Sol to *Foundations of Machine Learning, Ed2*

fomiuna

(Notations are consistent with those in the book.)

Chapter 2 The PAC Learning Framework

2.1 (Two-oracle variant of the PAC model.)
Assume that $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the standard PAC model. Notice that

$$\begin{aligned}
&\mathbb{P}_{x\sim\mathcal{D}}[h_S(x) \neq c(x)] \\
&= \mathbb{P}_{x\sim\mathcal{D}}\Big[h_S(x) = 0 \mid c(x) = 1\Big]\mathbb{P}_{x\sim\mathcal{D}}[c(x) = 1] + \mathbb{P}_{x\sim\mathcal{D}}\Big[h_S(x) = 1 \mid c(x) = 0\Big]\mathbb{P}_{x\sim\mathcal{D}}[c(x) = 0] \\
&= \mathbb{P}_{x\sim\mathcal{D}_+}[h_S(x) = 0]v_+ + \mathbb{P}_{x\sim\mathcal{D}_-}[h_S(x) = 1]v_-
\end{aligned}$$

where v_+, v_- are measure of sets of the two distributions, respectively. So from the following

$$\mathbb{P}_{S\sim\mathcal{D}^m}[R(h_S) \leq \min\{v_+, v_-\} \cdot \epsilon] \geq 1 - \delta, \quad \forall c \in \mathcal{C}, m \geq \mathrm{poly}(\epsilon^{-1}, \delta^{-1}, \mathrm{size}(c), n)$$

we could deduce that the $\mathcal{C}$ satisfies the definition of two-oracle PAC model. For the other side of the proof, considering any $c \in \mathcal{C}$, $\epsilon > 0$ and $\delta > 0$, there exist m_- and m_+ polynomial in ϵ^{-1}, δ^{-1}, size(c) and n, s.t. if we draw m_- negative examples or more and m_+ positive examples or more, with confidence $1 - \delta$, the hypothesis h_S output by $\mathcal{A}$ satisfies

$$\mathbb{P}_{x\sim\mathcal{D}_+}[h_S(x) = 0] \leq \epsilon, \quad \mathbb{P}_{x\sim\mathcal{D}_-}[h_S(x) = 1] \leq \epsilon$$

Now use the first equation again, we could get the standard PAC-learnable result. The last thing we need to do is to make sure that m_- negative examples and m_+ positive examples can be achieved with high probability, which could be proved by applying Chernoff bound when $v_-, v_+ > 0$ (h_0, h_1 are for the case where there is a zero value in v_- and v_+).

2.2 (PAC learning of hyper-rectangles.)
Just the same as two-dimension case: the algorithm selects the tightest hyper-rectangle that includes all positive samples. Then we consider $2n$ hyper-rectangle areas whose probability mass is at least $\epsilon/(2n)$.

2.3 (Concentric circles.)
W.L.O.G. we could assume that $\mathbb{P}_{x\sim\mathcal{D}}[x \in B_c] > \epsilon$, where $B_c = \{x \in \mathcal{X} : \|x\|_2 \leq c\}$ is the target concept (positive examples) and our algorithm simply returns the tightest disc which include all the positive examples with given labeled data S, denoted by B_S. Now let

$$h = \sup\big\{r \in (0, c) \mid \mathbb{P}_{x\sim\mathcal{D}}[x \in B_c \backslash B_r] > \epsilon\big\}$$

we could calculate that

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}[R(h_S) > \epsilon] &\leq \mathbb{P}_{S\sim\mathcal{D}^m}[B_S \cap (B_c \backslash B_h) = \emptyset] \\
&\leq \Big(\mathbb{P}_{x\sim\mathcal{D}}[x \in B_h]\Big)^m \leq (1 - \epsilon)^m \leq \mathrm{e}^{-\epsilon m}
\end{aligned}$$

Then solve m in $\mathrm{e}^{-\epsilon m} \leq \delta$ we could get the result.

2.4 (Non-concentric circles.)
Figure 2.5(b) has provided a counterexample for Gertrude's approach.

2.5 (Triangles.)
Just the same as rectangle case.

2.6 (Learning in the presence of noise — rectangles.)
The probability that R' misses region r_j could be bounded by

$$\begin{aligned}&\mathbb{P}_{x\sim\mathcal{D}}[x \notin r_j \vee (x \in r_j \wedge \text{label of } x \text{ flipped})]\\ =\;&\mathbb{P}_{x\sim\mathcal{D}}[x \notin r_j] + \eta\mathbb{P}_{x\sim\mathcal{D}}[x \in r_j]\\ =\;&1-(1-\eta)\mathbb{P}_{x\sim\mathcal{D}}[x\in r_j] \le 1-\epsilon(1-\eta)/4 \le 1-\epsilon(1-\eta')/4\end{aligned}$$

Applying union bound and the i.i.d. r.v. we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}[R(R')>\epsilon] \le 4\Big(1-\epsilon(1-\eta')/4\Big)^m \le 4\mathrm{e}^{-m\epsilon(1-\eta')/4}$$

By setting the RHS as δ we could deduce the PAC-learnable result.

2.7 (Learning in the presence of noise — general case.)
The label of a point disagrees with the one given by h is either because its label is correct and h misclassifies it, or because its label is incorrect and h classifies it correctly. Since the change of label is independent with h, h^*, we have

$$\begin{aligned}d(h) &= \mathbb{P}_{x\sim\mathcal{D}}[h(x)\neq \text{label}(x)]\\ &= \mathbb{P}_{x\sim\mathcal{D}}\Big[h(x)=h^*(x), h^*(x)\neq\text{label}(x)\Big] + \mathbb{P}_{x\sim\mathcal{D}}\Big[h(x)\neq h^*(x), h^*(x)=\text{label}(x)\Big]\\ &= (1-R(h))\eta + R(h)(1-\eta) = \eta + (1-2\eta)R(h)\end{aligned}$$

From the equation above it's obvious that $d(h^*)=\eta$. To show PAC-learning for algorithm L, that is

$$\begin{aligned}&\mathbb{P}_{S\sim\mathcal{D}^m}[R(h_S)\le\epsilon] \ge 1-\delta, \quad \forall m > \text{poly}(\epsilon^{-1},\delta^{-1},n)\\ \Leftarrow\;&\mathbb{P}_{S\sim\mathcal{D}^m}[d(h_S)-d(h^*)\le\epsilon'] \ge 1-\delta, \quad \text{where } \epsilon'=(1-2\eta')\epsilon,\ m \text{ sufficiently large}\end{aligned}$$

Now assume $\mathcal{H}_{\epsilon'} = \{h\in\mathcal{H} : d(h)-d(h^*)>\epsilon'\}$. Notice that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\forall h\in\mathcal{H}_{\epsilon'} : \widehat{d}(h)\neq\widehat{d}(h_S)\Big] \ge \mathbb{P}_{S\sim\mathcal{D}^m}\Bigg[\bigwedge_{h\in\mathcal{H}_{\epsilon'}} \widehat{d}(h)\ge\widehat{d}(h^*)\Bigg]$$

So if we could lower bound the RHS with a sufficiently large value, then we just finish our proof, which means to prove that, for any h, if $R(h)>\epsilon$, then with high probability $\widehat{d}(h)\ge\widehat{d}(h^*)$.

