ML Homework 4

Lai Zehua 2014012668

2017 10 22

Problem 1. *Figure out the relationship between Φ and $\mathcal{H}$.*

Proof. Recall that $\mathcal{H}$ is the hypothesis space $\subseteq \{f(x)|f(x): X \to [0,1]\}$. $\phi_f = I[f(x) \neq y]$, $\Phi = \{\phi_f | f \in \mathcal{H}\}$. If $\mathcal{H}$ can shatter set $\{x_1, ..., x_n\}$, then $\{(f(x_1), .., f(x_n))\} = [0,1]^d$, $\{(\phi_f(x_1), .., \phi_f(x_n))\} = [0,1]^d$, so Φ shatters set $\{x_1, ..., x_n\}$. On the contrary, if Φ shatters $\{x_1, ..., x_n\}$, then $\mathcal{H}$ shatters $\{x_1, ..., x_n\}$.

By the definition of VC dimension, $VC(\mathcal{H}) = VC(\Phi)$. □

Problem 2. *Compute the VC-Dimension of Linear Classifier.*

Proof. In $\mathbb{R}^d$, consider a set S of d+1 points $O = (0, 0, ..., 0)$, $S_1 = (1, 0, ..., 0)$, $S_2 = (0, 1, ..., 0)$, $S_d = (0, 0, ..., 1)$. If $S = A \cup B$, $A \cap B = \emptyset$. WLOG, $O \notin A$, $A = \{S_{i_1}, ..., S_{i_k}\}$, then the linear classifier: $x_{i_1} + ... + x_{i_k} > \frac{1}{2}$ give a partition of A and B. So linear classifier shatters S. $VC \geq d + 1$.

For any set $S = |d + 2|$. We want to show that S cannot shattered by linear classifier. We can prove it by induction on d.

For $d = 1$, it is clear that every three points of a line can not shattered by linear classifier.

For $d > 1$, WLOG, Let $S = \{O, S_1, ..., S_{d+1}\}$. If $S_1, ..., S_{d+1}$ spans a space of dimension $< d$, then by induction hypothesis, S cannot shattered by linear classifier. Otherwise, WLOG, assume $S_1, ..., S_d$ spans a space of dimension d. We can take them as a basis of $\mathbb{R}^d$, $S_1 = (1, 0, ..., 0)$, $S_2 = (0, 1, ..., 0)$, $S_d = (0, 0, ..., 1)$, $S_{d+1} = (a_1, ..., a_d)$. If $a_1 = 0$, then $S' = \{O, S_2, ..., S_{d+1}\}$ spans a $d-1$-dimension space, by induction hypothesis, S' cannot shattered by linear classifier, so does S. If $a_1 + ... + a_d = 1$, then $S' = \{S_1, ..., S_{d+1}\}$ spans a $d-1$-dimension space, by induction hypothesis, S' cannot shattered by linear classifier, so does S. So, we can assume all $a_i \neq 0$ and $a_1 + ... + a_d \neq 1$.

Condider the subset S' of S: $S_{d+1} \notin S'$. If $a_i > 0$, $S_i \in S'$. If $a_i < 0$, $S_i \notin S'$. If $a_1 + ... + a_d > 1$, then $O \notin S'$. If $a_1 + ... + a_d < 1$, then $O \in S'$. For a linear classifier $f(x) = \sum_{i=1}^d w_i x_i + b$, if $f(S') > 0$ and $f(S \backslash S') < 0$. We have the following: If $a_i > 0$, $f(S_i) = w_i + b > 0$. If $a_i < 0$, $f(S_i) = w_i + b < 0$. If $a_1 + ... + a_d > 1$, then $f(O) = b < 0$. If $a_1 + ... + a_d < 1$, then $f(O) = b > 0$. $f(S_{d+1}) \sum_{i=1}^d w_i a_i + b = \sum_{i=1}^d (w_i + b) a_i + (1 - a_1 - ... - a_n) b < 0$, a contradiction. So, there are no linear classifier such that $f(S') > 0$ and $f(S \backslash S') < 0$. S cannot shattered by linear classifier. □

Problem 3. *Given a matrix $A = (a_{ij})_{n \times m}$, show that $\min_i \max_j a_{ij} \geq \max_j \min_i a_{ij}$.*

Proof. assume $\max_j a_{ij} = a_{ik}$, $\min_i a_{ij} = a_{lj}$. Then $a_{ik} \geq a_{ij} \geq a_{lj}$, so $\min_i \max_j a_{ij} \geq \max_j \min_i a_{ij}$. □

Problem 4. *Let p be a distribtuion over [n] then let H be a family of subsets of [n]. Suppose the correpsonding family of indicator functions $F = \{I_S : S \in H\}$ has VC-dimension d. Independently take m samples from p, denoted by $X_1, X_2, ..., X_m$.*

(1) Prove that,

$$\mathrm{E}[\sup_{S \in H} |\frac{1}{m} \sum_{i=1}^m I[X_i \in S] - S(p)|] = O(\sqrt{\frac{d}{m}})$$

Where $S(p) = \sum_{i \in S} p_i$

(2) Show that if $m = O(\frac{n + \log \frac{1}{\delta}}{\epsilon^2})$ then with probability at least $1 - \delta$, the L_1-distance between the empirical distribution $\frac{1}{m} \sum_{i=1}^m \delta_{X_i}$ and p is less than ϵ. Where δ_{X_i} is the Dirac delta function.

