Home Assignment 3

Yu Wang (ndp689)

December 14, 2021

Contents

1 Kernels

1.1 Distance in feature space

$$\begin{aligned}\|\Phi(x)-\Phi(z)\| &= \sqrt{\|\Phi(x)-\Phi(z)\|^2} \\ &= \sqrt{<\Phi(x),\Phi(x)> -2 <\Phi(x),\Phi(z)> + <\Phi(z),\Phi(z)>} \\ &= \sqrt{<k(x,\cdot),k(x,\cdot)> -2 <k(x,\cdot),k(z,\cdot)> + <k(x,\cdot),k(z,\cdot)>} \\ &= \sqrt{k(x,x)-2k(x,z)+k(z,z)}\end{aligned}$$

1.2 Sum of kernels

Let $m \in \mathbb{N}$, $x_1, ..., x_m \in \mathcal{X}$, the kernel matrix of k_1 with respect to $x_1, ..., x_m$ is the $m \times m$ matrix $\mathbf{K1}$ with elements $K1_{ij} = k_1(x_i, x_j)$, and the kernel matrix of k_2 with respect to $x_1, ..., x_m$ is the $m \times m$ matrix $\mathbf{K2}$ with elements $K2_{ij} = k_2(x_i, x_j)$.

Since k_1, k_2 are positive-definite kernels, then we have:

$$\begin{aligned}\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j=1}^{m} c_i c_j K1_{ij} \\ = \sum_{i,j=1}^{m} c_i c_j k_1(x_i, x_j) \geq 0\end{aligned} \tag{1}$$

and

$$\begin{aligned}\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j=1}^{m} c_i c_j K2_{ij} \\ = \sum_{i,j=1}^{m} c_i c_j k_2(x_i, x_j) \geq 0\end{aligned} \tag{2}$$

Then, for $k(x,z) = a \cdot k_1(x,z) + b \cdot k_2(x,z)$ $(a, b \in \mathbb{R}^+)$, we define the kernel matrix of k with respect to $x_1, ..., x_m$ is the $m \times m$ matrix $\mathbf{K}$ with elements

$K_{ij} = k(x_i, x_j)$. Combined with equations (1) and (2), we have:

$$\begin{aligned}
\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j=1}^{m} c_i c_j K_{ij} \\
&= \sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \\
&= \sum_{i,j=1}^{m} c_i c_j \left(a \cdot k_1(x_i, x_j) + b \cdot k_2(x_i, x_j)\right) \\
&= a \cdot \sum_{i,j=1}^{m} c_i c_j k_1(x_i, x_j) + b \cdot \sum_{i,j=1}^{m} c_i c_j k_2(x_i, x_j) \geq 0
\end{aligned}$$

Thus, the kernel matrix satisfies $\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j=1}^{m} c_i c_j K_{ij} \geq 0$. So, k is the positive-definite kernel.

1.3 Rank of Gram matrix

Let $\mathbf{A}$ be a $d \times m$ matrix with respect to $x_1, ..., x_m \in \mathbb{R}^d$. Since $k(x, z) = x^T z$ for $x, z \in \mathbb{R}^d$ and m input patterns $x_1, ..., x_m \in \mathbb{R}^d$, then, we let the its Gram matrix $\mathbf{K} = \mathbf{A}^T \mathbf{A}$.

Assuming for a m columns matrix $\mathbf{Z}$ the $Null(\mathbf{Z}) = \{x \mid \mathbf{Z}x = \mathbf{0}, x \in \mathbb{R}^m\}$.

Let $x \in Null(\mathbf{A})$, so, $\mathbf{A}x = \mathbf{0}$, and $\mathbf{A}^T \mathbf{A}x = \mathbf{0}$. Then, $x \in Null(\mathbf{A}^T \mathbf{A})$, and we can get $Null(\mathbf{A}) \subseteq Null(\mathbf{A}^T \mathbf{A})$. Similarly, let $x \in Null(\mathbf{A}^T \mathbf{A})$, so, $\mathbf{A}^T \mathbf{A}x = \mathbf{0}$, and $x^T \mathbf{A}^T \mathbf{A}x = (\mathbf{A}x)^T (\mathbf{A}x) = \mathbf{0}$, thus $\mathbf{A}x = \mathbf{0}$. Then, $x \in Null(\mathbf{A})$, and we can get $Null(\mathbf{A}) \subseteq Null(\mathbf{A}^T \mathbf{A})$.