- Step 1: estimate the gap between $d(h^*)$ and $\widehat{d}(h^*)$ (single hypothesis).
  Since $\widehat{d}(h^*)$ is the sum of m i.i.d. r.v. and $\mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{d}(h^*)] = d(h^*)$, by Hoeffding's inequality we have
  $$\mathbb{P}_{S\sim\mathcal{D}^m}[\widehat{d}(h^*)-d(h^*)>\epsilon'/2] \le \mathrm{e}^{-m\epsilon'^2/2}$$
  Setting $\delta/2$ to the RHS yields
  $$\mathbb{P}_{S\sim\mathcal{D}^m}[\widehat{d}(h^*)-d(h^*)\le\epsilon'/2] \ge 1-\delta/2, \quad \forall m\ge\frac{2}{\epsilon'^2}\log\frac{2}{\delta} \tag{1}$$

- Step 2: estimate the gap between $d(h)$ and $\widehat{d}(h)$ (finite $\mathcal{H}$, inconsistent case).
  Consider
  $$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\exists h\in\mathcal{H} : d(h)-\widehat{d}(h)>\epsilon'/2\Big] &= \mathbb{P}_{S\sim\mathcal{D}^m}\Bigg[\bigvee_{h\in\mathcal{H}} d(h)-\widehat{d}(h)>\epsilon'/2\Bigg]\\ &\le |\mathcal{H}|\mathrm{e}^{-m\epsilon'^2/2}\end{aligned}$$

Setting $\delta/2$ to the RHS yields

$$\mathbb{P}_{S\sim\mathcal{D}^m}[d(h) - \widehat{d}(h) \leq \epsilon'/2,\ \forall h \in \mathcal{H}] \geq 1 - \delta/2, \quad \forall m \geq \frac{2}{\epsilon'^2}(\log|\mathcal{H}| + \log\frac{2}{\delta}) \tag{2}$$

- Step 3: Notice that we have decomposition

$$\widehat{d}(h) - \widehat{d}(h^*) = \big(\widehat{d}(h) - d(h)\big) + \big(d(h) - d(h^*)\big) + \big(d(h^*) - \widehat{d}(h^*)\big)$$

then combine Eq. (1) and (2) we could get $\forall m \geq \frac{2}{\epsilon^2(1-2\eta')^2}(\log|\mathcal{H}| + \log\frac{2}{\delta})$,

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^m}[\widehat{d}(h) \geq \widehat{d}(h^*),\ \forall h \in \mathcal{H}_{\epsilon'}] \\
\geq\ &\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\big(\widehat{d}(h) - d(h) \geq -\epsilon'\big) \wedge \big(d(h) - d(h^*) > \epsilon'\big) \wedge \big(d(h^*) - \widehat{d}(h^*) \geq -\epsilon'/2\big),\ \forall h \in \mathcal{H}_{\epsilon'}\Big] \\
\geq\ &\mathbb{P}_{S\sim\mathcal{D}^m}[\widehat{d}(h) - d(h) \geq -\epsilon',\ \forall h \in \mathcal{H}] + \mathbb{P}_{S\sim\mathcal{D}^m}[d(h^*) - \widehat{d}(h^*) \geq -\epsilon'/2] - 1 \\
\geq\ &(1-\delta/2) + (1-\delta/2) - 1 = 1 - \delta
\end{aligned}$$

which completes our proof.

Chapter 3 Rademacher Complexity and VC-Dimension

3.1 (Growth function of intervals in $\mathbb{R}$.)
Consider adding a new point x_{m+1} to the existed m points $x_1 < \cdots < x_m$ such that $x_{m+1} > x_m$. It's easy to check that x_{m+1} will bring one more classification for each dichotomy in

$$\{h \in \mathcal{H} : h \cap (x_m, +\infty) \neq \emptyset\}$$

which means $\Pi_{\mathcal{H}}(m+1) - \Pi_{\mathcal{H}}(m) = m + 1$. So we have $\Pi_{\mathcal{H}} = \frac{m(m+1)}{2} + 1$.

3.2 (Growth function and Rademacher complexity of thresholds in $\mathbb{R}$.)
Notice that

$$\mathcal{H} = \{(-\infty, a] : a \in \mathbb{R}\} \cup \{[a, +\infty) : a \in \mathbb{R}\}$$

So we have $\Pi_{\mathcal{H}}(m+1) - \Pi_{\mathcal{H}}(m) = 2$ and $\Pi_{\mathcal{H}}(m) = 2m$. By Massart's lemma

$$\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_S\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\right]\right] \leq \mathbb{E}_S\left[\frac{\sqrt{m}\sqrt{2\log\Pi_{\mathcal{H}}(m)}}{m}\right] = \sqrt{\frac{2\log(2m)}{m}}$$

3.3 (Growth function of linear combinations.)

(a) $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$ and $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}\}$ are linear separable by a hyperplane going through the origin iff $\exists \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ s.t.

$$\begin{aligned}
&\mathbf{w}_1 \cdot \mathbf{x} > 0, \forall \mathbf{x} \in X^+ \cup \{\mathbf{x}_{m+1}\}, \quad \mathbf{w}_1 \cdot \mathbf{x} < 0, \forall \mathbf{x} \in X^- \\
&\mathbf{w}_2 \cdot \mathbf{x} > 0, \forall \mathbf{x} \in X^+, \quad \mathbf{w}_2 \cdot \mathbf{x} < 0, \forall \mathbf{x} \in X^- \cup \{\mathbf{x}_{m+1}\}
\end{aligned}$$

Now consider mapping

$$f : t \mapsto (t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \cdot \mathbf{x}_{m+1}, \quad t \in [0,1]$$

Since f is continuous and $f(0) < 0, f(1) > 0$, there exists some $t_0 \in (0,1)$ s.t. $f(t_0) = 0$. So $\{X^+, X^-\}$ is separable by $t_0\mathbf{w}_1 + (1-t_0)\mathbf{w}_2$ which go through the origin and $\mathbf{x}_{m+1}$. The other side of proof is just an analogy.

(b) Let hyperplane $P_i := \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x}_i = 0\}$ for $\forall i \in [m]$. Then we have

$$\begin{aligned} C(m,d) &= |\{(\text{sgn}(\mathbf{w}\cdot\mathbf{x}_1), \dots, \text{sgn}(\mathbf{w}\cdot\mathbf{x}_m)) : \mathbf{w} \in \mathbb{R}^n\}| \\ &= \#\{\text{connected components of } \mathbb{R}^d \setminus \cup_{i=1}^m P_i\} \end{aligned}$$

Now consider $(m+1)$-th point x_{m+1}. Notice that P_{m+1} splits some connected components above into two parts. The increment is

$$\#\{\text{connected components of } P_{m+1} \setminus \cup_{i=1}^m P_i\} = C(m, d-1)$$

So we have equation $C(m+1,d) = C(m,d) + C(m,d-1)$. Applying $C(1,d) = 2$, by induction we could get

$$C(m,d) = 2\sum_{k=0}^{d-1}\binom{m-1}{k}$$

(c) A direction application of (b) yields the result.

3.4 (Lower bound on growth function.)
Consider set $[m]$ and its associated hypothesis class

$$\mathcal{H} = \{x \mapsto \mathbf{1}_S(x), \forall x \in [m] : S \subseteq [m], |S| \le d\}$$

It's obvious to see that $\text{VCdim}(\mathcal{H}) = d$ and $\Pi_{\mathcal{H}}(m) = \sum_{i=0}^d \binom{m}{i}$.

3.5 (Finer Rademacher upper bound.)
By applying Jensen's inequality we could get

$$\mathfrak{R}_m(\mathcal{G}) \le \mathbb{E}_S\left[\sqrt{\frac{2\log \Pi(\mathcal{G},S)}{m}}\right] \le \sqrt{\frac{2\log \mathbb{E}_S[\Pi(\mathcal{G},S)]}{m}}$$

(rmk: Is this really a "finer" bound?)

3.6 (Singleton hypothesis class.)
For (a) just check the definition. For (b), it's easy to check that both sides of Massart's inequality are zero for any single-element hypothesis set.

3.7 (Two function hypothesis class.)

(a) Easy to verify that the VC-dimension $d = 1$ for hypothesis $\mathcal{H}$, and

$$\begin{aligned} \widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\{h_{-1},h_{+1}\}}\sum_{i=1}^m \sigma_i h(x_i)\right] \\ &= \frac{1}{m2^{m-1}}\sum_{i=0}^{\lceil m/2-1\rceil}\binom{m}{i}(m-2i) \le 1 - \frac{4m}{2^m}, \quad \forall m \ge 5 \end{aligned}$$

(b) The VC-dimension $d = 1$, and $\widehat{\mathfrak{R}}_S(\mathcal{H}) = 1/m$ since

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\{h_{-1},h_{+1}\}}\sum_{i=1}^m \sigma_i h(x_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\{h_{-1},h_{+1}\}}\sigma_1 h(x_1)\right] = 1$$

3.8 (Rademacher identities.)

