# FML HW 2

Yucong Lei

Collaborated with Mengjian Hua(mh5113), Haiyang Wang(hw1927)

## 1 Problem A

### 1.1 A.1

**Proof**

$$\begin{aligned}\hat{\mathfrak{R}}_S(H) &= \mathbb{E}_\sigma[sup_{h\in H}\frac{1}{m}\sum_{1\le i\le m}\sigma_i h(z_i)] \\ &\ge \mathbb{E}_\sigma[\frac{1}{m}\sum_{1\le i\le m}\sigma_i h(z_i)], \forall h \in H \\ &= \frac{1}{m}\sum_{1\le i\le m}\mathbb{E}_\sigma[\sigma_i h(z_i)] = 0\end{aligned}$$

Hence it is nonnegative.

### 1.2 A.2

**Proof** Let $\Phi_i$ denote a continuous function such that:

$$\Phi_i(x) = 0, x \in (-\infty, 1]$$
$$\Phi_i(x) = x - 1, x \in (1, \infty)$$

It is easy to see that $\Phi_i$ is 1-Lipschitz, since:

$$|\Phi_i(x) - \Phi_i(y)| = 0, x, y \in (-\infty, 1]$$
$$|\Phi_i(x) - \Phi_i(y)| = |x - y|, x, y \in (1, \infty)$$
$$|\Phi_i(x) - \Phi_i(y)| = |x - 1| \le |x - 0| \le |x - y|, x \in (1, \infty), y \in (-\infty, 1]$$

In addition, we note that $\Phi_i$ has the following properties:

$$\Phi_i(0) = 0$$
$$\Phi_i(2) = 1$$
$$\Phi_i(1) = 0$$

Thus, $\forall h_1, h_2$, two binary classifiers, with values taken from 0,1, we have:

$$\Phi_i(h_1(z) + h_2(z)) = h_1(z)h_(z)$$

We can now apply Talagrand's contraction lemma:

$$\begin{aligned}\hat{\mathfrak{R}}_S(\Phi_i \circ (H_1 + H_2)) &\leq 1 * \hat{\mathfrak{R}}_S((H_1 + H_2)) \\ &\leq \hat{\mathfrak{R}}_S(H_1) + \hat{\mathfrak{R}}_S(H_2) \\ \rightarrow \hat{\mathfrak{R}}_S(H) = \hat{\mathfrak{R}}_S(H_1 * H_2) &= \hat{\mathfrak{R}}_S(\Phi_i \circ (H_1 + H_2)) \leq \hat{\mathfrak{R}}_S(H_1) + \hat{\mathfrak{R}}_S(H_2)\end{aligned}$$

# 2 Problem B

## 2.1 B.1

Consider the following neural network diagram (Figure 1). Each input data $\alpha_i \in \mathbb{R}^n$, where $1 \leq i \leq m$, we obtain a binary vector of dimension k in the intermediate layer of a given neural network, denoted as $\beta_i \in \{0,1\}^k, 1 \leq i \leq m$.

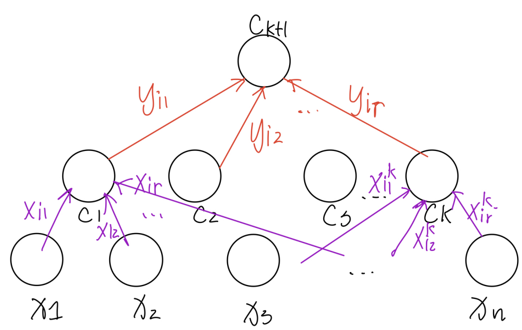


Figure 1:

Let us fix the input values, $\alpha_i \in \mathbb{R}^n$, where $1 \leq i \leq m$, and range all possible neural networks. Then the maximal distinct values are produced by ranging from different intermediate layer structures, that is, of which $r$ binary values out of the $k$ nodes are chosen as input value of the concept function, and ranging from different choices of concept values. Hence the following inequality hold:

$|\{(h(\alpha_1), ..., (h(\alpha_m)), h \in H\}|$

$\leq |\{(c \circ d(\beta_1), ..., (c \circ d(\beta_m)), c \in C, d \in \{A \in Mat(n, \mathbb{R}); A_{i,j} = 0, 1; \sum_j A_{i_p,j} = 1, 1 \leq p \leq r\}\}|$

