Aashsih Kumar pcr902

December 10, 2019

Contents

1 The Role of Independence

Let $X_1, ..., X_n$ be dependent random variables such that $X_i = X_1$ and $X_0 \in \{0,1\}$

$$E[X] = 0 * \frac{1}{2} + 1 * \frac{1}{2} \tag{1}$$

$$= \frac{1}{2} \tag{2}$$

So, there are only two cases,
Case 1 : all the X are 0

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} 0 \tag{3}$$

$$= 0 \tag{4}$$

then,

$$|E[X] - \frac{1}{n}\sum_{i=1}^{n} X_i| = |\frac{1}{2} - 0| \tag{5}$$

$$= \frac{1}{2} \tag{6}$$

Case 2 : all the X are 1

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} 1 \tag{7}$$
$$= 1 \tag{8}$$

then,

$$|E[X] - \frac{1}{n}\sum_{i=1}^{n} X_i| = |\frac{1}{2} - 1| \tag{9}$$
$$= \frac{1}{2} \tag{10}$$

So,

$$P(|E[X] - \frac{1}{n}\sum_{i=1}^{n} X_i| \geq \frac{1}{2}) = \frac{2}{2} \tag{11}$$
$$= 1 \tag{12}$$

2 How to Split a Sample into Training and Test Set

2.1

We have a fixed split of dataset S into S^{test} and S^{train}, where n^{test} is size of S^{test}.
We have trained a model $\hat{h}^*_{S^{train}}$
According to theorem 3.1 in lecture notes , Generalization bound for single hypothesis, we have

$$P\left(L(h) \geq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \leq \delta \tag{13}$$

Using the above stated theorem we can say

$$P\left(L(\hat{h}^*_{S^{train}}) \geq \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n^{test}}}\right) \leq \delta \tag{14}$$

By changing signs of the inequality, we have

$$P\left(L(\hat{h}^*_{S^{train}}) \leq \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n^{test}}}\right) \geq 1 - \delta \tag{15}$$

2.2

We consider m splits $\{(S_1^{test}, S_1^{train}), ..., (S_m^{test}, S_m^{train})\}$, where the size of the test sets are $n_1, ..., n_n$ and $h_1^*, ..., h_n^*$ are m prediction models where h_i^* is trained on S_i^{train}.
We need to show

$$P\left(\exists h_i^* \in H : L(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \leq \delta \tag{16}$$

we have,

$$P\left(\exists h_i^* \in H : L(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \tag{17}$$

taking union bound

$$P\left(\exists h_i^* \in H : L(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \leq \sum_{i=1}^{m} P\left(L(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \tag{18}$$

From comparing above equation with hoeffding inequality below,

$$P\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \leq \delta \tag{19}$$

we can say that

$$P\left(\exists h_i^* \in H : L(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \leq \sum_{i=1}^{m} P\left(L(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \tag{20}$$

$$\leq \sum_{i=1}^{m} \frac{\delta}{m} \tag{21}$$

$$= \delta \tag{22}$$

Hence, proved
By changing the sign of the above inequality, we get

$$P\left(\exists h_i^* \in H : L(\hat{h}_i^*) \leq \hat{L}(\hat{h}_i^*, S_i^{test}) + \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}}\right) \geq 1 - \delta \tag{23}$$

2.3

We can define the prior , $\pi(h) = \frac{1}{3i-1}$, where i is the index of the test set and split is done in a way such that as i increases the size of train set decreases.

3 Occam's Razor

3.1 a high-probability bound for L(h) that holds for all h $\in H_d$.

According to corollary 3.1 in lecture notes,

$$P\left(\exists h \in H : L(h) \geq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}\right) \leq \delta \tag{24}$$

where H is a finite set of hypothesis with class of size M.
Let d be the depth We know $|\sum|$ = 27,
$\sum^d$ be the space of the string of length d
H_d be the space op functions from $\sum^d$ to {0,1}
for string of length 1, we would have 27 choices and for length d we would have 27^d
And we have then we have two choices for languages, so in total we have M = 2^{27^d}

$$M = 2^{27^d} \tag{25}$$

$$\pi(h) = \frac{1}{M} \tag{26}$$

$$\pi(h) = \frac{1}{2^{27^d}} \tag{27}$$

By using above equation 23 and subsituting the value of M, we get

$$P\left(\exists h \in H_d : L(h) \geq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}}\right) \leq \delta \tag{28}$$

By changing the sign in the inequality, we get

$$P\left(\exists h \in H_d : L(h) \leq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}}\right) \geq 1 - \delta \tag{29}$$

3.2 a high-probability bound for L(h) that holds for all h $\in H$.