(a) If $\alpha \geq 0$

$$\sup_{h\in\alpha\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i) = \alpha \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)$$

otherwise if $\alpha < 0$, then

$$\sup_{h\in\alpha\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i) = (-\alpha) \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} (-\sigma_i) h(x_i)$$

Notice that $\sigma_i, -\sigma_i$ are same distribution, so we have $\mathfrak{R}_m(\alpha\mathcal{H}) = |\alpha|\mathfrak{R}_m(\mathcal{H})$.

(b) Notice that

$$\begin{aligned}
\mathfrak{R}_m(\mathcal{H}+\mathcal{H}') &= \frac{1}{m}\mathbb{E}_{S,\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \sum_{i=1}^{m} \sigma_i(h(x_i)+h'(x_i))\Bigg] \\
&= \frac{1}{m}\mathbb{E}_{S,\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \sum_{i=1}^{m} \sigma_i h(x_i) + \sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \sum_{i=1}^{m} \sigma_i h'(x_i)\Bigg] \\
&= \frac{1}{m}\mathbb{E}_{S,\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)\Bigg] + \frac{1}{m}\mathbb{E}_{S,\boldsymbol{\sigma}}\Bigg[\sup_{h'\in\mathcal{H}'} \sum_{i=1}^{m} \sigma_i h'(x_i)\Bigg] = \mathfrak{R}_m(\mathcal{H}) + \mathfrak{R}_m(\mathcal{H}')
\end{aligned}$$

(c) Notice that

$$\begin{aligned}
&\mathfrak{R}_m(\{\max(h,h') : h\in\mathcal{H}, h'\in\mathcal{H}'\}) \\
=& \frac{1}{2m}\mathbb{E}_{S,\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \sum_{i=1}^{m} \sigma_i\big(h(x_i)+h'(x_i)+|h(x_i)-h'(x_i)|\big)\Bigg] \\
\leq& \frac{1}{2}\big(\mathfrak{R}_m(\mathcal{H})+\mathfrak{R}_m(\mathcal{H}')\big) + \frac{1}{2m}\mathbb{E}_{S,\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \sum_{i=1}^{m} \sigma_i |h(x_i)-h'(x_i)|\Bigg]
\end{aligned} \tag{3}$$

By definition of supremum, for any $\epsilon > 0$, there exists $h_1, h_2, \in \mathcal{H}$ and $h_1', h_2', \in \mathcal{H}'$ s.t.

$$\begin{aligned}
u_{m-1}(h_1,h_1') + |h_1(x_m)-h_1'(x_m)| &\geq (1-\epsilon)\Big(\sup_{h\in\mathcal{H},h\in\mathcal{H}'} u_{m-1}(h,h') + |h(x_m)-h'(x_m)|\Big) \\
u_{m-1}(h_2,h_2') - |h_2(x_m)-h_2'(x_m)| &\geq (1-\epsilon)\Big(\sup_{h\in\mathcal{H},h\in\mathcal{H}'} u_{m-1}(h,h') - |h(x_m)-h'(x_m)|\Big)
\end{aligned}$$

where $u_{m-1}(h,h') = \sum_{i=1}^{m-1} \sigma_i |h(x_i)-h'(x_i)|$. Thus, we have

$$\begin{aligned}
&(1-\epsilon)\mathbb{E}_{\sigma_m}\Big[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} u_{m-1}(h,h') + \sigma_m|h(x_m)-h'(x_m)|\Big] \\
\leq& \frac{1}{2}\big(u_{m-1}(h_1,h_1') + |h_1(x_m)-h_1'(x_m)|\big) + \frac{1}{2}\big(u_{m-1}(h_2,h_2') - |h_2(x_m)-h_2'(x_m)|\big) \\
\leq& \frac{1}{2}\Big(u_{m-1}(h_1,h_1') + u_{m-1}(h_2,h_2') + s\big(h_1(x_m)-h_1'(x_m)-(h_2(x_m)-h_2'(x_m))\big)\Big) \\
\leq& \frac{1}{2}\big(u_{m-1}(h_1,h_1') + s(h_1(x_m)-h_1'(x_m))\big) + \frac{1}{2}\big(u_{m-1}(h_2,h_2') - s(h_2(x_m)-h_2'(x_m))\big) \\
\leq& \frac{1}{2}\sup_{h\in\mathcal{H},h'\in\mathcal{H}'}\big(u_{m-1}(h,h') + s(h(x_m)-h'(x_m))\big) + \frac{1}{2}\sup_{h\in\mathcal{H},h'\in\mathcal{H}'}\big(u_{m-1}(h,h') - s(h(x_m)-h'(x_m))\big) \\
=& \mathbb{E}_{\sigma_m}\Big[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} u_{m-1}(h,h') + \sigma_m(h(x_m)-h'(x_m))\Big]
\end{aligned}$$

where $s = \operatorname{sgn}(h_1(x_m) - h'_1(x_m) - (h_2(x_m) - h'_2(x_m)))$. Due to the arbitrariness of ϵ and by induction, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'}\sum_{i=1}^{m}\sigma_i|h(x_i)-h'(x_i)|\right] &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'}\sum_{i=1}^{m}\sigma_i(h(x_i)-h'(x_i))\right] \\ &= m\big(\widehat{\mathfrak{R}}_S(\mathcal{H})+\widehat{\mathfrak{R}}_S(\mathcal{H}')\big)\end{aligned} \tag{4}$$

Combine Eq. (3) and (4) we have

$$\mathfrak{R}_m(\{\max(h,h') : h\in\mathcal{H}, h'\in\mathcal{H}'\}) \leq \mathfrak{R}_m(\mathcal{H}) + \mathfrak{R}_m(\mathcal{H}')$$

3.27 (VC-dimension of neural networks.)

(a) Let $\Pi_u(m)$ denote the growth function at a node u in the intermediate layer. For a fixed set of values at the intermediate layer, using the concept class $\mathcal{C}$ the output node can generate at most $\Pi_{\mathcal{C}}(m)$ distinct labelings. There are $\prod_u \Pi_u(m)$ possible sets of values at the intermediate layer since, by definition, for a sample of size m, at most $\Pi_u(m)$ distinct values are possible at each u. Thus we have

$$\Pi_{\mathcal{H}}(m) \leq \Pi_{\mathcal{C}}(m)\prod_u \Pi_u(m)$$

(b) For any intermediate node u, $\Pi_u(m) = \Pi_{\mathcal{C}}(m)$. Thus $\Pi_{\mathcal{H}}(m) \leq \Pi_{\mathcal{C}}^k(m)$. By Sauer's lemma, $\Pi_{\mathcal{H}}(m) \leq (\frac{\mathrm{e}m}{d})^{dk}$. Let $m = 2kd\log_2(\mathrm{e}k)$. In view of the inequality given by the hint and $\mathrm{e}k > 4$, this implies $m > dk\log_2(\frac{\mathrm{e}m}{d})$, that is $2^m > (\frac{\mathrm{e}m}{d})^{dk}$. Thus,

$$\mathrm{VCdim}(\mathcal{H}) \leq 2kd\log_2(\mathrm{e}k)$$

(c) For threshold functions, the VC-dimension of $\mathcal{C}$ is r, thus, the VC-dimension of $\mathcal{H}$ is upper bounded by $2kr\log_2(\mathrm{e}k)$.