$\leq \Pi_C(m)^{\binom{k}{r}}$

$\rightarrow \Pi_H(m) \leq \Pi_C(m)^{\binom{k}{r}}$

## 2.2 B.2

Let $a = 2^l \geq \binom{k}{r} \geq 1$. By the corollary of Sauer's lemma in the textbook, we have:

$$\begin{aligned} \Pi_H(m) &\leq \Pi_C(m)^{\binom{k}{r}} \\ &\leq (\frac{em}{d})^{da} \\ &\leq (\frac{8m}{d})^{da} \end{aligned}$$

Let $x = ad$, $y = \frac{8}{d}$, and $m = 2xlog_2(xy) = 2adlog_2(8a)$. We have:

$$\begin{aligned} x * y &> 4 \\ m &\geq 1 \\ m &\in \mathbb{Z} \end{aligned}$$

Hence by the hint, we have: $m > xlog_2(ym)$.

$$\begin{aligned} \Pi_H(m) &\leq (\frac{8m}{d})^{da} \\ &\leq 2^{ad*log_2 \frac{8m}{d}} \\ &< 2^m \end{aligned}$$

Hence, $VCdim(H) < m = 2adlog_2(8a)$.

## 2.3 B.3

We first try to compute $d$ in terms of $k, r$. Then we bound $VCdim(H)$ by the inequality obtained in the last question. We will make use of the fact that the VC dimension of hyperplanes in $\mathbb{R}^r$ is (r+1).

First we prove that the VC dimension of hyperplanes through the origin in $\mathbb{R}^r$ is no less than $r$, by showing that there exists $r$ samples, $\alpha_1, ..., \alpha_r \in \mathbb{R}^r$, such that:

$$|\{(c(\alpha_1), ..., c(\alpha_r)), c \in C\}| = 2^r$$

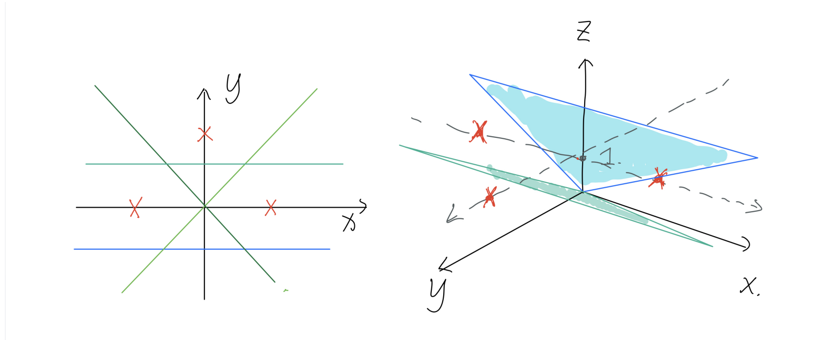


Figure 2:

Because the VC dimension of hyperplanes in $\mathbb{R}^{r-1}$ is $r$, there exists $r$ samples, $\beta_1, ..., \beta_r \in \mathbb{R}^{r-1}$, such that:

$$|\{(d(\beta_1), ..., d(\beta_r)), d \in \{sgn(w * \beta + b), w \in \mathbb{R}^{r-1}, b \in \mathbb{R}\}\}| = 2^r$$

We map all $\beta$ to $(\beta, 1) \in \mathbb{R}^r$, and note that this mapping is injective. Thus:

$$|\{(c(\beta_1, 1), ..., c(\beta_r, 1)), c \in \{sgn((w, b) * \alpha), w \in \mathbb{R}^{r-1}, b \in \mathbb{R}\}\}| = 2^r$$

We next prove that the VC dimension of hyperplanes through the origin in $\mathbb{R}^r$ has to be less than $(r+1)$, by showing that any set of $r+1$ points in $\mathbb{R}^r$ can only be completely classified by planes including a plane that does not pass through the origin.

We prove by contradiction. Suppose there are such $r+1$ samples in $\mathbb{R}^{r+1}$, $\alpha_1, ..., \alpha_{r+1} \in \mathbb{R}^r$, completely classified by hyperplanes through the origin. Then, we pick such a hyperplane through the origin that makes the label of every sample be the same, that is, they all lie in one side of a hyperlane through the origin. Assume without loss of generality, that this hyperplane is $x_r = 0$, and each $\alpha_i$'s r-coordinate is greater than 0.