Combined with the above two assumptions, we can get $Null(\mathbf{A}) = Null(\mathbf{A}^T \mathbf{A})$, then,

$$Nullity(\mathbf{A}) = Nullity(\mathbf{A}^T \mathbf{A}) \tag{3}$$

According to the Rank–nullity theorem, for $\mathbf{A}^T \mathbf{A}$, we have:

$$m = Rank(\mathbf{A}^T \mathbf{A}) + Nullity(\mathbf{A}^T \mathbf{A}) \tag{4}$$

And for $\mathbf{A}$, we have:

$$m = Rank(\mathbf{A}) + Nullity(\mathbf{A}) \tag{5}$$

So, according to the equations (3), (4) and (5), we have:

$$Rank(\mathbf{A}) = Rank(\mathbf{A}^T \mathbf{A}) = Rank(\mathbf{K})$$

Then, $Rank(\mathbf{K}) = Rank(\mathbf{A}) \leq min\{m, d\}$

2 Early stopping

1. In which of the following cases is $\hat{L}(h_{t^*}, S_{val})$ an unbiased estimate of $L(h_{t^*})$ and in which cases is it not.

   (a) In this case $\hat{L}(h_{t^*}, S_{val})$ is an unbiased estimate of $L(h_{t^*})$. Because we have $h_{t^*} = h_{100}$. Then, the choosing of t^* does not depend on S_{val}.

   (b) In this case $\hat{L}(h_{t^*}, S_{val})$ is not an unbiased estimate of $L(h_{t^*})$. Because we have $t^* = arg \min_{t \in \{1,\ldots,T\}} \hat{L}(h_t, S_{val})$. Then, the choosing of t^* depends on S_{val}.

   (c) In this case $\hat{L}(h_{t^*}, S_{val})$ is not an unbiased estimate of $L(h_{t^*})$. Because the training procedure of this case stops when no improvement in $\hat{L}(h_t, S_{val})$ is observed for a significant number of epochs. Then, the choosing of t^* depends on S_{val}.

2. Derive a high-probability bound (a bound that holds with probability at least $1 - \delta$) on $L(h_{t^*})$.

   (a) Predefined stopping

   Since $\hat{L}(h_{t^*}, S_{val})$ is an unbiased estimate of $L(h_{t^*})$, then, we need to prove that $E\left[\hat{L}(h_{t^*}, S_{val})\right] = L(h_{t^*})$:

$$
\begin{aligned}
E\left[\hat{L}(h_{t^*}, S_{val})\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} \ell(h_{t^*}(X_i), Y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[\ell(h_{t^*}(X_i), Y_i)\right] \\
&= L(h_{t^*})
\end{aligned}
$$

   Where n is the size of S_{val}.

   Then, according to the Hoeffding's inequality, we have:

$$
P\left(L(h_{t^*}) - \hat{L}(h_{t^*}, S_{val}) \geq \epsilon\right) \leq e^{-2n\epsilon^2}
$$

And then, let $\delta = e^{-2n\epsilon^2}$, so, $\epsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$. We have:

$$P\left(L(h_{t^*}) - \hat{L}(h_{t^*}, S_{val}) \geq \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \leq \delta \iff$$

$$P\left(L(h_{t^*}) - \hat{L}(h_{t^*}, S_{val}) \leq \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \geq 1 - \delta \iff$$

$$P\left(L(h_{t^*}) \leq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \geq 1 - \delta$$

Then, we have a bound on $L(h_{t^*})$ in terms of $\hat{L}(h_{t^*}, S_{val})$, δ, and n that holds with probability at least $1 - \delta$.