3.28 (VC-dimension of convex combinations.)
Following the hint, we can think of this family of functions as a one hidden layer neural network, where the hidden layer is represented by the functions $h_t \in \mathcal{H}$, and the top layer is a threshold function characterized by $(\alpha_1,\ldots,\alpha_T)$. Denote this class of threshold functions by Δ_T. By problem 3.27 we could bound

$$\Pi_{\mathcal{F}_T}(m) \leq \Pi_{\Delta_T}(m)\Pi_{\mathcal{H}}^T(m)$$

Since the VC-dimension of Δ_T is at most T, and we may further denote the VC-dimension of $\mathcal{H}$ by d. Applying Sauer's lemma to the growth function yields

$$\Pi_{\Delta_T}(m) \leq \left(\frac{\mathrm{e}m}{T}\right)^T,\ \Pi_{\mathcal{H}}(m) \leq \left(\frac{\mathrm{e}m}{d}\right)^d \quad\Rightarrow\quad \Pi_{\mathcal{F}_T}(m) \leq \left(\frac{\mathrm{e}m}{T}\right)^T\left(\frac{\mathrm{e}m}{d}\right)^{Td}$$

Let $m = \max\{4T\log_2(2\mathrm{e}), 2Td\log_2(\mathrm{e}T)\}$, we have

$$\left(\frac{\mathrm{e}m}{T}\right)^T\left(\frac{\mathrm{e}m}{d}\right)^{Td} < 2^{4T\log_2(2\mathrm{e})+2Td\log_2(\mathrm{e}T)}$$

so that the VC-dimension of $\mathcal{F}_T$ is bounded by $2T(2\log_2(2\mathrm{e}) + d\log_2(\mathrm{e}T))$.

Chapter 4 Model Selection

4.1 For any hypothesis set $\mathcal{H}$, show that the following inequality holds:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\widehat{R}_S(h_S^{\mathrm{ERM}})\right] \leq \inf_{h\in\mathcal{H}} R(h) \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[R(h_S^{\mathrm{ERM}})\right]$$

Chapter 5 Support Vector Machines

5.1 (Soft margin hyperplanes.)

(a) Let $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}_+^m$ and the Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m \xi_i^p + \sum_{i=1}^m \alpha_i\big(1 - \xi_i - y_i(\mathbf{w}\cdot\mathbf{x}_i + b)\big) + \sum_{i=1}^m \beta_i(-\xi_i)$$

then the primal problem is

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Consider

$$\begin{cases} \nabla_{\mathbf{w}}\mathcal{L} & = 0 \\ \partial\mathcal{L}/\partial b & = 0 \\ \partial\mathcal{L}/\partial \xi_i & = 0,\ \forall i \in [m] \end{cases} \quad \Rightarrow \quad \begin{cases} \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i & = 0 \\ \sum_{i=1}^m \alpha_i y_i & = 0 \\ Cp\xi_i^{p-1} - \alpha_i - \beta_i & = 0,\ \forall i \in [m] \end{cases}$$

So we could write $\mathcal{L}$ as

$$\begin{aligned} \mathcal{L} &= \frac{1}{2}\Big\|\sum_{i=1}^m \alpha_i y_i \mathbf{x}_i\Big\|^2 + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i y_i\Big(\sum_{j=1}^m \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x}_i\Big) + \sum_{i=1}^m \Big(C\xi_i^p - (\alpha_i + \beta_i)\xi_i\Big) \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{1\leq i,j\leq m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j - C'\sum_{i=1}^m \big(\alpha_i + \beta_i\big)^{\frac{p}{p-1}} \end{aligned}$$

where $C' = (p-1)\Big/(Cp^p)^{\frac{1}{p-1}}$. So the Lagrange duality is

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{1\leq i,j\leq m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j - C'\sum_{i=1}^m \big(\alpha_i + \beta_i\big)^{\frac{p}{p-1}} \tag{5}$$

(b) When $p = 1$, the last term will take zero value. When $p = 2$, Eq. (5) becomes a quadratic programming problem, which is convex.

5.2 (Tighter Rademacher bound.)
Consider two sequences $\{\rho_k\}_{k=1}^\infty \subseteq (0, +\infty)$ and $\{\epsilon_k\}_{k=1}^\infty \subseteq (0, 1)$. For any fixed $k \geq 1$, we have

$$\mathbb{P}\Big[\sup_{h\in\mathcal{H}} R(h) - \widehat{R}_{S,\rho_k}(h) - \frac{2}{\rho_k}\mathfrak{R}_m(\mathcal{H}) > \epsilon_k\Big] \leq \mathrm{e}^{-2m\epsilon_k^2}$$

Now choose $\epsilon_k = \epsilon + \sqrt{\frac{\log k}{m}}$, by union bound it holds

$$\begin{aligned} &\mathbb{P}\Big[\sup_{h\in\mathcal{H},k\geq 1} R(h) - \widehat{R}_{S,\rho_k}(h) - \frac{2}{\rho_k}\mathfrak{R}_m(\mathcal{H}) - \sqrt{\frac{\log k}{m}} > \epsilon\Big] \\ \leq &\sum_{k=1}^\infty \mathrm{e}^{-2m\epsilon_k^2} = \sum_{k=1}^\infty \exp\Big\{-2m\Big(\epsilon + \sqrt{\frac{\log k}{m}}\Big)^2\Big\} \\ \leq &\sum_{k=1}^\infty \mathrm{e}^{-2m\epsilon^2}\mathrm{e}^{-2\log k} = \frac{\pi^2}{6}\mathrm{e}^{-2m\epsilon^2} < 2\mathrm{e}^{-2m\epsilon^2} \end{aligned}$$

Then choose $\rho_k = \gamma^{-k}$. For any $\rho \in (0, 1]$, let $k' = \lfloor \log_\gamma \frac{\gamma}{\rho} \rfloor$ and we have $\rho_{k'} = \gamma^{-k'} < \rho \leq \gamma^{-(k'-1)} = \gamma\rho_{k'}$. So it holds

$$\log k' \leq \log\log_\gamma \frac{\gamma}{\rho}, \quad \widehat{R}_{S,\rho_{k'}}(h) \leq \widehat{R}_{S,\rho}(h), \quad \frac{2}{\rho_{k'}} \leq \frac{2\gamma}{\rho}$$

which means

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{h\in\mathcal{H},\rho\in(0,1]} R(h) - \widehat{R}_{S,\rho}(h) - \frac{2}{\rho}\mathfrak{R}_m(\mathcal{H}) - \sqrt{\frac{\log\log_\gamma \frac{\gamma}{\rho}}{m}} > \epsilon\Big] \\
\le\ &\mathbb{P}\Big[\sup_{h\in\mathcal{H},k\ge 1} R(h) - \widehat{R}_{S,\rho_k}(h) - \frac{2}{\rho_k}\mathfrak{R}_m(\mathcal{H}) - \sqrt{\frac{\log k}{m}} > \epsilon\Big] < 2\mathrm{e}^{-2m\epsilon^2}
\end{aligned}$$

5.3 (Importance weighted SVM.)
The primal problem could be stated as

$$\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^m p_i\xi_i \\
\text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \ge 1 - \xi_i \wedge \xi_i \ge 0,\ i\in[m]
\end{aligned}$$

and the dual problem is

$$\begin{aligned}
\max_{\boldsymbol{\alpha}\ge 0} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{1\le i,j\le m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \\
\text{subject to} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \wedge 0 \le \alpha_i \le p_i,\ i\in[m]
\end{aligned}$$

5.4 (Sequential minimal optimization.)

(a) Easy to check that

$$\begin{aligned}
&\sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{1\le i,j\le m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \\
=\ &\alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2\mathbf{x}_1\cdot\mathbf{x}_1 - \frac{1}{2}\alpha_2^2\mathbf{x}_2\cdot\mathbf{x}_2 - \alpha_1\alpha_2 y_1 y_2 \mathbf{x}_1\cdot\mathbf{x}_2 \\
&\quad - \alpha_1 y_1 \sum_{i=3}^m \alpha_i y_i \mathbf{x}_i\cdot\mathbf{x}_1 - \alpha_2 y_2 \sum_{i=3}^m \alpha_i y_i \mathbf{x}_i\cdot\mathbf{x}_2 + \sum_{i=3}^m \alpha_i - \sum_{3\le i,j\le m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \\
=\ &\alpha_1 + \alpha_2 - \frac{1}{2}K_{11}\alpha_1^2 - \frac{1}{2}K_{22}\alpha_2^2 - sK_{12}\alpha_1\alpha_2 - y_1\alpha_1 v_1 - y_2\alpha_2 v_2 + \widetilde{C}
\end{aligned}$$

where

$$K_{ij} = \mathbf{x}_i\cdot\mathbf{x}_j,\ i,j\in[m], \qquad v_i = \sum_{j=3}^m \alpha_j y_j K_{ij},\ i\in[2], \qquad s = y_1 y_2,$$

$$\widetilde{C} = \sum_{i=3}^m \alpha_i - \sum_{3\le i,j\le m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \quad \text{are terms that do not depend on either } \alpha_1 \text{ or } \alpha_2$$

So the optimization problem reduces to

$$\begin{aligned}
\max_{\alpha_1,\alpha_2\ge 0} \quad & \Psi_1(\alpha_1,\alpha_2) = \alpha_1 + \alpha_2 - \frac{1}{2}K_{11}\alpha_1^2 - \frac{1}{2}K_{22}\alpha_2^2 - sK_{12}\alpha_1\alpha_2 - y_1\alpha_1 v_1 - y_2\alpha_2 v_2 \\
\text{subject to} \quad & 0 \le \alpha_1,\alpha_2 \le C \wedge \alpha_1 + s\alpha_2 = \gamma
\end{aligned}$$

where $\gamma = -y_1\sum_{i=3}^m y_i\alpha_i$.

