Aparna prasad + kfv856

December 10, 2019

Contents

1 The Role of Independence

let $x_1, ...x_n$ be a Bernoulli random variable with parameter $\frac{1}{2}$ and

$$x_1 = x_2 = x_3.... = x_n$$

where x_1, x_2, x_3 are Bernoulli random variable

$$\frac{1}{n}\sum_{i=1}^{n} = x_1 \in \{0, 1\}$$

$$P(x_i = 1) = p(x_i = 0) = p = \frac{1}{2}$$

where i=1,....n we have $\mu = E[x_i]$ where $E[x_i] = \frac{1}{2}$ so

$$\mu = E[x_i] = \frac{1}{2}$$

$$|\mu - \frac{1}{n}\sum_{i=1}^{n} x_i| \geq \frac{1}{2}$$

when$x_i = 0$ we have $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$ and when$x_i = 1$ we have $\frac{1}{n}\sum_{i=1}^{n} x_i = 1$
under the conditions above we have,

$$p\Big(|\mu - \frac{1}{n}\sum_{i=1}^{n} x_i| \geq \frac{1}{2}\Big) = 1$$

2 How to Split a Sample into Training and Test Set

1. we have a fixed split of S into S^{train} and S^{test} where size of $S^{test} = n^{test}$and we have a single hypothesis case here where $|H| = 1$
   based on theorem 3.1 from lecture notes we have
   From which we can derive the bound on $L(\hat{h}^* s_{train})$ in terms of $L(\hat{h}^* s_{train}, s^{test})$ and

n^{test} that holds with probability at least as $1-\delta$

$$L(\hat{h}^* s_{train}) \leq L(\hat{h}^* s_{train}, s^{test}) + \sqrt{\frac{ln(\frac{1}{\delta})}{2n^{test}}}$$

2. we train m prediction models where h_i^* is trained on s^{train} and calculate test loss on i-th model on i-th test set. H consists of a finite number of hypotheses M. where $|H| = M$

$$p\Big(\exists h \in H : \hat{L}(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, s_i^{test}) + \sqrt{\frac{ln(\frac{M}{\delta})}{2n_i}}\Big) \leq \delta$$

Below we have the where the first inequality is by the union bound and the second is by Hoeffding's inequality

$$p\Big(\exists h \in H : \hat{L}(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, s_i^{test}) + \sqrt{\frac{ln(\frac{M}{\delta})}{2n_i}}\Big) \leq \sum_{h\in H} p\Big(\hat{L}(\hat{h}_i^*) \geq \hat{L}(\hat{h}_i^*, s_i^{test}) + \sqrt{\frac{ln(\frac{M}{\delta})}{2n_i}}\Big)$$

$$\leq \sum_{h\in H} \frac{\delta}{M} = \delta$$

The bound on $\hat{L}(\hat{h}_i^*)$ in terms of $\hat{L}(\hat{h}_i^*, s_i^{test})$ with probability $1-\delta$ for all $h \in H$

$$\hat{L}(\hat{h}_i^*) \leq \hat{L}(\hat{h}_i^*, s_i^{test}) + \sqrt{\frac{ln(\frac{M}{\delta})}{2n_i}}$$

3. from theorem 3.3from leacture notes,let H be a countable hypothesis set, $\hat{L}(\hat{h}_i^*, s_i^{test})$ with probability $1-\delta$ for all $h \in H$

$$\hat{L}(\hat{h}_i^*) \leq \hat{L}(\hat{h}_i^*, s_i^{test}) + \sqrt{\frac{ln(\frac{1}{\pi(h)\delta})}{2n_i}}$$

where the prior$\pi(h)$ will be $\frac{1}{2^{d+1}}$ such that,

$$\hat{L}(\hat{h}_i^*) \leq \hat{L}(\hat{h}_i^*, s_i^{test}) + \sqrt{\frac{ln(\frac{1}{(\frac{1}{2^{d+1}})\delta})}{2n_i}}$$

since $\sum_0^\infty \frac{1}{2^{d+1}} = 1$ the assumption $\sum_h \pi(h) \leq 1$

3 Occam's Razor

1. Let $d \in \mathbb{N}_0$ and
   let $\Sigma_d =$ the space of strings of length d.

   and $\mathcal{H}_d$ = the space of functions from Σ_d from $\{0,1\}$, where Σ_d is the input string and $\{0,1\}$ is the prediction.

   Therefore, there are is the number of ways to choose d elements from Σ with replacement. so,

   $$|\Sigma_d| = |\Sigma|^d = 27^d \tag{1}$$

   There is a one-to-one correspondence between the elements of $\mathcal{H}_d$ and the power set $\mathcal{P}(\Sigma_d)$ of Σ_d, then

   $$|\mathcal{H}_d| = |\mathcal{P}(\Sigma_d)| = 2^{|\Sigma_d|} = 2^{27^d} \tag{2}$$

   Since $\mathcal{H}_d$ is finite, we can use **Theorem 3.5** to conclude that with probability $1-\delta$ for all $h \in \mathcal{H}_d$

   $$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{|\mathcal{H}_d|}{\delta}}{2n}} = \hat{L}(h,S) + \sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{3}$$

   where S is some labeled sample of strings from Σ_d, and $|S| = n$.

   the term $\sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}}$ grows exponentially as a function of d. Therefore, the d should be small or the sample size n should be large for a useful bound

2. . Let $d \in \mathbb{N}_0$ and
   let $\Sigma_d =$ the space of strings of length d.

   and $\mathcal{H}_d$ = the space of functions from Σ_d from $\{0,1\}$, where Σ_d is the input string and $\{0,1\}$ is the prediction.