We then note that no two sample points determine a line through the origin. To see why this is the case, we assume two samples are on a line passing through the origin that is:

$$(x_1, ..., x_r) = a(y_1, ..., y_r), \exists a \neq 1, 0, x_r, y_r > 0.$$

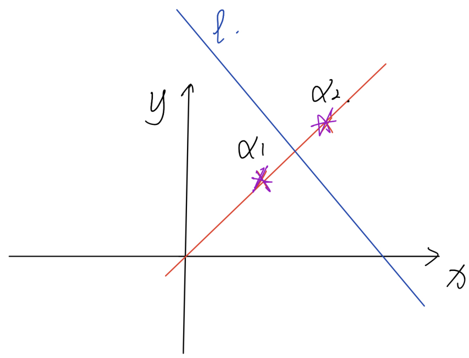


Figure 3:

If there exists a hyperplane that separates them, that is:

$$\begin{aligned}
&\exists w \in \mathbb{R}^r, b \in \mathbb{R}: \\
&w * x + b > 0 \\
&w * y + b < 0 \\
\rightarrow &\exists \lambda \in (0,1) \\
&w * (\lambda x + (1-\lambda)y) + b = 0 \\
\rightarrow &b = -w * (\lambda x + (1-\lambda)y) \\
&= -w * \frac{(||\lambda x + (1-\lambda)y||)x}{||x||} \neq 0
\end{aligned}$$

Hence we have proved that no two sample points in the $r+1$ sample can give a line which passes through the origin. What is significant about this can seen later on.

We define a mapping:

$$\phi : (x_1, ..., x_r), x_r > 0 \mapsto (\frac{x_1}{x_r}, ..., \frac{x_{r-1}}{x_r})$$

$$\begin{aligned}
&|\{(c(\alpha_1), ..., c(\alpha_{r+1})), c \in \{sgn(w * \alpha), w \in \mathbb{R}^r\}\}| = 2^{r+1} \\
\rightarrow &|\{(d(\phi(\alpha_1)), ..., d(\phi(\alpha_{r+1}))), d \in \{sgn(w * (\phi(\alpha), 1)), w \in \mathbb{R}^r\}\}| = 2^{r+1}
\end{aligned}$$

That is, there are $r+1$ sample points in $\mathbb{R}^{r-1}$, i.e, $\phi(\alpha_1), ..., \phi(\alpha_{r+1})$ such that they can be completely classified by hyperplanes in $\mathbb{R}^{r-1}$. Thus a contradiction to the fact that $VCdim(H^{r-1}) = r$. Hence we have proved that $VCdim(C) = r$.

$$\begin{aligned}
VCdim(H) &< m = 2adlog_2(8a) \\
&< 2arlog_2(8a)
\end{aligned}$$

# 3 Problem C

## 3.1 C.1

## 3.2 C.2,3

## 3.3 C.4

See abalone.data.processing.FML.HW2.ipynb

## 3.4 C.5

$C^* = 512, d^* = 9$. See abalone.data.processing.FML.HW2.ipynb

## 3.5 C.6

### 3.5.1

Consider the following map:

$$\hat{x} = (y_1 K(x, x_1), y_2 K(x, x_2), ..., y_m K(x, x_m))$$

Which takes a sample $x \in \mathbb{R}^m$ to another point in the same Euclidean space. Then problem (1) can be rewritten as:

$$\begin{aligned}
&min_{\alpha,b,\xi} \frac{1}{2}||\alpha||_2^2 + C \sum_{1 \le i \le m} \xi_i \\
&y_i(\alpha * \hat{x_i} + b) \ge 1 - \xi_i, 1 \le i \le m \\
&\xi_i, \alpha_i \ge 0
\end{aligned}$$

That is, by thinking of $\hat{x}$ as a result of transformation of the original sample point $x$ under the kernel function, we reformulated the problem as the primal optimization problem of SVMs for the transformed vector $\hat{x}$.