(b) Easy to verify that

$$\Psi_2(\alpha_2) = \Big(K_{12}-\frac{1}{2}K_{11}-\frac{1}{2}K_{22}\Big)\alpha_2^2+\Big(s(K_{11}-K_{12})\gamma+y_2(v_1-v_2)-s+1\Big)\alpha_2+\Big(1-\frac{1}{2}K_{11}\gamma-y_1v_1\Big)\gamma$$

Notice that $\eta = K_{11} + K_{22} - 2K_{12} \geq 0$, Ψ_2 is minimized at

$$\alpha_2 = \frac{s(K_{11} - K_{12})\gamma + y_2(v_1 - v_2) - s + 1}{\eta}$$

without the constraints of Ψ_1.

(c) Easy to verify that

$$\begin{aligned}
v_1 - v_2 &= \sum_{j=3}^{m} \alpha_j^* y_j (K_{1j} - K_{2j}) \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) - \alpha_1^* y_1 K_{11} - \alpha_2^* y_2 K_{12} + \alpha_1^* y_1 K_{12} + \alpha_2^* y_2 K_{22} \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) + \alpha_2^* y_2 \eta - \alpha_2^* y_2 (K_{11} - K_{12}) - \alpha_1^* y_1 (K_{11} - K_{12}) \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) + \alpha_2^* y_2 \eta - s y_2 \gamma (K_{11} - K_{12})
\end{aligned}$$

(d) Now we apply (c) to (b), that is

$$\begin{aligned}
\alpha_2 &= \frac{1}{\eta}\Big(s(K_{11} - K_{12})\gamma + y_2\big(f(\mathbf{x}_1) - f(\mathbf{x}_2)\big) + \alpha_2^*\eta - s\gamma(K_{11} - K_{12}) - s + 1\Big) \\
&= \alpha_2^* + \frac{y_2}{\eta}\Big(f(\mathbf{x}_1) - f(\mathbf{x}_2) - y_1 + y_2\Big) \\
&= \alpha_2^* + y_2 \frac{(y_2 - f(\mathbf{x}_2)) - (y_1 - f(\mathbf{x}_1))}{\eta}
\end{aligned}$$

(rmk: This is the update formula of SMO, which consider two variables together in order to solve the difficulty when updating under the constraint.)

(e) The clipping is required to make sure that $\alpha_1, \alpha_2 \in [0, C]$. When $s = +1$ we have $\alpha_1 + \alpha_2 = \gamma$, so

$$\begin{aligned}
\alpha_2 \geq 0 \wedge \alpha_2 = \gamma - \alpha_1 \geq \gamma - C \;\Rightarrow\; L = \max\{0, \gamma - C\} \\
\alpha_2 \leq C \wedge \alpha_2 = \gamma - \alpha_1 \leq \gamma \;\Rightarrow\; H = \min\{C, \gamma\}
\end{aligned}$$

5.6 (Sparse SVM.)

(a) Let $\mathbf{x}_i' = \big(y_1\mathbf{x}_1 \cdot \mathbf{x}_i, \ldots, y_m\mathbf{x}_m \cdot \mathbf{x}_i\big)$, then the optimization problem becomes

$$\begin{aligned}
\min_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\sum_{i=1}^{m}\xi_i \\
\text{subject to} \quad & y_i(\boldsymbol{\alpha} \cdot \mathbf{x}_i' + b) \geq 1 - \xi_i \wedge \alpha_i, \xi_i \geq 0,\ i \in [m]
\end{aligned}$$

This is just the standard form of the primal SVM optimization problem on samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, modulo the non-negativity constraints on α_i.

(b) Let $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathbb{R}^m_{\geq 0}$ and the Lagrangian is

$$\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \mathbf{p}, \mathbf{q}, \mathbf{r}) = \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m} p_i\Big(1 - \xi_i - y_i\Big(\sum_{j=1}^{m}\alpha_j y_j \mathbf{x}_j \cdot \mathbf{x}_i + b\Big)\Big) + \sum_{i=1}^{m} q_i(-\xi_i) + \sum_{i=1}^{m} r_i(-\alpha_i)$$

Consider

$$\begin{cases} \partial\mathcal{L}/\partial\alpha_i &= 0 \\ \partial\mathcal{L}/\partial b &= 0 \\ \partial\mathcal{L}/\partial\xi_i &= 0 \end{cases} \quad\Rightarrow\quad \begin{cases} \alpha_i - y_i\mathbf{x}_i\cdot\sum_{j=1}^m p_j y_j \mathbf{x}_j - r_i &= 0 \\ \sum_{i=1}^m p_i y_i &= 0 \\ C - p_i - q_i &= 0 \end{cases}$$

Plugging in the expressions above in $\mathcal{L}$ gives

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \mathbf{p}, \mathbf{q}, \mathbf{r}) \\
=\ & \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\sum_{i=1}^m \xi_i + \sum_{i=1}^m p_i(1-\xi_i) - \sum_{i=1}^m \alpha_i(\alpha_i - r_i) - \sum_{i=1}^m p_i y_i b - \sum_{i=1}^m q_i\xi_i - \sum_{i=1}^m r_i\alpha_i \\
=\ & -\frac{1}{2}\|\boldsymbol{\alpha}\|^2 + \sum_{i=1}^m p_i
\end{aligned}$$

Due to complementary slackness, we have

$$r_i\alpha_i = 0 \ \Rightarrow\ r_i y_i \mathbf{x}_i \cdot \sum_{j=1}^m p_j y_j \mathbf{x}_j + r_i^2 = 0, \quad \forall i \in [m]$$

Now we could calculate

$$\begin{aligned}
\|\boldsymbol{\alpha}\|^2 &= \sum_{i=1}^m \left(y_i\mathbf{x}_i \cdot \sum_{j=1}^m p_j y_j \mathbf{x}_j \right)^2 - \sum_{i=1}^m r_i^2 \\
&= \sum_{i=1}^m \left(\sum_{1\le j,k\le m} p_j p_k (y_j\mathbf{x}_j \cdot y_i\mathbf{x}_i)(y_k\mathbf{x}_k \cdot y_i\mathbf{x}_i) \right) - \sum_{i=1}^m r_i^2 \\
&= \sum_{1\le j,k\le m} p_j p_k K_{jk} - \sum_{i=1}^m r_i^2
\end{aligned}$$

where $K_{jk} = \sum_{i=1}^m (y_j\mathbf{x}_j \cdot y_i\mathbf{x}_i)(y_k\mathbf{x}_k \cdot y_i\mathbf{x}_i)$. Putting everything together, the dual optimization problem is

$$\begin{aligned}
\max_{\mathbf{p},\mathbf{r}} \quad & \sum_{i=1}^m p_i - \frac{1}{2}\sum_{1\le i,j\le m} p_i p_j K_{ij} + \frac{1}{2}\sum_{i=1}^m r_i^2 \\
\text{subject to} \quad & \sum_{i=1}^m p_i y_i = 0 \wedge 0 \le p_i \le C \wedge r_i \ge 0,\ i \in [m]
\end{aligned}$$

(c) Just like the induction of (b).

