Machine Learning
Assignment 3
Template

Mao Xiqing (tls868)

December 7, 2020

Contents

1 The Role of Independence

Let Bernoulli random variables $X_1, X_2, \ldots, X_n, (X_i \in [0,1])$ be a sequence of a Markov chain, the emission probabilities $P(X_i = 1) = 0.5$ and $P(X_i = 0) = 0.5$. Let $P_{i,j} = P(X_2 = j | X_1 = i)$ denote the transition probabilities: $P_{00} = 1, P_{11} = 1, P_{10} = 0, P_{01} = 0$. Thus

$$\begin{aligned} E(X_n) &= P(X_n = 1) \\ &= \frac{P_{01}}{P_{01} + P_{10}} - (P_{00} + P_{11} - 1)^{n-1}(\frac{P_{01}}{P_{01} + P_{10}} - 0.5) \\ &= \frac{1}{2}, \end{aligned} \tag{1}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \begin{cases} 1, & X_1 = 1, \\ 0, & X_1 = 0. \end{cases} \tag{2}$$

Hence

$$|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i| = \frac{1}{2} \longrightarrow \mathbb{P}(|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i| \geq \frac{1}{2}) = 1 \tag{3}$$

2 How to Split a Sample into Training and Test Set

2.1 Warm up

Since samples $(X_i, Y_i) \in S_{train}$ and samples $(X, Y) \in S_{test}$ both come from S and have the same distribution, we have

$$\begin{aligned} \mathbb{E}[L(\hat{h}^*_{S^{train}}, S^{test})] &= \mathbb{E}\left[\frac{1}{n^{test}}\sum_{i=1}^{n^{test}} \ell(\hat{h}^*_{S^{train}}(X_i), Y_i)\right] \\ &= \frac{1}{n^{test}}\sum_{i=1}^{n^{test}} \mathbb{E}\left[\ell(\hat{h}^*_{S^{train}}(X_i), Y_i)\right] \\ &= \frac{1}{n^{test}}\sum_{i=1}^{n^{test}} L(\hat{h}^*_{S^{train}}) \\ &= L(\hat{h}^*_{S^{train}}) \end{aligned} \tag{4}$$

Hoeffding's Inequality can be used to bound them:

$$\mathbb{P}\left(L(\hat{h}^*_{S^{train}}) - \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) \geq \epsilon\right) \leq e^{-2n(test)\epsilon^2}. \tag{5}$$

Let $\delta = e^{-2n\epsilon^2}$, and then $\epsilon = \sqrt{ln(1/\delta)/2n^{test}}$, thus

$$\begin{aligned} &\mathbb{P}(L(\hat{h}^*_{S^{train}}) \geq \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) + \sqrt{ln(\frac{1}{\delta})/2n^{test}}) \leq \delta. \\ \longrightarrow\ &\mathbb{P}(L(\hat{h}^*_{S^{train}}) \leq \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) + \sqrt{ln(\frac{1}{\delta})/2n^{test}} \geq 1 - \delta. \end{aligned} \tag{6}$$

Now $\hat{L}(h^*_{S^{train}})$ in terms of $\hat{L}(h^*_{S^{train}}, S^{test}_i)$ and n^{test} that holds with probability at least $1 - \delta$.

2.2 m possible splits

Let $\mathcal{H} = \{\hat{h}^*_1, \ldots, \hat{h}^*_m\}$ be a finite hypothesis space, and let $\hat{h}^*_i$ be the best hypothesis with a balance sizes of the test n^{test}_i. Since we selected hypothesis from $\mathcal{H}$ based on S^{test}_i, for each hypothesis $h_i \in \mathcal{H}$ individually $\mathbb{E}[\hat{L}(h_i, S^{test}_i)] = L(h_i)$, but $\mathbb{E}[\hat{L}(\hat{h}^*_i, S^{test}_i)] \neq L(\hat{h}^*_i)$. So we also need to apply Union bound for this case:

$$\begin{aligned} \mathbb{P}\left(\exists h^*_i \in \mathcal{H} : L(h^*_i) \geq \hat{L}(h^*_i, S^{test}_i) + \sqrt{\frac{ln(\frac{1}{\delta_i})}{2n^{test}_i}}\right) &\leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(L(h^*_i) \geq \hat{L}(h^*_i, S^{test}_i) + \sqrt{\frac{ln(\frac{1}{\delta_i})}{2n^{test}_i}}\right) \\ &\leq \sum_{h \in \mathcal{H}} \delta_i \\ &= \sum_{h \in \mathcal{H}} \frac{\delta}{m} = \delta \longrightarrow \\ \mathbb{P}\left(\exists h^*_i \in \mathcal{H} : L(h^*_i) \leq \hat{L}(h^*_i, S^{test}_i) + \sqrt{\frac{ln(\frac{m}{\delta})}{2n^{test}_i}}\right) &\geq 1 - \delta. \end{aligned} \tag{7}$$

Now $\hat{L}(h^*_i)$ in terms of $\hat{L}(h^*_i, S^{test}_i)$ and n_i that holds for all h^*_i simultaneously with probability at least $1 - \delta$.