According to theorem 3.3 from lecture notes, we have

$$P\left(\exists h \in H : L(h) \geq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta \tag{30}$$

Let d(h) be the depth of the hypothesis h
We know $|\sum|$ = 27,
$\sum^d$ be the space of the string of length d
H_d be the space op functions from $\sum^d$ to {0,1}
$H = \bigcup_{i=0}^{\infty} H_d$
$|H_d| = 2^{27^d}$

$$\pi(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{27^{d(h)}}} \tag{31}$$

$\sum_{h \in H} \pi(h) \leq 1$
By subsituting the value of $\pi(h)$ in equation 30, we get

$$P\left(\exists h \in H : L(h) \geq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{2^{d(h)+1} 2^{27^{d(h)}}}{\delta}}{2n}}\right) \leq \delta \tag{32}$$

By changing the sign of the above inequality, we get

$$P\left(\exists h \in H : L(h) \leq \hat{L}(h,s) + \sqrt{\frac{\ln \frac{27^{d(h)+1} 2^{27^{d(h)}}}{\delta}}{2n}}\right) \geq 1 - \delta \tag{33}$$

3.3 Trade off between picking short strings and long strings

The L(h) depends on two terms - $\hat{L}(h,s)$ and $\sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}$
$\hat{L}(h,s)$ depends on the number of hypothesis, i.e, 2^{27^d} (calculated above), which decreases with increase in number of hypothesis. Thus decreases with the increase in d and favors large d.
$\sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}$ increases with the increase in d and favors small d.
So, there is a trade-off between picking small and large d

3.4

4 Kernels

4.1 Distance in feature space

To show :

$$||\Phi(x) - \Phi(z)|| = \sqrt{k(x,x) - 2k(x,z) + k(z,z)} \tag{34}$$

We have ,$||\Phi(x) - \Phi(z)||$

$$||\Phi(x) - \Phi(z)|| = \sqrt{\langle \Phi(x) - \Phi(z), \Phi(x) - \Phi(z) \rangle} \tag{35}$$
$$= \sqrt{\langle \Phi(x), \Phi(x) \rangle - 2\langle \Phi(x), \Phi(z) \rangle + \langle \Phi(z), \Phi(z) \rangle} \tag{36}$$

We know that k(x,y) = $\langle \Phi(x), \Phi(y) \rangle$. So,

$$||\Phi(x) - \Phi(z)|| = \sqrt{k(x,x) - 2k(x,z) + k(z,z)} \tag{37}$$

4.2 Sum of kernels

Let k_1, k_2 : X x X $\rightarrow$ R be positive-definite kernels.
To show : k(x,z) = $k_1(x,z) + k_2(x,z)$ is also positive-definite
For positive definite kernel K, $\forall c_1, ..., c_m \in$ R : $\sum_{i,j=1}^{m} c_i c_j K_{i,j} \geq 0$
$K_{i,j} = k(x_i, z_j)$
So, using above equation we can say

$$\sum_{i,j=1}^{m} c_i c_j K_1(x_i, z_j) \geq 0 \tag{38}$$
$$\sum_{i,j=1}^{m} c_i c_j K_2(x_i, z_j) \geq 0 \tag{39}$$

Adding above two equations, we get

$$\sum_{i,j=1}^{m} c_i c_j K_1(x_i, z_j) + \sum_{i,j=1}^{m} c_i c_j K_2(x_i, z_j) \geq 0 \tag{40}$$
$$\sum_{i,j=1}^{m} c_i c_j (K_1(x_i, z_j) + K_2(x_i, z_j)) \geq 0 \tag{41}$$
$$\sum_{i,j=1}^{m} c_i c_j K(x_i, z_j) \geq 0 \tag{42}$$

Hence, k(x,z) is positive-definite

4.3 Rank of Gram matrix

The Gram or kernel matrix of k with respect to $x_1, ..., x_m$ is the m x m matrix K with elements $K_{i,j} = k(x_i, x_j)$
We know that

$$k(x_i, x_j) = \langle \phi(x_i), \Phi(x_j) \rangle \tag{43}$$

here, Φ is identity, so

$$
\begin{aligned}
k(x_i, x_j) &= \langle x_i, x_j \rangle && (44) \\
&= x_i^T x_j && (45)
\end{aligned}
$$

$x_i \in R^d$ and there are m input patterns, and

$$K = X^T X \tag{46}$$

According to rank properties, rank of $X^T X$ is same as rank of X.

$$
\begin{aligned}
RankOf(K) &= RankOf(X^T X) && (47) \\
&= RankOf(X) && (48) \\
&\leq min(m, d) && (49)
\end{aligned}
$$

m and d are dimensions of X matrix