Chapter 6 Kernel Methods

6.1

Let $\mathbb{H}$ be the associated RKHS and Φ is the feature mapping, we have

$$\sum_{i,j\in I} a_i a_j K'(x_i, x_j) = \sum_{i,j\in I} a_i a_j \frac{K(x_i,x_j)}{\alpha(x_i)\alpha(x_j)} = \left\| \sum_{i\in I} \frac{a_i}{\alpha(x_i)}\Phi(x_i) \right\|_{\mathbb{H}}^2 \ge 0$$

for $\forall a_i \in \mathbb{R}, x_i \in \mathcal{X}, i \in I, |I| < \infty$, which means K' is PDS.

6.2

(a) Easy to check

$$\sum_{i,j\in I} a_i a_j \cos(x_i - x_j) = \Big(\sum_{i\in I} a_i \cos x_i\Big)^2 + \Big(\sum_{i\in I} a_i \sin x_i\Big)^2 \geq 0$$

(b) Just the same as (a).

(c) Since $\cos^n x$ could be expressed by a linear combination of $1, \cos x, \ldots, \cos nx$, we could just do the same as (a) and (b).

(d) Consider any $\forall a_i \in \mathbb{R}, x_i > 0, i \in I, |I| < \infty$, define

$$f(t) = \sum_{i,j\in I} a_i a_j (x_i + x_j)^{-1} t^{x_i + x_j}, \quad t > 0$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t) = \sum_{i,j\in I} a_i a_j t^{x_i + x_j - 1} = \Big(\sum_{i\in I} a_i t^{x_i - \frac{1}{2}}\Big)^2 \geq 0$$

So $\sum_{i,j\in I} a_i a_j K(x_i, x_j) = f(1) \geq f(0) = 0$, which means K is PDS.

(e) K is PDS since

$$\sum_{i,j\in I} a_i a_j \cos\angle(\mathbf{x}_i, \mathbf{x}_j) = \Big\|\sum_{i\in I} \frac{a_i}{\|\mathbf{x}_i\|}\mathbf{x}_i\Big\|^2 \geq 0$$

(f) Notice that $\sin^2(x-x') = \frac{1}{2}\big(1-\cos(2x-2x')\big)$, so $\sin^2(x-x')$ is NDS and $K(x,x') = \exp(-\lambda \sin^2(x-x'))$ is PDS.

(g) We have

$$\|\mathbf{x}-\mathbf{y}\| = \frac{1}{2\Gamma(\frac{1}{2})}\int_0^\infty t^{-\frac{3}{2}}\Big(1 - \mathrm{e}^{-t\|\mathbf{x}-\mathbf{y}\|^2}\Big)\mathrm{d}t$$

Notice that the integrand is NDS for all $t > 0$, so $\|\mathbf{x}-\mathbf{y}\|$ is NDS and $K(\mathbf{x}, \mathbf{y}) = \mathrm{e}^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}}$ is PDS.

(h) Notice that

$$\sum_{i,j\in I} a_i a_j \min(x_i, x_j) = \sum_{i,j\in I}\int_0^1 a_i \mathbf{1}_{(0,x_i)}(t)\cdot a_j \mathbf{1}_{(0,x_j)}(t)\mathrm{d}t = \int_0^1 \Big(\sum_{i\in I} a_i \mathbf{1}_{(0,x_i)}(t)\Big)^2 \mathrm{d}t \geq 0$$

and

$$\sum_{i,j\in I} a_i a_j (1 - \max(x_i, x_j)) = \sum_{i,j\in I}\int_0^1 a_i \mathbf{1}_{(x_i,1)}(t)\cdot a_j \mathbf{1}_{(x_j,1)}(t)\mathrm{d}t \geq 0$$

and $\min(x,y) - xy = \min(x,y)(1 - \max(x,y))$, we know that $K(x,y) = \min(x,y) - xy$ is PDS.

(i) Applying binomial theorem

$$K(\mathbf{x}, \mathbf{x}') = \Big(1 - \mathbf{x}\cdot\mathbf{x}'\Big)^{-\frac{1}{2}} = \sum_{k=0}^{\infty}\binom{-\frac{1}{2}}{k}(-1)^k(\mathbf{x}\cdot\mathbf{x}')^k$$

and notice that

$$\binom{-\frac{1}{2}}{k}(-1)^k = \frac{\frac{1}{2}(\frac{1}{2}+1)\cdots(\frac{1}{2}+k-1)}{k!} > 0, \quad \forall k > 0$$

so K is PDS.

(j) The same as (g) applying

$$\frac{1}{1 + \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}} = \int_0^\infty \mathrm{e}^{-t\left(1 + \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)}\mathrm{d}t$$

(k) The same as (h).

6.4 (Symmetric difference kernel.)
Put all the elements of $\mathcal{X}$ in a ordered sequence as $x_1, x_2, \ldots, x_{|\mathcal{X}|}$, then any set $A \in 2^{\mathcal{X}}$ could be regarded as a $|\mathcal{X}|$-dimension 0-1 vector $\mathbf{x}_A$, satisfying

$$(\mathbf{x}_A)_i = \begin{cases} 0, & \text{if } x_i \notin A, \\ 1, & \text{if } x_i \in A \end{cases}$$

And we have

$$\sum_{i,j \in I} a_i a_j |S_i \cap S_j| = \sum_{i,j \in I} a_i a_j \mathbf{x}_{S_i} \cdot \mathbf{x}_{S_j} = \left\| \sum_{i \in I} a_i \mathbf{x}_{S_i} \right\|^2 \geq 0$$

So if we define $K'(A,B) = \exp(|A \cap B|)$, we know that K' is PDS. Notice that

$$K(A,B) = \exp\left(-\frac{1}{2}|A\Delta B| \right) = \frac{\exp(|A \cap B|)}{\exp(\frac{1}{2}|A|)\exp(\frac{1}{2}|B|)} = \frac{K'(A,B)}{\sqrt{K'(A,A)K'(B,B)}}$$

so K is the result of the normalization of PDS kernel K'.

6.5 (Set kernel.)
Since K_0 is PDS, let ϕ_0 be the feature mapping of K_0, we have

$$\sum_{i,j \in I} a_i a_j K'(S_i, S_j) = \sum_{i,j \in I} \left\langle a_i \sum_{x \in S_i} \phi_0(x), a_j \sum_{x \in S_j} \phi_0(x) \right\rangle_{\mathbb{H}} = \left\| \sum_{i \in I} a_i \sum_{x \in S_i} \phi_0(x) \right\|_{\mathbb{H}}^2 \geq 0$$

which means K' is also PDS.

6.6

(a) See **6.2** (f).

(b) Notice that

$$K(x,y) = \log(x+y) = \int_0^\infty \frac{\mathrm{e}^{-t} - \mathrm{e}^{-(x+y)t}}{t} \mathrm{d}t$$

and $\mathrm{e}^{-(x+y)}$ is PDS.

6.7 6.8 6.10
(Proposition). *If K is NDS, then for any $0 < \alpha \leq 1$, K^α is NDS. This could be shown by*

$$K^\alpha(x,y) = \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty t^{-\alpha-1}(1 - \mathrm{e}^{-tK(x,y)})\mathrm{d}t$$

6.12 (Explicit polynomial kernel mapping.)
$K(\mathbf{x}, \mathbf{x}')$ could be expressed as two vectors' inner product, with every entry of the $\mathbf{x}$-associated ($\mathbf{x}'$ could follow a same way) vector being the following form

$$C_{\boldsymbol{\alpha}} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_N^{\alpha_N}$$

where $\boldsymbol{\alpha}$ is the vector of exponents, $C_{\boldsymbol{\alpha}}$ is an associated constant. Notice that for the degree r, there are

$$\left|\left\{\boldsymbol{\alpha} : \sum_{i=1}^{N} \alpha_i = r\right\}\right| = \binom{N+r-1}{r}$$

solutions. So the dimension of $\mathbf{x}$ is

$$\sum_{r=0}^{d} \binom{N+r-1}{r} = \binom{N+d}{d}$$

Easy to check that the weight assigned to k_i is $\binom{d}{i} c^{d-i}$.

