FML-FA20-HW2-yg1758-Yixiang-Gao

yg1758

October 2020

Problem A1

By definition, for any hypothesis $\mathcal{H}$ and sample $\mathcal{S}$ we have the empirical Rademacher complexity defined as:

$$\hat{\mathcal{R}}_{\mathrm{S}}(\mathcal{H}) = \mathbb{E}_{\sigma}[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i)]$$

By sub-additivity of the supremum, we have

$$\hat{\mathcal{R}}(\mathcal{S}) \geq \sup_{h \in H} \mathbb{E}_{\sigma}[\frac{1}{m} \sum_{i=1}^{m} \sigma h(z_i)]$$

By linearity of Expectation, it's clear that the sum of expectation of random-noise σ_i's is 0, or

$$\mathbb{E}_{\sigma}[\frac{1}{m} \sum_{i=1}^{m} \sigma g(z_i)] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[\sigma_i] g(z_i) = 0$$

Hence we conclude that

$$\hat{\mathcal{R}_S}(\mathcal{H}) \geq 0$$

problem A2

Consider the function $f(h_1 + h_2) = h_1 h_2$. If $h_1 + h_2$ takes values of $0, 1$, then $f(h_1 + h_2) = 0$, and if $h_1 + h_2 = 2$, then $f(h_1 + h_2) = 1$. That is, $f(0) = 0$, $f(1) = 0$ and $f(2) = 1$
Then consider $|f(x) - f(y)|$, where x, y can be $0, 1, 2$. It implies that if $|f(x) - f(y)| = 0$, then $0 \leq |x - y| \leq 1$, and if $|f(x) - f(y)| = 1$, then $1 \leq |x - y| \leq 2$, and we can conclude that $|f(x) - f(y)| < |x - y|$, which means f is 1-Lipschitz. And, $f(H_1 + H_2) = H_1 H_2$

By Talagrand's Lemma:

$$\begin{aligned}
\hat{\mathcal{R}}(H) = \hat{\mathcal{R}}(H_1 H_2) \leq \hat{\mathcal{R}}(H_1 + H_2) &= \mathbb{E}_{\sigma}[\sup_{h_1 \in H_1, h_2 \in H_2} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (h_1(z_i) + h_2(z_i))] && (1) \\
&= \mathbb{E}_{\sigma}[\sup_{h_1 \in H_1, h_2 \in H_2} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h_1(z_i) + \frac{1}{m} \sum_{i=1}^{m} \sigma_i h_2(z_i)] && (2) \\
&= \mathbb{E}_{\sigma}[\sup_{h_1 \in H_1} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h_1(z_i)] + \mathbb{E}_{\sigma}[\sup_{h_2 \in H_2} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h_2(z_i)] && (3) \\
&= \hat{\mathcal{R}}(H_1) + \hat{\mathcal{R}}(H_2) && (4)
\end{aligned}$$

Problem B1

The goal is to find an upper bound of the growth function $\Pi_H(m)$.

First, consider each single intermediate layer node u, which has m inputs. The growth function is therefore $\Pi_C(m)$, which is the largest different number of different results each individual nodes can get. Then, consider that, if each node has $\Pi_C(m)$ different results, then, collectively, the intermediate layer's upper bound is $(\Pi_C(m))^k$.

Next, consider the final layer. For every single possible outcome of intermediate layer, the most possible number of outcomes from the output layer would be determined by the input layer's value, and can have at most $(\Pi_C(m))^k$ outcomes for some not well-behaved class C. Hence, the totally upper bound of growth function is $\Pi_H(m) \leq (\Pi_C(m))^k * (\Pi_C(m)) = (\Pi_C(m))^{k+1}$.

Problem B2

By the upper bound we have from $B1$, $\Pi_H(m) \leq (\Pi_C(m))^k * (\Pi_C(m)) = (\Pi_C(m))^{k+1}$, then it follows

$$\log_2(\Pi_H(m)) \quad \leq \quad \log_2(\Pi_C(m))^{k+1} = (k+1)\log_2(\Pi_C(m))$$

Then, by Sauer's lemma,

$$\begin{aligned}(k+1)\log_2(\Pi_C(m)) &\leq (k+1)\log_2(\frac{em}{d})^d \\ &= (k+1)d\log_2(\frac{em}{d})\end{aligned}$$

Now, let $x = d(k+1)$, $m = 2d(k+1)\log_2(e(k+1))$, $y = \frac{e}{d}$. Then it follows that $m \geq \log_2 e > 1$, and $xy = (k+1)e > 4$, then

$$m > x\log_2(my) = d(k+1)\log_2(\frac{em}{d}) \quad \text{and} \quad 2^m > \frac{em}{d}^{d(k+1)} \geq \Pi_H(m)$$

