FOML Assignment-5

AI21BTECH11011 Gunjit Mittal

5.1 Q1 Loss function $L_c(h)$ is given by

$$L_c(h) = |\{i \in \{1, \dots, m\} : h(x_i) \neq y_i\}|$$

ERM problem for domain Z and input sample $S \in Z^m$ is given by

$$L_S(h) = \frac{1}{|S|} \sum_{p \in S} \ell(h, p)$$

The sample complexity of learning a finite class is upper bounded by $m_H(\epsilon, \delta) = c \log(c|H|/\delta)/\epsilon^c$

Where $c = 1$ in realizable case and $c = 2$ in the non realizable one

Performing exhaustive search the run time is $K|H|c \log(c|H|/\delta)/\epsilon^c$

Let us consider the example in which $H_1 \subseteq H_2 \dots \subseteq H_K$ and $|H_i| = 2^i \ \forall i \in K$. Learning H_K in Agnostic-Pac model provides the following bound for an ERM hypothesis h:

$$L_D(h) = \min_{h \in H_K} L_D(h) + \sqrt{\frac{2(K_H + \log(1/\delta))}{m}}$$

Using model selection, assuming j is the minimal index which contains $h^* \in \underset{h \in H}{\operatorname{argmin}} L_D(h)$ and fixing $r \in \{1, \dots, K\}$

By Hoeffding's inequality we have

$$|L_D(\hat{h}_k) - L_V(\hat{h}_k)| \le \sqrt{\frac{1}{2\alpha m}\log\left(\frac{4}{\delta}\right)}$$

is true with probability $\ge 1 - \frac{\delta}{2k}$

Applying the union bound we have with probability $\ge 1 - \frac{\delta}{2}$

$$\begin{aligned} L_D(\hat{h}) &\le L_V(\hat{h}) + \sqrt{\frac{1}{2\alpha m}\log\left(\frac{4k}{\delta}\right)} \\ &\le L_V(\hat{h}_k) + \sqrt{\frac{1}{2\alpha m}\log\left(\frac{4k}{\delta}\right)} \\ &\le L_D(\hat{h}_k) + 2\sqrt{\frac{1}{2\alpha m}\log\left(\frac{4k}{\delta}\right)} \\ &= L_D(\hat{h}_k) + \sqrt{\frac{2}{\alpha m}\log\left(\frac{4k}{\delta}\right)} \end{aligned}$$

for $k \ge j$ we have

$$L_D(\hat{h}) \le L_D(\hat{h}_j) + \sqrt{\frac{2}{\alpha m}\log\left(\frac{4k}{\delta}\right)}$$

Similarly we have with probability $> 1 - \frac{\delta}{2}$

$$L_D(\hat{h}_j) \le L_D(h^*) + \sqrt{\frac{2}{(1-\alpha)m}\log\left(\frac{4|H_j|}{\delta}\right)}$$

Combining last two we have with probability $> 1 - \delta$

$$L_D(\hat{h}) \le L_D(h^*) + \sqrt{\frac{2}{\alpha m}\log\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\log\left(\frac{4|H_j|}{\delta}\right)}$$

Substituting $|H_j| = 2^j$ we get

$$L_D(\hat{h}) \leq L_D(h^*) + \sqrt{\frac{2}{\alpha m}\log\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\left(j + \log\left(\frac{4}{\delta}\right)\right)}$$

From this we can observe that when the optimal index j is much smaller than k ~~using~~ model selection is the better approach

5.3 ~~Q1~~ Q16.3 According to the representer theorem the minimiser of training error lies in ~~span~~ span $\{\psi(x_1), \psi(x_2) \ldots \psi(x_m)\}$.

$\therefore$ The ERM problem can be rewritten as

$$\min_{\alpha \in \mathbb{R}^m} \lambda \left\| \sum_{i=1}^{m} \alpha_i \psi(x_i) \right\|^2 + \frac{1}{2m}\sum_{i=1}^{m}\left(\left\langle \sum_{j=1}^{m} \alpha_j \psi(x_j), \psi(x_i)\right\rangle - y_i\right)^2$$

Using gram matrix we can write it as

$$\min_{\alpha \in \mathbb{R}^m} \lambda \alpha^T G \alpha + \frac{1}{2m}\sum_{i=1}^{m}\left(\langle \alpha, G_i \rangle - y_i\right)^2$$

Since it is convex we differentiate it and get

$$(\lambda' G + G G^T)\alpha - G y = 0 \qquad (\lambda' = m\lambda)$$

As G is symmetric

$$G(\lambda' I + G^T)\alpha = G y$$

If G is invertible

$$(\lambda' I + G^T)\alpha = y$$

Since G is positive semi definite and $\lambda > 0$

$\lambda' I + G$ is positive definitive

$\therefore$ $(\lambda' I + G^T)$ is invertible

$\therefore$ the minimiser will be

$$\alpha = (\lambda' I + G^T)^{-1} y$$

Q6.4 Let $\psi = \{1, 2, \ldots N\} \to R^N$

$$\psi_j = (1^j, 0^{N-j})$$

where 1^j is the one vectors in R^j
and 0^{N-j} is the zeroes vector in R^{N-j}
then

$$\langle \psi_i, \psi_j \rangle = \langle (1^i, 0^{N-i}), (1^j, 0^{N-j}) \rangle$$
$$= \min\{i, j\}$$
$$= K(i, j)$$

Q2 Q6.1 The cauchy schwartz inequality for PDS kernels state that for a PDS kernel K, for any $x, x' \in X$

$$K(x, x')^2 \leq K(x, x) \; K(x', x')$$

For the kernel $K'(x, y) = \dfrac{K(x, y)}{\alpha(x)\,\alpha(y)}$

~~Let G be kernel of~~
Let G be the gram matrix of the above kernel
Considering any column matrix C
with column vectors $c_1, c_2, \ldots c_m \in R^{m \times 1}$

$$C^T G C = \sum_{i,j=1}^{m} \frac{c_i c_j K(x_i, y_i)}{\alpha(x_i)\,\alpha(y_i)}$$

since both the numerators and denominators ~~evolute~~ evolute to > 0

$C^T G C > 0$

which implies $K'(x,y)$ is PDS

Q6.2 a) $K(x,y) = \cos(x-y)$ over $R \times R$

$\cos(x-y) = \cos x \cos y + \sin x \sin y$

$K(x,y)$ can be written as inner product of

$\phi(x) = \begin{bmatrix} \cos x \\ \sin y \end{bmatrix}$ ~~Sin~~ and $\phi(y) = \begin{bmatrix} \cos y \\ \sin y \end{bmatrix}$

b) $K(x,y) = \cos(x^2 - y^2)$ over $R \times R$

$$C^T G C = \sum_{i,j=1}^{m} c_i c_j \cos(x_i^2 - y_i^2)$$

Taking $x_i^2 = x_i'$ & $y_i^2 = y_i'$

$$= \sum_{i,j=1}^{m} c_i c_j \cos(x_i' - y_i')$$

which is the same as proving $\cos(x-y)$ is PDS

c) $K(x,y) = \sum_{i=1}^{N} \cos^n (x_i^2 - y_i^2)$

since each term in $K(x,y)$ is PDS

$K(x,y)$ will also be PDS

e) $K(x, x') = \cos \angle (x, x')$

$$c^T G c = \sum_{i,j=1}^{N} c_i c_j \cos(\angle(x_i, x_j'))$$

~~wh~~ can be written as $\cos(\alpha_i - \alpha_j)$ for some α_i & α_j

so it becomes the same as a)

hence $K(x, x') = \cos \angle (x, x')$ is PDS