(b) Non-adaptive stopping

Let $\mathcal{H} = \{h_1, h_2, ..., h_T\}$, $|\mathcal{H}| = T$, and $t^* = arg \min_{t \in \{1,...,T\}} \hat{L}(h_t, S_{val})$. Then, according to the corollary, we have

$$P\left(L(h_{t^*}) \geq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{\ln \frac{T}{\delta}}{2n}}\right) \leq \delta \iff$$

$$P\left(L(h_{t^*}) \leq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{\ln \frac{T}{\delta}}{2n}}\right) \geq 1 - \delta$$

Then, we have a bound on $L(h_{t^*})$ in terms of $\hat{L}(h_{t^*}, S_{val})$, δ, T, and n that holds with probability at least $1 - \delta$.

(c) Adaptive stopping

We define $\pi(\mathcal{H}_t) = \frac{1}{t(t+1)}$ and $|\mathcal{H}_t| = t$, then $\pi(h) = \pi(\mathcal{H}_t)\frac{1}{|\mathcal{H}_t|}$. So, we have

$$\sum_{h \in \mathcal{H}} \pi(h) = \sum_{t=1}^{\infty} \frac{1}{t(t+1)} \sum_{h \in \mathcal{H}_t} \frac{1}{t} = \sum_{t=1}^{\infty} \frac{1}{t(t+1)} = 1$$

Then, according to the Occam's razor theorem, we have,

$$P\left(\exists h \in \mathcal{H} : L(h_t) \geq \hat{L}(h_t, S_{val}) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta \iff$$

$$P\left(\exists h \in \mathcal{H} : L(h_t) \geq \hat{L}(h_t, S_{val}) + \sqrt{\frac{\ln \frac{t^2(t+1)}{\delta}}{2n}}\right) \leq \delta \iff$$

$$P\left(\forall h \in \mathcal{H} : L(h_t) \leq \hat{L}(h_t, S_{val}) + \sqrt{\frac{\ln \frac{t^2(t+1)}{\delta}}{2n}}\right) \geq 1 - \delta$$

So, for the best model h_{t^*}, we have,

$$P\left(L(h_{t^*}) \leq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{\ln \frac{(t^*)^2(t^*+1)}{\delta}}{2n}}\right) \geq 1 - \delta$$

Then, we have a bound on $L(h_{t^*})$ in terms of $\hat{L}(h_{t^*}, S_{val})$, δ, t^*, and n that holds with probability at least $1 - \delta$.

3. Since ℓ be bounded in $[0, 1]$, then $L(h_t) - \hat{L}(h_t, S_{val})$ is bounded in 1. So, we have,

$$\sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}} \leq 1 \iff$$
$$\pi(h) \geq \frac{e^{-2n}}{\delta}$$

Since $\pi(h) = \pi(\mathcal{H}_t)\frac{1}{|\mathcal{H}_t|} = \frac{1}{t(t+1)}\frac{1}{t}$, then, we have

$$\frac{1}{t(t+1)}\frac{1}{t} \geq \frac{e^{-2n}}{\delta} \iff$$
$$t^2(t+1) \leq \delta e^{2n}$$

So, $T_{max}^2(T_{max} + 1) = \delta e^{2n}$

4. If use the series $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$, $\pi(h) = \frac{1}{2^t}\frac{1}{t}$. From Point 3, we have,

$$\frac{1}{2^t}\frac{1}{t} \geq \frac{e^{-2n}}{\delta} \iff$$
$$t \cdot 2^t \leq \delta e^{2n}$$

So, $T_{max}2^{T_{max}} = \delta e^{2n}$. For the $T_{max}(T_{max}+1)$ and the $2^{T_{max}}$ when they have the same upper bound, we can know that the maximum value T_{max} of the latter is smaller, which means that with the series $\sum_{i=1}^{\infty}\frac{1}{2^i} = 1$ we can run significantly less epochs.