6.15 (Image classification kernel.)
Notice that

$$\begin{aligned}
\sum_{i,j\in I} a_i a_j K_\alpha(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j\in I} \int_{\mathbb{R}} \sum_{k=1}^{N} a_i \mathbf{1}_{(0,|x_{ik}|^\alpha)}(x) \cdot a_j \mathbf{1}_{(0,|x_{jk}|^\alpha)}(x) \mathrm{d}x \\
&= \sum_{k=1}^{N} \int_{\mathbb{R}} \left(\sum_{i\in I} a_i \mathbf{1}_{(0,|x_{ik}|^\alpha)}(x)\right)^2 \mathrm{d}x \geq 0
\end{aligned}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iN})$.

6.16 (Fraud detection.)
Notice that

$$K(U,V) = \mathbb{P}[U \wedge V] - \mathbb{P}[U]\mathbb{P}[V] = \int_\Omega \mathbf{1}_U(x)\mathbf{1}_V(x)\mathrm{d}F(x) - \int_\Omega \mathbf{1}_U(x)\mathrm{d}F(x) \int_\Omega \mathbf{1}_V(x)\mathrm{d}F(x)$$

where F is the c.d.f. of the r.v. X. So we have

$$\sum_{i,j\in I} a_i a_j K(S_i, S_j) = \int_\Omega \left(\sum_{i\in I} a_i \mathbf{1}_{S_i}(x)\right)^2 \mathrm{d}F(x) - \left(\int_\Omega \sum_{i\in I} a_i \mathbf{1}_{S_i}(x)\mathrm{d}F(x)\right)^2 = \mathbb{D}[S] \geq 0$$

where $S \sim \sum_{i\in I} a_i \mathbf{1}_{S_i}(X)$.

6.17 (Relationship between NDS and PDS kernels.)
(Schoenberg's Theorem). *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric kernel. Then K is NDS iff $\exp(-tK)$ is a PDS kernel for all $t > 0$.*
Pf. If $\exp(-tK)$ is PDS, then $-\exp(-tK)$ is NDS. Since

$$K(x,x') = \lim_{t\to 0+} \frac{1-\exp(-tK(x,x'))}{t}$$

so K is NDS. For the other part, we assume that K is NDS. Fix x_0, define

$$K'(x,x') = K(x,x_0) + K(x',x_0) - K(x,x') - K(x_0,x_0)$$

then K' is PDS. Now

$$\mathrm{e}^{-tK(x,x')} = \mathrm{e}^{tK'(x,x')}\mathrm{e}^{-tK(x,x_0)}\mathrm{e}^{-tK(x',x_0)}\mathrm{e}^{tK(x_0,x_0)}$$

Notice that

$$\sum_{i,j\in I} a_i a_j \mathrm{e}^{-tK(x_i,x_0)}\mathrm{e}^{-tK(x_j,x_0)}\mathrm{e}^{tK(x_0,x_0)} = \left(\sum_{i\in I} a_i \mathrm{e}^{-t(K(x_i,x_0)-\frac{1}{2}K(x_0,x_0))}\right)^2 \geq 0$$

So e^{-tK} is PDS. $\square$

6.18 (Metrics and Kernels.)

(a) Since K is NDS, we fix a $x_0 \in \mathcal{X}$ and define

$$K'(x,x') = \frac{1}{2}\Big(K(x,x_0) + K(x',x_0) - K(x,x') - K(x_0,x_0)\Big), \quad \forall x,x' \in \mathcal{X}$$

then K' is PDS. By theorem of RKHS, there exists a Hilbert space $\mathbb{H} \subseteq \mathbb{R}^{\mathcal{X}}$ and an associated feature mapping ϕ from $\mathcal{X}$ to $\mathbb{H}$ s.t.

$$\phi(x) = K'(x,\cdot), \quad \langle \phi(x), \phi(x')\rangle = K'(x,x'), \quad \forall x,x' \in \mathcal{X}$$

Now we calculate

$$\begin{aligned}\|\phi(x) - \phi(x')\|^2 &= K'(x,x) + K'(x',x') - 2K'(x,x') \\ &= K(x,x_0) + K(x',x_0) - (K(x,x_0) + K(x',x_0) - K(x,x')) \\ &= K(x,x')\end{aligned}$$

so $\sqrt{K}$ defines a metric on $\mathcal{X}$ since

$$\|\phi(x) - \phi(z)\| + \|\phi(z) - \phi(y)\| \geq \|\phi(x) - \phi(y)\| \quad \Rightarrow \quad \sqrt{K(x,z)} + \sqrt{K(z,y)} \geq \sqrt{K(x,y)}$$

(b) For $p > 2$, if $\exp(-|x-x'|^p)$ is PDS, then $|x-x'|^p$ is NDS, which means $|x-x'|^{p/2}$ defines a metric on $\mathbb{R}$, this is not true since the triangle inequality does not hold when $p/2 > 1$.

(c) (Remain unsolved.)

6.21 (Mercer's condition.)
(Proposition). *Let $\mathcal{X} \subseteq \mathbb{R}^N$ be a compact set and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel function satisfying*

$$\iint_{\mathcal{X}\times\mathcal{X}} c(x)c(x')K(x,x')\mathrm{d}x\mathrm{d}x' \geq 0$$

for any $c \in L^2(\mathcal{X})$, then K is PDS.
Pf. If K is not PDS, we could suppose there exist $a_i \in \mathbb{R}, x_i \in \mathcal{X}$ with index set $i \in I, |I| < \infty$, satisfying

$$\sum_{i,j\in I} a_i a_j K(x_i,x_j) = -\delta < 0$$

By continuity of K, there is an open neighborhood U_i of x_i such that

$$\sum_{i,j\in I} a_i a_j K(z_i,z_j) \leq -\delta/2$$

for all $z_i \in U_i$. Then we could approximate $\sum_{i\in I} \frac{a_i}{m(U_i)}\mathbf{1}_{U_i}$ by a continuous function c with arbitrary accuracy, where $m(U_i)$ represents the measure of set U_i. $\square$

6.22 (Anomaly detection.)

(a) Let $\boldsymbol{\alpha} \in \mathbb{R}^m_{\geq 0}$ and the Lagrangian $\mathcal{L}(\mathbf{c}, r, \boldsymbol{\alpha}) = r^2 + \sum_{i=1}^m \alpha_i(\|\phi(x_i) - \mathbf{c}\|^2 - r^2)$. Notice that

$$\begin{aligned}
\sum_{i=1}^m \alpha_i \|\phi(x_i) - \mathbf{c}\|^2 &= \sum_{i=1}^m \alpha_i \|\phi(x_i)\|^2 - 2\Big\langle \mathbf{c}, \sum_{i=1}^m \alpha_i \phi(x_i) \Big\rangle + \sum_{i=1}^m \alpha_i \|\mathbf{c}\|^2 \\
&= \sum_{i=1}^m \alpha_i \|\phi(x_i)\|^2 + \left\| \sqrt{\sum_{i=1}^m \alpha_i}\,\mathbf{c} - \frac{1}{\sqrt{\sum_{i=1}^m \alpha_i}} \sum_{i=1}^m \alpha_i \phi(x_i) \right\|^2 - \frac{1}{\sum_{i=1}^m \alpha_i} \left\| \sum_{i=1}^m \alpha_i \phi(x_i) \right\|^2 \\
&\Rightarrow \text{The min is achieved at } \mathbf{c} = \frac{1}{\sum_{i=1}^m \alpha_i} \sum_{i=1}^m \alpha_i \phi(x_i)
\end{aligned}$$

Besides,

$$\partial \mathcal{L}/\partial r = 2r + \sum_{i=1}^m \alpha_i(-2r) = 0 \quad \Rightarrow \quad \sum_{i=1}^m \alpha_i = 1$$

Applying those above we could get the dual problem

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^m \alpha_i K(x_i, x_i) - \sum_{1 \leq i,j \leq m} \alpha_i \alpha_j K(x_i, x_j) \\
\text{subject to} \quad & \boldsymbol{\alpha} \geq 0 \wedge \sum_{i=1}^m \alpha_i = 1
\end{aligned}$$

(rmk: In other words the location of this sphere only depends on points x_i with non-zero coefficients α_i. These points are analogous to the support vectors of SVM.)