(3)The Kolmogorov's distance between two distributions p and q is $\max_I |p(\{1,...,i\}) - q(\{1,...,i\})|$, i.e. the largest discrepency between their CDFs. Such that, if $m = O(\frac{n+\log\frac{1}{\delta}}{\epsilon^2})$ then with probability at least $1-\delta$, the Kolmogorov's distance between $\frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ and p is less than ϵ.

Proof. (1) We can introduce random variables $X_1', X_2', ..., X_m'$ as in the proof of symmetrization and we can fix the set $\{X_1, X_2, ..., X_n, X_1', X_2', ..., X_n'\}$ when we calculate the expectation. Denote $\phi(S) = \frac{1}{m}\sum_{i=1}^m I[X_i \in S] - \frac{1}{m}\sum_{i=1}^m I[X_i' \in S]$. It is clear that $\mathbb{P}[\phi(S) > a] = \mathbb{P}[\phi(S) < -a]$.

$$
\begin{aligned}
&\mathrm{E}_{X_i}[\sup_{S\in H}|\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)|]\\
=&\mathrm{E}_{X_i}[\sup_{S\in H}|\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - \mathrm{E}_{X_i'}[\frac{1}{m}\sum_{i=1}^m I[X_i' \in S]|]]\\
\leq&\mathrm{E}_{X_i,X_i'}[\sup_{S\in H}|\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - \frac{1}{m}\sum_{i=1}^m I[X_i' \in S]|]\\
=&\frac{1}{\lambda}\log\exp(\lambda \mathrm{E}_{X_i,X_i'}[\sup_{S\in H}|\phi(S)|])\\
\leq&\frac{1}{\lambda}\log \mathrm{E}_{X_i,X_i'}\exp(\lambda[\sup_{S\in H}|\phi(S)|])\\
\leq&\frac{1}{\lambda}\log \mathrm{E}_{X_i,X_i'}\sum_{S\in H} 2\exp(\lambda\phi(S))\\
=&\frac{1}{\lambda}\log\sum_{S\in H} 2\mathrm{E}_{X_i,X_i'}\exp(\frac{\lambda}{m}[\sum_{i=1}^m(I[X_i \in S] - I[X_i' \in S])])\\
\leq&\frac{1}{\lambda}\log\sum_{S\in H} 2\exp(\frac{\lambda^2}{2m})
\end{aligned}
$$

The first and second inequality is Jensen's inequality. The third inequality is straight up and the forth inequality is Hoeffding inequality. If we set $\lambda = \sqrt{2mlog(2|H|)}$, then

$$
\begin{aligned}
&\mathrm{E}_{X_i}[\sup_{S\in H}|\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)|]\\
\leq&\frac{1}{\lambda}\log 2|H|\exp(\frac{\lambda^2}{2m})\\
=&\sqrt{\frac{2log(2|H|)}{m}} \leq \sqrt{\frac{2log(2(\frac{en}{d})^d)}{m}}\\
=&O(\sqrt{\frac{d}{m}})
\end{aligned}
$$

(2)

$$
\begin{aligned}
&\| \frac{1}{m}\sum_{i=1}^m \delta_{X_i} - p \|_{L_1}\\
=&\sum_{j=1}^n |\frac{1}{m}\sum_{i=1}^m I[X_i \in \{j\}] - p_j|\\
=&|\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)| + |\frac{1}{m}\sum_{i=1}^m I[X_i \in S'] - S'(p)|
\end{aligned}
$$

for some $S \cap S' = \emptyset$, $S \cup S' = [n]$, so $\parallel \frac{1}{m}\sum_{i=1}^m \delta_{X_i} - p \parallel_{L_1} \leq 2\sup_{S\in P([n])} |\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)|$

$$\begin{aligned}
&\mathbb{P}(\parallel \frac{1}{m}\sum_{i=1}^m \delta_{X_i} - p \parallel_{L_1} \geq \epsilon) \\
\leq&\mathbb{P}(\sup_{S\in P([n])} |\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)| \geq \frac{\epsilon}{2}) \\
\leq&O(2^n e^{-O(m\epsilon^2)}) \\
=&O(\delta)
\end{aligned}$$