Hence

$$VC_{dim}(H) < 2(k+1)d\log_2(e(k+1))$$

Problem B3

Observe that the VC dimension of threshold functions is r. Hence, it follows that

$$VC_{dim}(H) < 2(k+1)r\log_2(e(k+1))$$

Problem C (4)

Graphs of CV error with plus or minus 1 std. Choose C from [2^{-10},2^{10}] Graphs as the following:

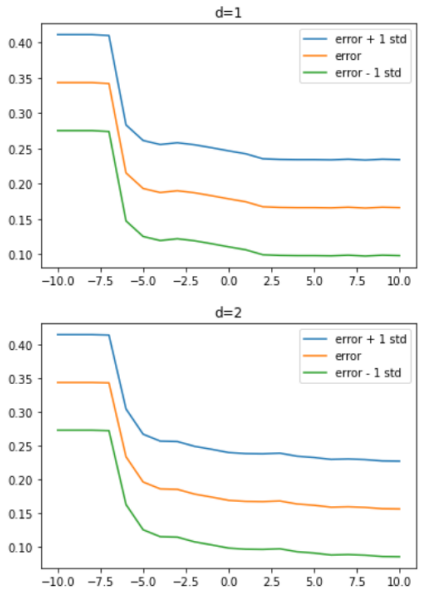

Figure 1: d = 1, d = 2, CV error with +/- 1 std

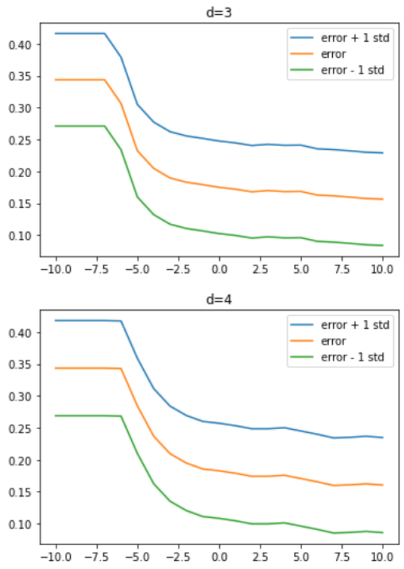

Figure 2: d = 3, d = 4, CV error with +/- 1 std

Problem C (5)

Graphs of test errors and CV-errors as d increases

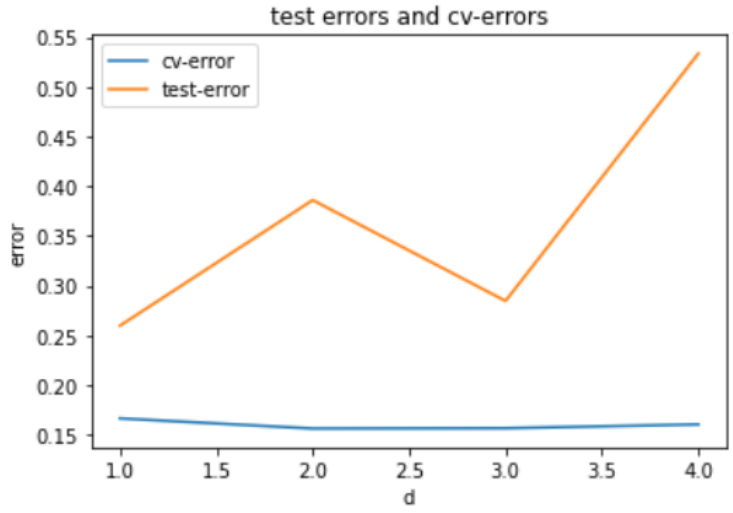

Figure 3: test errors and cv-errors as d increases

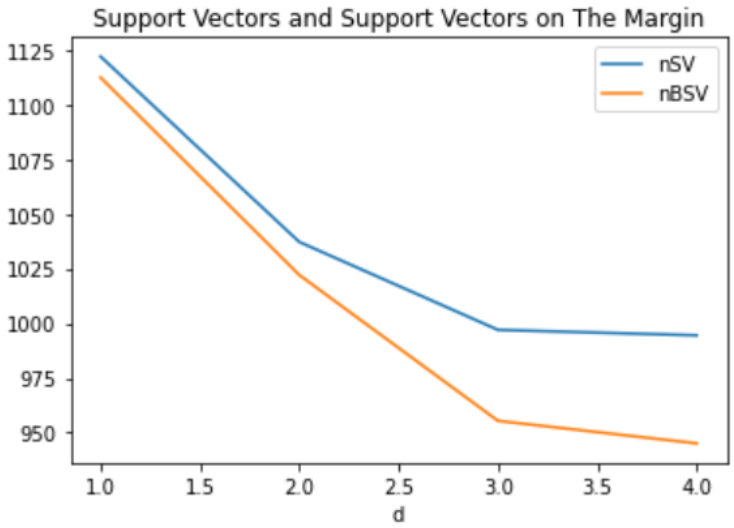

Figure 4: number of support vectors and number of support vectors on margin

Problem C6 (a)