(b) Since the solution $\mathbf{c}$ is the convex combination of $\phi(x_i)$, we have

$$\begin{aligned}
\|\mathbf{c}\| &\leq \sum_{i=1}^m \alpha_i \|\phi(x_i)\| \leq \sup_x \|\phi(x)\| \leq M \\
\|\phi(x_i) - \mathbf{c}\| &\leq \max_{i,j \in [m]} \|\phi(x_i) - \phi(x_j)\| \leq 2 \sup_x \|\phi(x)\| \leq 2M
\end{aligned}$$

which means the solution of (a) could be found in $\mathcal{H}$ with $\Lambda \leq M$ and $R \leq 2M$.

(c) (Remain unsolved.)

(d) The deduction is same as (a).

Chapter 7 Boosting

7.1 (VC-dimension of the hypothesis set of AdaBoost.)
See **3.28**.

7.12 (Empirical margin loss boosting.)

(a) Obviously,

$$\widehat{R}_{S,\rho}(f) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{(-\infty,0]} \left(y_i \sum_{t=1}^T \alpha_t h_t(x_i) - \rho \sum_{t=1}^T \alpha_t \right) \leq \frac{1}{m} \sum_{i=1}^m \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i) + \rho \sum_{t=1}^T \alpha_t \right)$$

(b) G_ρ is a sum of convex function and exp is differentiable.

(c) Initialize $\mathcal{D}_1$ with uniform distribution over S, and for $t \in [T]$ do

$$\begin{aligned}
h_t &= \underset{h \in \text{base classifiers}}{\operatorname{argmin}} \quad \epsilon_t := \mathbb{P}_{i \sim \mathcal{D}_t}[h(x_i) \neq y_i] \\
\alpha_t &= \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t} - \frac{1}{2} \log \frac{1+\rho}{1-\rho} \\
Z_t &= 2\sqrt{\frac{\epsilon_t(1-\epsilon_t)}{1-\rho^2}} \quad \text{(normalization factor)} \\
\mathcal{D}_{t+1}(i) &= \frac{\mathcal{D}_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}, \quad i \in [m]
\end{aligned}$$

And we return the result as $\operatorname{sgn}(\sum_{t=1}^T \alpha_t h_t)$.

(d) For the coordinate descent algorithm to make progress at each round, the step size selected along the descent direction must be non-negative, that is to say

$$\frac{1-\rho}{1+\rho} \cdot \frac{1-\epsilon_t}{\epsilon_t} > 1 \quad \Rightarrow \quad \epsilon_t < \frac{1-\rho}{2}$$

(e) See (c). Nothing to say.

(f) i. Notice that

$$\begin{aligned}
\widehat{R}_{S,\rho}(f) &\leq \frac{1}{m}\sum_{i=1}^m \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i) + \rho \sum_{t=1}^T \alpha_t\right) \\
&= \frac{1}{m}\sum_{i=1}^m \left(m \prod_{t=1}^T Z_t\right)\mathcal{D}_{T+1}(i)\exp\left(\rho\sum_{t=1}^T \alpha_t\right) \\
&= \exp\left(\rho \sum_{t=1}^T \alpha_t\right)\prod_{t=1}^T Z_t
\end{aligned}$$

ii. Let $u = \frac{1-\rho}{1+\rho}$ and recall the definition of Z_t

$$\begin{aligned}
Z_t &= \sum_{i=1}^m \mathcal{D}_t(i)\exp(-\alpha_t y_i h_t(x_i)) = \mathrm{e}^{-\alpha_t}(1-\epsilon_t) + \mathrm{e}^{\alpha_t}\epsilon_t \\
&= \sqrt{\frac{\epsilon_t(1+\rho)}{(1-\epsilon_t)(1-\rho)}}(1-\epsilon_t) + \sqrt{\frac{(1-\epsilon_t)(1-\rho)}{\epsilon_t(1+\rho)}}\epsilon_t = (u^{\frac{1}{2}} + u^{-\frac{1}{2}})\sqrt{\epsilon_t(1-\epsilon_t)}
\end{aligned}$$

so by applying i. we have

$$\widehat{R}_{S,\rho}(f) \leq \prod_{t=1}^T \mathrm{e}^{\rho\alpha_t} \prod_{t=1}^T (u^{\frac{1}{2}} + u^{-\frac{1}{2}})\sqrt{\epsilon_t(1-\epsilon_t)} = \left(u^{\frac{1+\rho}{2}} + u^{-\frac{1-\rho}{2}}\right)^T \prod_{t=1}^T \sqrt{\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}}$$

iii. By using the inequality

$$\left(u^{\frac{1+\rho}{2}} + u^{-\frac{1-\rho}{2}}\right)\sqrt{\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}} \leq 1 - 2\frac{(\frac{1-\rho}{2}-\epsilon_t)^2}{1-\rho^2}, \quad \frac{1-\rho}{2} - \epsilon_t > 0$$

we have

$$\widehat{R}_{S,\rho}(f) \leq \left(1 - 2\frac{(\frac{1-\rho}{2}-\epsilon_t)^2}{1-\rho^2}\right)^T \leq \exp\left(-\frac{2\gamma^2 T}{1-\rho^2}\right)$$

when for all $t \in [T]$, $\frac{1-\rho}{2} - \epsilon_t > \gamma > 0$. Thus, if the upper bound is less that $1/m$, then $\widehat{R}_{S,\rho}(f) = 0$ and every training point has margin at least ρ. The inequality $\exp(-\frac{2\gamma^2 T}{1-\rho^2}) < 1/m$ is equivalent to $T > \frac{(\log m)(1-\rho^2)}{2\gamma^2}$.

Chapter 8 On-Line Learning

(rmk: Here recommend a better reading material: A Modern Introduction to Online Learning, authored by Francesco Orabona.)

Chapter 9 Multi-Class Classification

Appendix

(McDiarmid's Inequality). Consider independent r.v. $X_1, \dots, X_n \in \mathcal{X}$ and a mapping $\phi : \mathcal{X}^n \to \mathbb{R}$. If for all $i \in [n]$, and for all $x_1, \dots, x_n, x_i' \in \mathcal{X}$, the function ϕ satisfies

$$|\phi(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - \phi(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i$$

then for any $t \geq 0$,

$$\mathbb{P}\Big[\phi(\mathbf{x}) - \mathbb{E}[\phi(\mathbf{x})] \geq t\Big] \leq \exp\Big(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\Big), \quad \mathbb{P}\Big[\phi(\mathbf{x}) - \mathbb{E}[\phi(\mathbf{x})] \leq -t\Big] \leq \exp\Big(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\Big)$$

(Talagrand's Inequality). Assume $X = \times_{i=1}^n X_i$ is a product space endowed with a product probability measure $\mathbb{P}$. For a subset $A \subseteq X$, the α-weighted Hamming distance between $x \in X$ and A is defined as

$$d_\alpha(x, A) = \inf_{y \in A} d_\alpha(x, y) = \inf_{y \in A} \sum_{i=1}^n \alpha_i \mathbf{1}_{x_i \neq y_i}, \quad \alpha \in \mathbb{R}_+^n$$

where $\mathbf{1}_w$ is indicator function for event w. The Talagrand's inequality states

$$\mathbb{P}[x \in A]\mathbb{P}[\rho(x, A) \geq t] \leq \exp\Big(-\frac{1}{4}t^2\Big), \quad \forall t > 0$$

where $\rho(x, A) := \sup_{\|\alpha\|_2 = 1} d_\alpha(x, A)$ is Talagrand's convex distance.

(Corallary). *Let $\Psi_1, \dots, \Psi_m$ be ℓ-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$ and $\sigma_1, \dots, \sigma_m$ be Rademacher r.v.. Then, for any hypothesis set $\mathcal{H}$ of real-valued functions, the following inequality holds*

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i(\Psi_i \circ h)(x_i)\right] \leq \ell \widehat{\mathfrak{R}}_S(\mathcal{H})$$