5. In this question we compare the adaptive procedure with non-adaptive.

   (a) For t^*:

$$P\left(L(h_{t^*}) \leq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{\ln\frac{(t^*)^2(t^*+1)}{\delta}}{2n}}\right) \geq 1-\delta$$

   For T^*:

$$P\left(L(h_{T^*}) \leq \hat{L}(h_{T^*}, S_{val}) + \sqrt{\frac{\ln\frac{T}{\delta}}{2n}}\right) \geq 1-\delta$$

   Since we know that the adaptive bound for epoch t^* is lower than the adaptive bound for epoch T^*, and $T^* \leq T$. Then, we have

$$\begin{aligned}
L(h_{t^*}) &\leq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{\ln\frac{(t^*)^2(t^*+1)}{\delta}}{2n}} \\
&\leq \hat{L}(h_{T^*}, S_{val}) + \sqrt{\frac{\ln\frac{(T^*)^2(T^*+1)}{\delta}}{2n}} \\
&\leq \hat{L}(h_{T^*}, S_{val}) + \sqrt{\frac{\ln\frac{T^2(T+1)}{\delta}}{2n}}
\end{aligned}$$

   Since $\delta \leq \frac{T}{T+1}$. Then $T \leq \frac{1}{\delta(T+1)}$ and $T+1 \leq \frac{T}{\delta}$, we have

$$\begin{aligned}
L(h_{t^*}) &\leq \hat{L}(h_{T^*}, S_{val}) + \sqrt{\frac{\ln(\frac{T(T+1)}{\delta} \times \frac{1}{\delta(T+1)})}{2n}} \\
&\leq \hat{L}(h_{T^*}, S_{val}) + \sqrt{\frac{\ln(\frac{T}{\delta})^2}{2n}} \\
&= \hat{L}(h_{T^*}, S_{val}) + \sqrt{2} \times \sqrt{\frac{\ln\frac{T}{\delta}}{2n}} \\
&\leq \sqrt{2}\left(\hat{L}(h_{T^*}, S_{val}) + \sqrt{\frac{\ln\frac{T}{\delta}}{2n}}\right) \\
&= \sqrt{2}L(h_{T^*})
\end{aligned}$$

So, the adaptive bound can be at most a multiplicative factor of $\sqrt{2}$ larger than the non-adaptive bound.

(b)

(c)

3 Learning by discretization

1. We define $\pi(\mathcal{H}_{d(h)}) = \frac{1}{2^{d(h)}}$ and $|\mathcal{H}_{d(h)}| = 2^{f(n)}$, then $\pi(h) = \pi(\mathcal{H}_{d(h)})\frac{1}{|\mathcal{H}_{d(h)}|}$, and $\sum_{h\in\mathcal{H}}\pi(h) = \sum_{d(h)=1}^{\infty}\frac{1}{2^{d(h)}}\sum_{h\in\mathcal{H}_{d(h)}}\frac{1}{2^{f(n)}} = \sum_{d(h)=1}^{\infty}\frac{1}{2^{d(h)}} = 1$

   Then, for the $L(h)$ and its unbiased estimate $\hat{L}(h,S)$, according to the Occam's razor theorem, we have,

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta \iff$$

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{(d^2(h) + d(h))\ln(2) + \ln\frac{1}{\delta}}{2n}}\right) \leq \delta \iff$$

$$P\left(\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h,S) + \sqrt{\frac{(d^2(h) + d(h))\ln(2) + \ln\frac{1}{\delta}}{2n}}\right) \geq 1-\delta$$

2. we can choose h^* by $h^* = arg\min_h\left(\hat{L}(h,S) + \sqrt{\frac{(d^2(h)+d(h))\ln(2)+\ln\frac{1}{\delta}}{2n}}\right)$

3. Since ℓ be bounded in $[0,1]$, then $L(h_t) - \hat{L}(h_t,S)$ is bounded in 1. So, we have,

$$\sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}} \leq 1 \iff$$
$$\pi(h) \geq \frac{e^{-2n}}{\delta}$$

   Since $\pi(h) = \frac{1}{2^{d(h)}}\frac{1}{2^{d^2(h)}}$, then, we have

$$\frac{1}{2^d}\frac{1}{2^{d^2}} \geq \frac{e^{-2n}}{\delta} \iff$$
$$d + d^2 \leq \log_2(\delta e^{2n})$$

   Then the max number of cells is d_{max}^2, where $d_{max} + d_{max}^2 = \log_2(\delta e^{2n})$

4. From Point 1, we know that: d(h) in the bound increase as the density of the grid increases, while n in the bound decrease as the density of the grid increases.