(a) consider a different way to view the term $y_j K(xi, xj)$ as $x_i^* = [(y_1 K(x_i, x_1) \dots y_n K(x_i, x_n)]$. (a different way of looking at the product with kernel), then the problem changes to

$$\min_{\alpha,\beta} \frac{1}{2} \sum_{i=1}^{m} \alpha_i^2 + C \sum_{i=1}^{m} \xi_i$$

subject to $\quad y_i(\alpha \cdot x_i^* + b) \geq 1 - \xi_i, i \in [1, m]$

Problem C6 (b)

The optimization does not require the kernel function to be positive definite. With this specific form of optimization problem, we can show that it's always convex.

The objective function is obviously convex. Consider the constrain: let > 0, and consider the following two tuples $(\alpha_1, b_1, \xi_1), (\alpha_2, b_2, \xi_2)$, and $(\alpha_3, b_3, \xi_3) = \lambda(\alpha_1, b_1, \xi_1) + (1-\lambda)(\alpha_2, b_2, \xi_2)$. Then

$$\begin{aligned} y_i(\sum_{k=1}^{m}(\alpha_{3j}y_jK(x_i,x_k)+b_3) &= y_i(\sum_{k=1}^{m}(\lambda\alpha_{1k}+(1-\lambda)\alpha_{2k})y_kK(x_i,x_k)+\lambda b_1+(1-\lambda)b_2 \\ &= \lambda y_i(\sum_{k=1}^{m}(\alpha_{1k}y_kK(x_i,x_k)+b_1)+(1-\lambda)y_i(\sum_{k=1}^{m}(\alpha_{2k}y_kK(x_i,x_k)+b_2) \\ &\leq \lambda(1-\xi_{1i})+(1-\lambda)(1-\xi_{2i}) = 1-\xi_{3i} \end{aligned}$$

Hence we proved that the constraint is also convex. The optimization problem is therefore convex regardless of the property of kernel.

Problem C6 (c)

For this problem, the Lagrangian is

$$L = \frac{1}{2}||\alpha||^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\lambda_i[y_i(\sum_{j=1}^{m}\alpha_j x_j^* + b) - 1 - \xi_i] - \sum_{i=1}^{m}\phi_i\alpha_i - \sum_{i=1}^{m}\mu_i\xi_i$$

Then the F.O.C conditions will follow:

$$\nabla_{\alpha i}L = \alpha_i - \sum_{i=1}^{m}\lambda_i y_i x_i - \phi_i = 0 \rightarrow \alpha_i = \sum_{i=1}^{m}\lambda_i y_i x_i + \phi_i$$

$$\nabla_b L = \sum_{i=1}^{m} l_i y_i = \sum_{i=1}^{m}\lambda_i y_i = 0$$

$$\nabla_{\xi i}L = C - \lambda_i - \mu_i \rightarrow C = \lambda_i + \mu_i$$

In addition we have the condition $\lambda_i[(\sum_i^m y_i(\alpha_i x_i^* + b) - 1 + \xi_i] = 0$. From scratches, we have $\mu_i\xi_i = 0$ and $\alpha_i\phi_i = 0$. Which implies that

$$\sum_{i}^{m}\lambda_i[(\sum_{i}^{m}y_i(\alpha_i x_i^* + b) - 1 + \xi_i] = \sum_{i=1}^{m}\alpha_i(\alpha_i - \phi_i) + b\sum_{i=1}^{m}\lambda_i y_i - \sum_{i=1}^{m}\lambda_i + \lambda_i^m\lambda_i\xi_i$$

Finally

$$\begin{aligned} L &= \frac{1}{2}||\alpha||^2 + C\sum_{i=1}^{m} - \sum_{i=1}^{m}\lambda_i[y_i(\sum_{j=1}^{m}\alpha_j x_j^* + b) - 1 - \xi_i] - \sum_{i=1}^{m}\phi_i\alpha_i - \sum_{i=1}^{m}\mu_i\xi_i \\ &= \sum_{i=1}^{m}\lambda_i - \frac{1}{2}||\sum_{i=1}^{m}\lambda_i y_i x_i + \phi_i||^2 \end{aligned}$$

Which is the dual-problem we wanted.