(3) Let $H = \{\emptyset, \{1\}\, \{1,2\} ..., \{1,2,...,n\}\}$, $|H| = n+1$. It is clear that $\parallel \frac{1}{m}\sum_{i=1}^m \delta_{X_i} - p \parallel_{Kolmogorov} = \sup_{S\in H} |\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)|$.

$$\begin{aligned}
&\mathbb{P}(\parallel \frac{1}{m}\sum_{i=1}^m \delta_{X_i} - p \parallel_{Kolmogorov} \geq \epsilon) \\
=&\mathbb{P}(\sup_{S\in H} |\frac{1}{m}\sum_{i=1}^m I[X_i \in S] - S(p)| \geq \epsilon) \\
\leq&O((n+1))e^{-O(m\epsilon^2)}) \\
=&O(\delta)
\end{aligned}$$

□

Problem 5. *Show the dual and primal programming in Lagrange Duality theory has the same optimal value.(if one of them exists.)*

Proof. We want to prove that:

$$\begin{cases} min f(x) \\ s.t. g_i(x) \leq 0 \\ h_i(x) = 0 \end{cases}$$

$$\Longleftrightarrow \begin{cases} min_x max_{\mu,\lambda} f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) \\ s.t. \lambda_i \geq 0 \end{cases}$$

If $g_i(x) > 0$ or $h_i(x) \neq 0$, then $max_{\mu,\lambda} f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) = +\infty$. When $g_i(x) \leq 0$ and $h_i(x) = 0$, $min_x max_{\mu,\lambda} f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) = min_x f(x)$ which is the left hand side. So the optimal values are the same (if exists).

□

Problem 6. *Show that KKT conditions are necessary and if f, g_i are convex and each h_i is linear, then it's also sufficient for (X^*, λ^*, μ^*) to be the optima of primal and dual programmings.*

Proof. Let $L(x,\mu,\lambda) = f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x)$. (X^*, λ^*, μ^*) solve the problem $max_{\mu,\lambda} min_x L(x,\mu,\lambda) = max_{\mu,\lambda} min_x L(x,\mu,\lambda)$ subject to $\lambda_i \geq 0$. Then by taking derivative, we get $\nabla_x L(x,\lambda,\mu)|_{x^*,\lambda^*,\mu^*} = 0$. $h_i(x^*) = 0, g_i(x^*) \leq 0$ and $\lambda_i^* \geq 0$ are obvious. If $\lambda_i > 0$, then $g_i(x^*) = 0$, so $\lambda_i g_i(x^*) = 0$. KKT conditions are sufficient.

If f, g_i are convex, each h_i is linear and KKT conditions are satisfied. Then $L(x,\mu,\lambda)$ is convex in x. Because $\nabla_x L(x,\lambda,\mu)|_{x^*,\lambda^*,\mu^*} = 0$, $L(x^*,\mu^*,\lambda^*) = min_x L(x,\mu^*,\lambda^*)$. $max_{\mu,\lambda} min_x L(x,\mu,\lambda) \geq L(x^*,\mu^*,\lambda^*) = f(x^*) \geq min f(x) = min_x max_{\mu,\lambda} L(x,\mu,\lambda)$. So (X^*, λ^*, μ^*) are the optima of primal and dual programmings. □

Problem 7. *Assume $p : [n] \to [0,1]$ is a distribution over $[n] = \{1, 2, ..., n\}$. Suppose $m' \sim Poi(m)$ is a random variable has Poisson distribution, show that if we take m' samples indepdently from p and let X_i denote the occurrences of i, then $X_i \sim Poi(mp_i)$ and $X_1, ..., X_n$ are independent.*

Proof. $P(m' = k) = \frac{m^k}{k!}e^{-m}$, $P(X_i = k) = \sum_{j=k}^{\infty} \frac{m^j}{j!}e^{-m}C_j^k p_i^k (1-p_i)^{j-k} = \frac{(mp_i)^k}{k!}e^{-m}\sum_{i=0}^{\infty}\frac{(m-mp_i)^j}{j!} = \frac{(mp_i)^k}{k!}e^{-mp_i}$. So $X_i \sim Poi(mp_i)$.

$P(X_1 = x_1, ..., X_n = x_n) = \frac{m^{x_1+\ldots+x_n}}{(x_1+\ldots+x_n)!}e^{-m}\frac{(x_1+\ldots+x_n)!}{x_1!\ldots x_n!}\prod(p_i)_i^x = \prod \frac{(mp_i)_i^x}{x_i!}e^{-mp_i} = P(X_1 = x_1)...P(X_n = x_n)$, so $X_1, ..., X_n$ are independent.

$\square$