If we fix ϵ and δ, the best hypothesis with a balance sizes of the test n_i can be found:

$$n_i = \frac{1}{2\epsilon^2} ln(m/\delta). \tag{8}$$

3 Occam's Razor

The size of $\mathcal{H}_d$ is 2^{27^d}, where d is the length of strings.

1. We define $\pi(h) = 1/2^{27^{d(h)}}$. Then

$$\mathbb{P}\left(\exists h \in \mathcal{H}_d : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln(2^{27^{d(h)}}/\delta)}{2n}}\right) \leq \delta, \tag{9}$$

where $\sum_{h\in\mathcal{H}_d} \pi(h) < 1$. Hence

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln 2^{27^{d(h)}}/\delta)}{2n}}\right) \geq 1-\delta, \tag{10}$$

2. We define $\pi(h) = \frac{1}{2^{d(h)+1}}(1/2^{27^{d(h)}})$. Then

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h) + \sqrt{\frac{\ln(2^{d(h)+1} \times 2^{27^{d(h)}}/\delta)}{2n}}\right) \leq \delta, \tag{11}$$

where $\sum_{d=0}^{\infty} \frac{1}{2^{d+1}} = 1$ and $\sum_{h\in\mathcal{H}_d} 1/2^{27^{d(h)}} < 1$, thus $\sum_{h\in\mathcal{H}_d} \pi(h) < 1$. Hence

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \leq \hat{L}(h) + \sqrt{\frac{\ln(2^{d(h)+1} \times 2^{27^{d(h)}}/\delta)}{2n}}\right) \geq 1-\delta, \tag{12}$$

3. We want to minimize expected loss of the empirically best hypothesis, which is making $\pi(h)$ to be as large as possible for every h, which means making d is relatively small. However, it is difficult to tell language when d is very small, for example, is i a danish word 'in', or just letter i? And is *museum* English or Danish? Some prior information can be use to make a smaller d while have a good classifying, for example, it is *uncommon* for English users only type an i, so it can be classified to Danish.

4. Since $d \longrightarrow \infty$ and $\delta \in [0,1]$, we have

$$\sqrt{\frac{\ln(2^{d(h)+1} \times 2^{27^{d(h)}}/\delta)}{2n}} \geq \sqrt{\frac{\ln(2^{d(h)+1+27^{d(h)}})}{2n}} \geq \sqrt{\frac{\ln(2^{2n})}{2n}} = \sqrt{ln(2)} \approx 0.8, \tag{13}$$

which has no contradiction with $L(h) \geq 0.25$.

4 Kernels

4.1 Distance in feature space

$$
\begin{aligned}
\|\phi(x)-\phi(z)\| &= \sqrt{\|\phi(x)-\phi(z)\|^2} = \sqrt{[\phi(x)-\phi(z)][\phi(x)-\phi(z)]} \\
&= \sqrt{<\phi(x),\phi(x)> + <\phi(z),\phi(z)> -2<\phi(x),\phi(z)>} \\
&= \sqrt{<k(x,\cdot),k(x,\cdot)> + <k(z,\cdot),k(z,\cdot)> -2<k(x,\cdot),k(z,\cdot)>} \\
&= \sqrt{k(x,x)-2k(x,z)+k(z,z)}
\end{aligned} \tag{14}
$$

4.2 Sum of kernels

From the question: $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive-definite kernels, we have:

$$
\begin{aligned}
k_1(x,z) = \phi_1(x)\cdot\phi_1(z), \\
k_2(x,z) = \phi_2(x)\cdot\phi_2(z),
\end{aligned} \tag{15}
$$

and Gram matrix: $[K_{ij}]_{m\times m} = [K(x_i,x_j)]_{m\times m}$. Thus, for any $c_1, c_2, \ldots, c_m \in \mathbb{R}$:

$$
\begin{aligned}
k(x,z) &= k_1(x,z)+k_2(x,z) \longrightarrow \\
\sum_{i,j=1}^{m} c_i c_j k(x_i,x_i) &= \sum_{i,j=1}^{m} c_i c_j [k_1(x_i,x_j)+k_2(x_i,x_j)] \\
&= \sum_{i,j=1}^{m} c_i c_j k_1(x_i,x_j) + \sum_{i,j=1}^{m} c_i c_j k_2(x_i,x_j) \\
&= \sum_{i,j=1}^{m} c_i c_j \phi_1(x_i)\cdot\phi_1(x_j) + \sum_{i,j=1}^{m} c_i c_j \phi_2(x_i)\cdot\phi_2(x_j) \\
&= \sum c_i\phi_1(x_i)\sum c_j\phi_1(x_j) + \sum c_i\phi_2(x_i)\sum c_j\phi_2(x_j) \\
&= \|\sum c\phi_1(x)\|^2 + \|\sum c\phi_2(x)\|^2 \geq 0.
\end{aligned} \tag{16}
$$

Hence, $k(x,z)$ is also positive-definite.

4.3 Rank of Gram matrix

The linear kernel $k(x,z)$ is equivalent to dot product $<x,z>$ for $x,z \in \mathbb{R}^d$. Let A be a $d\times m$ matrix with $x_1,\ldots,x_m \in \mathbb{R}^d$, and Gram matrix $\mathbf{G} = A^TA$.

Assuming that $\mathbf{z} \in Nul(A)$, then$A\mathbf{z} = 0$, and then $A^TA\mathbf{z} = 0 \longrightarrow \mathbf{z} \in Nul(A^TA) \longrightarrow Nul(A) \subseteq Nul(A^TA)$.

Assuming that $\mathbf{z} \in Nul(A^TA)$, then $A^TA\mathbf{z} = 0$, and then $\mathbf{z}^TA^TA\mathbf{z} = (A\mathbf{z})^TA\mathbf{z} =< A\mathbf{z}, A\mathbf{z} >= 0 \longrightarrow A\mathbf{z} = 0 \longrightarrow Nul(A^TA) \subseteq Nul(A)$, thus $Nul(A) = Nul(A^TA)$.

A^TA and A have the same d columns, so from Rank Theorem we have $rank(A) + dim(Nul(A)) = rank(A^TA) + dim(Nul(A^TA))$. Hence, $rank(A) = rank(A^TA) = rank(\mathbf{G})$.

Hence $rank(\mathbf{G}) \leq min\{m, d\}$.

4.4 Cauchy-Schwarz

It is obverse that

$$\langle x + \alpha y, x + \alpha y\rangle = \langle x, x\rangle + \alpha^2\langle y, y\rangle + 2\alpha\langle x, y\rangle \geq 0. \tag{17}$$

(1). If $\langle x, x\rangle = \langle y, y\rangle = 0$, and $\alpha = \pm 1$, let $x = \frac{1}{2}y$. Then:

$$\begin{aligned} 2\alpha\langle x, y\rangle = 2\alpha\langle \frac{1}{2}y, y\rangle = 2\alpha \times \frac{1}{2}\langle y, y\rangle = \langle y, y\rangle = 0 \\ \longrightarrow \langle x, y\rangle = 0. \end{aligned} \tag{18}$$

so we have

$$\begin{aligned} |\langle x, y\rangle| &= |\langle \frac{1}{2}y, y\rangle| = \frac{1}{2}|\langle y, y\rangle| \\ &= \frac{1}{2}\|y\|\|y\| = \|x\|\|y\| \\ &= 0 \\ \longrightarrow |\langle x, y\rangle|^2 &= \|x\|^2\|y\|^2 = \langle x, x\rangle\langle y, y\rangle = 0. \end{aligned} \tag{19}$$

(2). If $\langle y, y\rangle \neq 0$, let$\langle x, x\rangle = a, \langle y, y\rangle = b, 2\langle x, y\rangle = c$ and $\alpha = -c/2b$. Then:

$$\begin{aligned} \langle x + \alpha y, x + \alpha y\rangle &= a + \alpha^2 b + \alpha c = a + (\frac{c}{2b})^2 b - \frac{c}{2b}c \\ &= a + \frac{c^2}{4b} - \frac{c^2}{2b} = a - \frac{c^2}{4b} > 0 \longrightarrow 4ab > c^2 \end{aligned} \tag{20}$$

thus $4\langle x, x\rangle \cdot \langle y, y\rangle > (2\langle x, y\rangle)^2 \longrightarrow 4\langle x, x\rangle\langle y, y\rangle > 4\langle x, y\rangle^2 \longrightarrow \langle x, y\rangle^2 < \langle x, x\rangle\langle y, y\rangle$.

Hence $|\langle x, y\rangle^2| \leq \langle x, x\rangle\langle y, y\rangle$.