### 3.5.2

No, the positive-definiteness is not necessary, since the target function is convex, and the domain defined by the constrains are a convex region, because the domain function is linear with respect to $\alpha, \xi, b$, and hence convex. To see this:

$$\forall(\alpha, b, \xi), (\tilde{a}, \tilde{b}, \tilde{\xi}), \forall\lambda \in [0,1]$$
$$y_i(\sum_{1\leq j\leq m} \alpha_j y_j K(x_i, x_j) + b) + \xi_i - 1 \geq 0$$
$$y_i(\sum_{1\leq j\leq m} \tilde{\alpha_j} y_j K(x_i, x_j) + \tilde{b}) + \tilde{\xi_i} - 1 \geq 0$$
$$\rightarrow y_i(\sum_{1\leq j\leq m} (\lambda\alpha_j + (1-\lambda)\tilde{\alpha_j})y_j K(x_i, x_j) + (\lambda b + (1-\lambda)\tilde{b})) + (\lambda\xi + (1-\lambda)\tilde{\xi_i}) - 1$$
$$= \lambda(y_i(\sum_{1\leq j\leq m} \alpha_j y_j K(x_i, x_j) + b) + \xi_i - 1 \geq 0) + (1-\lambda)(y_i(\sum_{1\leq j\leq m} \tilde{\alpha_j} y_j K(x_i, x_j) + \tilde{b}) + \tilde{\xi_i} - 1)$$
$$\geq 0$$

Since the intersection of convex sets are convex, and that $\xi_i, \alpha_i \geq 0$ are convex, our domain of optimization problem is convex.

#### 3.5.3

First construct the Lagrange multiplier function:

$$L = (\frac{1}{2}||\alpha||_2^2 + C\sum_{1\leq i\leq m}\xi_i) - \sum_{1\leq i\leq m}\lambda_i(y_i(\alpha * \hat{x_i} + b) - 1 + \xi_i) - \mu * \alpha - \sigma * \xi$$
$$\lambda_i, \mu_i, \sigma_i \geq 0, \forall i$$

Differentiating the function with respect to $\alpha, b, \xi$, we get:

$$\nabla_\alpha L = \alpha - \sum_{1\leq i\leq m}\lambda_{1i}\hat{x_i} - \mu = 0$$
$$\nabla_b L = -\sum_{1\leq i\leq m}\lambda_i y_i = 0$$
$$\nabla_{\xi_i} L = C - \lambda_i - \sigma_i = 0$$

which then gives the following:

$$\alpha = \sum_{1\leq i\leq m}\lambda_{1i}\hat{x_i} + \mu$$
$$\sum_{1\leq i\leq m}\lambda_i y_i = 0$$
$$C = \lambda_i + \sigma_i$$

Plug them back in the Lagrange function, we get:

$$\begin{aligned}
L = \frac{1}{2}|| \sum_{1\le i\le m} \lambda_{1i}\hat{x_i} + \mu|| + C\sum_{1\le i\le m}\xi_i \\
- \sum_{1\le i\le m}\lambda_i[y_i(\mu * \hat{x_i} + \sum_{1\le j\le m}\lambda_j y_j \hat{x_j} * \hat{x_i} + b) - 1 + \xi_i] \\
- (\mu * \mu + \mu * (\sum_{1\le j\le m}\lambda_j y_j \hat{x_j})) - \sigma * \xi \\
= (\frac{1}{2}(||\mu||^2 + ||\sum_{1\le i\le m}\lambda_{1i}\hat{x_i}||^2) + \sum_{1\le i\le m}\lambda_i y_i(\mu * \hat{x_i})) \\
+ C\sum_{1\le i\le m}\xi_i - \sum_{1\le i\le m}(\lambda_i + \sigma_i)\xi_i + \sum_{1\le i\le m}\lambda_i \\
- (2\sum_{1\le i\le m}\lambda_i y_i(\mu * \hat{x_i}) + \sum_{1\le i,j\le m}\lambda_i\lambda_j y_i y_j(\hat{x_i} * \hat{x_j}) + ||\mu||^2) \\
= -\frac{1}{2}||\mu + \sum_{1\le i\le m}\lambda_i y_i \hat{x_i}||^2 + \sum_{1\le i\le m}\lambda_i
\end{aligned}$$

Hence the dual problem is:

$$\begin{aligned}
max_{\mu,\lambda\in\mathbb{R}^m} - \frac{1}{2}||\mu + \sum_{1\le i\le m}\lambda_i y_i \hat{x_i}||^2 + \sum_{1\le i\le m}\lambda_i \\
\mu_i \ge 0, 0 \le \lambda_i \le C, \forall i, \sum_{1\le j\le m}\lambda_j y_j = 0
\end{aligned}$$