputs are points in the plane.

(d) Assume that the margin is known to be at least γ when the single hidden layer neural networks has ℓ_1 norm equal to 1. Provide a bound on the error of an algorithm that maximizes the margin of the network. For simplicity, use the bounds without doing structural risk minimization in this question. (Hint: It may be useful to upper bound the $0 - 1$ loss with the hinge loss for the analysis.)

(e) Compare the bound using the margin and using the VC-dimension. When is each bound better?

3. **Estimation Error for ℓ_1 vs ℓ_2 regularization**
In the example discussed in class, we want to find a sequence of characters in a file that indicate whether the file contains a virus – a "signature". Consider a variant, where string indicating the positive class has length exactly d. That is, the target function we want to learn is $f_v(x) = 1$ iff v is a substring of x, and $f_v(x) = -1$ otherwise, where v is a string of length d. Naturally, the string v is unknown, otherwise there is no learning problem. We will use linear classifiers to try to learn the target function. As features, we will use indicator functions $\psi_u(x)$, where $\psi_u(x)$ is an indicator function that takes the value 1 if u is a substring of x and 0 otherwise. Here u ranges over all possible strings of length d. Let the file length be F. For simplicity, use the bounds without doing structural risk minimization.

(a) Give the estimation error bound of using hard support vector machine. That is, describe how you can represent the target function as a linear function with magnitude at least 1 on all inputs and $\|\mathbf{w}\|_2 \leq B$ for the weights. Let H be the class of linear function satisfying $\|\mathbf{w}\|_2 \leq B$. Then compute an upper bound for $L_D(h)$ assuming h belongs to H and has magnitude at least 1. For simplicity, consider the homogeneous case (bias implemented by having a feature that always has value 1). (Hint: It may be useful to upper bound the $0 - 1$ loss with the hinge loss for the analysis.)

(b) Now describe how you can represent the target function as a linear function with magnitude at least 1 on all inputs and $\|\mathbf{w}\|_1 \leq B$ for the weights. Let H be the class of linear function satisfying $\|\mathbf{w}\|_1 \leq B$. Then compute an upper bound for $L_D(h)$ assuming h belongs to H and achieves magnitude at least 1. For simplicity, consider the homogeneous case (bias implemented by having a feature that always has value 1). Assume that the number of possible characters is C. Which bound is better, compared to the hard SVM case?

(c) Consider the case when many substrings may be correct, i.e. the function should output 1 if any of the substrings in a subset S of substrings appears, and -1 otherwise. In this case, which bound is better, assuming $|S|$ is much larger than F?