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 $|h|$.

Solution: In the MDL derivation, we write H as a countable union of singleton classes, $H = \cup_{n \in \mathbb{N}} \{H_n\}$. Let hypothesis h be weighted $w(h)$. Let Y_i denote $\ell(x_i, y_i, h)$. From Haussler's bound, we get that with probability at least $1 - w(h)\delta$, the function h_n satisfies

$$L_D(h) - L_S(h) < \alpha(\epsilon + L_S(h) + L_D(h)).$$

$$L_D(h) < \frac{1+\alpha}{1-\alpha} L_S(h) + \frac{\alpha\epsilon}{1-\alpha}.$$

Setting $w(h)\delta = 2\exp\left(-\frac{\alpha^2 \epsilon m}{M}\right)$ and using the union bound, we get with probability at least $1-\delta$,

$$L_D(h) < \frac{1+\alpha}{1-\alpha} L_S(h) + \frac{M}{\alpha(1-\alpha)m}\left(\ln\frac{2}{\delta} + \ln(1/w(h))\right)$$

for all h. With the MDL weighting, $w(h) = 1/2^{|h|}$. Using $\ln(2^{|h|}) = |h|\ln 2 < |h|$, we get

$$L_D(h) < \frac{1+\alpha}{1-\alpha} L_S(h) + \frac{M}{\alpha(1-\alpha)m}\left(\ln\frac{2}{\delta} + |h|\right).$$

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

Solution: Structural risk minimization suggests minimizing the upper bound to $L_D(h)$. The bound derived in section (a) suggests finding h that minimizes

$$L_S(h) + \lambda|h|,$$

where λ is a constant. This is different from the previous regularized cost function which has the form

$$L_S(h) + \lambda\sqrt{|h|}.$$

(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.

Solution: The new bound has the form $L_D(h) < c_1 L_S(h) + c_2|h|/m$ while the previous bound has the form $L_D(h) < L_S(h) + c\sqrt{|h|/m}$. If $L_S(h)$ is zero (or very small), then the new bound which converges at a rate $O(1/m)$ to zero (or a very small value) would be preferable to the previous bound that converges at the rate $O(1/\sqrt{m})$. If $L_S(h)$ is large, the previous bound may be better as it converges to $L_S(h)$ instead of $c_1 L_S(h)$, where $c_1 > 1$. However, it converges at a rate $O(1/\sqrt{m})$.

No marks allocated for these further observations: Looking further at the bounds, what useful insights can the two bounds tell us regarding how we may want to regularize? Consider the case where $|h| = 0.01m$. In that case, $\sqrt{|h|/m} = 0.1$ while $|h|/m = 0.01$. In fact, we can increase $|h|$ to $0.1m$ for $|h|/m$ to reach 0.1. In other words, we may be able to regularize less or use a larger hypothesis with less concerns about overfitting in the case where there is a noiseless target function compared to the case where the target function is noisy.

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).

Solution: Select k points on the unit circle as the set S. For any subset $S' \subseteq S$, we can construct a convex polygon with $|S'|$ vertices where the vertices correspond exactly to the set S'. Since the points are on the unit circle, the polygon would exclude the set $S - S'$. As the set S' is arbitrary, the set of k points is shattered.

(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.

Solution: As discussed in class, the output layer can represent the AND function. By setting the output unit to be represent the AND function of the appropriate subset of k units, the single hidden layer neural network can represent a convex polygon with k vertices and hence shatter the same points as the convex polygon. The VC-dimension is hence at least k.

(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.

Solution: The VC-dimension of a hidden unit is 3 since it is a linear threshold unit with 2-dimensional input. By Sauer's lemma, the number of functions on m points is no more than $N \leq (em/3)^3$. Let this set of vectors (functions restricted to the m points) be $\{v_1, \ldots, v_N\}$. Let $V = \{v_1, \ldots, v_N\} \cup \{-v_1, \ldots, -v_N\}$. Then the size of V is no more than $2(em/3)^3$. With the norm of the output weights set at 1, the class becomes the convex hull of the of V. The convex hull has the same Rademacher complexity as the original class. Using Massart's lemma, we can bound the Rademacher by $\max_i \|v_i\|_\infty \sqrt{\frac{2\log(|V|)}{m}} \leq \sqrt{\frac{6\log m+2}{m}}$.

(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.)

Solution: From SSBD Theorem 26.5,

$$L_D(h) - L_S(h) \leq 2E_{S' \sim D^m} R(\ell \circ H \circ S') + c\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

We apply this on the hinge loss which upper bounds the $0-1$ loss. When the magnitude of the function is at least 1 on the training set, the empirical risk using the hinge loss is 0. We rescale the output weights so that the minimum magnitude of the function is 1, giving ℓ_1 norm of $1/\gamma$. This rescales the Rademacher complexity to $\frac{1}{\gamma}\sqrt{\frac{6\log m+2}{m}}$. For the hinge loss $\rho = 1$, so composing with the hinge loss does not increase the Rademacher complexity. The hinge loss may add 1 to the magnitude of the function, so $c \leq 1 + \frac{1}{\gamma}$. Applying the Radamacher and bound on c we get

$$L_D(h) \leq \frac{2}{\gamma}\sqrt{\frac{6\log m + 2}{m}} + \left(1 + \frac{1}{\gamma}\right)\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

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

Solution: The margin bound does not depend on the number of hidden units. Since the VC-dimension is at least the number of hidden units, from the fundamental theorem the bound that uses VC-dimension would grow with the number of hidden units. So when the number of hidden units is large but the margin is large, the margin-based bound would be better. On the other hand, the VC-dimension is $O(n\log n)$ where n is the number of hidden units. So, if the margin is small but the number of hidden units is small, the VC-dimension bound may be better.

Observations on the question (no marks allocated). The analysis suggests that we may want to analyse the estimation error in different ways depending on whether we expect to have a neural network with large margin or a neural network with a small number of hidden units. It also suggests different possible optimization criterion for good performance: maximizing margin, or minimizing size (in practice we may try both and see which works 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.)

Solution: Set the bias to -1, set the output weight of the feature corresponding to the target string to 2, and set all other weights to 0. We have magnitude of 1 for all inputs, ℓ_2 norm B of the weights of $\sqrt{5}$ and ℓ_2 norm R of the input is upper bounded by $\sqrt{F}$ where F is the length of the file. When the magnitude of the function is 1, the hinge loss has value 0. Using the hard SVM error bound for the hinge loss (SSBD Theorem 26.12), we get

$$\begin{aligned} L_D(h) \quad & \leq 2BR\sqrt{\tfrac{1}{m}} + (1+BR)\sqrt{\tfrac{2\log(2/\delta)}{m}} \\ & \leq 2\sqrt{5F}\sqrt{\tfrac{1}{m}} + (1+\sqrt{5F})\sqrt{\tfrac{2\log(2/\delta)}{m}} \end{aligned}$$

(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?

Solution: The ℓ_1 norm of the weight B is 3, while the ℓ_∞ norm of the inputs R is 1. Assume that the number of possible characters is C, so that the total number of features is C^d. The bound on the expected error (SSBD Theorem 26.15) is

$$\begin{aligned} L_D(h) \quad & \leq \tfrac{2BR}{\sqrt{m}}\sqrt{2\log(2C^d)} + \tfrac{1+BR}{\sqrt{m}}\sqrt{2\ln(2/\delta)} \\ & \leq \tfrac{6}{\sqrt{m}}\sqrt{2(d+1)\log(C)} + \tfrac{4}{\sqrt{m}}\sqrt{2\ln(2/\delta)}. \end{aligned}$$

Generally, we expect F to be large in comparison to the other parameters, so in cases of extreme sparsity like this one, the ℓ_1 regularization has better bound.

(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?

Solution: In this case the ℓ_1 norm of the weight B_1 is $2|S| + 1$ and the ℓ_∞ norm R_∞ of the input is 1, giving $B_1 R_\infty = 2|S| + 1$. Meanwhile, the ℓ_2 norm of the weight, B_2 is $\sqrt{4|S| + 1}$ and the ℓ_2 norm, R_2 of the input is $\sqrt{F}$, giving $B_2 R_2 = \sqrt{(4|S| + 1)F}$. As $|S|$ is much larger than F, the bound for ℓ_2 regularization is better.

Observations on the question (no marks allocated). The analysis shows that the estimation error bound may be different depending on the properties of the problem. Correspondingly, we may want to do different regularization for different problems depending on its properties. In practice, we may not know which regularization is better, so may try multiple types of regularization that tend to work well and use techniques such as cross validation to select among them.