   Therefore, we can use **Theorem 3.5** from lectures to conclude that for all $h \in \mathcal{H}$

   $$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{2^{d(h)+1}2^{27^{d(h)}}}{\delta}}{2n}} \tag{4}$$

   where $p : \mathcal{H} \to (0,1)$ is some function defined independently of S with $\sum_{h\in\mathcal{H}} p(h) \leq 1$.

   $$\pi(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{27^{d(h)}}} \tag{5}$$

   Since $\sum_{d=0}^{\infty} \frac{1}{2^{d+1}} = 1$, then $\sum_{d=0}^{\infty} \pi(h) \leq 1.1\, h \in \mathcal{H}$

3. if we look at the term $\sqrt{\frac{\ln \frac{2^{d(h)+1}2^{2^{d(h)}}}{\delta}}{2n}}$,we see that it grows exponentially as a function of $d(h)$. However,the term $\hat{L}(h,S)$ should decrease as a function of $d(h)$, to make its predictions,a h with a higher $d(h)$ uses longer strings.

4. From the informal lower bound we have expected loss to be $\frac{1}{4} = 0.25$. In the construction of the lower bound the size was 2^{2n}

   from the point 2 (i.e) equation (4) above we have a term that rises exponentially as a function of d(h)

4 Kernels

1. Distance in feature space
   Given a kernel k on input space h defining RKHS. let $\phi : x \to H$ denote the corresponding feature map. let $x, z \in X$, the distance of $\phi(x)$ and $\phi(z)$ is given by,

$$||\phi(x) - \phi(z)||^2 = (\phi(x) - \phi(z), \phi(x) - \phi(z)) \tag{6}$$

$$(\phi(x) - \phi(z), \phi(x) - \phi(z)) = (\phi(x), \phi(x)) + (\phi(z), \phi(z)) - 2(\phi(x), \phi(z)) \tag{7}$$

   applying (6) and (7) and Taking a square root,

$$||\phi(x) - \phi(z)|| = \sqrt{(\phi(x), \phi(x)) + (\phi(z), \phi(z)) - 2(\phi(x), \phi(z))}$$

   we know that.

$$k(x_1, x_2) = (\phi(x_1), \phi(x_2))$$

   so finally we get,

$$||\phi(x) - \phi(z)|| = \sqrt{k(x,z), k(z,z) - 2k(x,z)}$$

2. Sum of kernels

   let $k_1, k_2 : X \times X \to R$ be positive definite kernels.
   let $x_1, ...x_m$ be $\in X$ and A and B be the gram matrix ofk_1 and k_2 with respect to$x_1, ...x_m$.
   Since k_1 and k_2 are kernels, A and B are positive definite matrices.
   let $a_1, ...a_m \in R$ then

$$\sum_{i,j}^{m} a_i a_j A_{ij} \geq 0 \tag{8}$$

$$\sum_{i,j}^{m} a_i a_j B_{ij} \geq 0 \tag{9}$$

let us consider the function$k_3 : X \in X \to R$defined by,

$$k_3(x, y) = k_1(x, y) + k_2(x, y) \tag{10}$$

let C be the Gram matrix ofk_3 then,$C_{ij} = k_3(x_i, x_j)$

$$k_3(x_i, x_j) = k_1(x_i, x_j) + k_2(x_i, x_j) \tag{11}$$

$$C_{ij} = A_{ij} + B_{ij} \tag{12}$$

let $a_1, ..a_m \in R$

$$\sum_{i,j}^{m} a_i a_j C_{ij} = \sum_{i,j}^{m} a_i a_j (A_{ij} + B_{ij})$$

$$= \sum_{i,j}^{m} a_i a_j A_{ij} + \sum_{i,j}^{m} a_i a_j B_{ij} \geq 0$$

which means that C is also a positive definite. the Gram matrix of the function wrt to $x_1, ...x_m$ is alo s positive definite

3. Rank of a Gram Matrix

From the nullity rank theorem,matrices X with real elements let us consider $x \in N(A)$ where N(A) is a null space of matix , so

$$Ax = 0 \quad and \quad A^T Ax = 0$$

,such that $x \in N(A^T, A)$. therefore,

$$N(A^T, A) = N(A)$$

$$dim(N(A^T, A)) = dim(N(A))$$

from this we know that $rank(A^T A) = rank(A)$ so for matrix X, we we have,

$$rank(X) = rank(X^T X) \tag{13}$$

From the generel result ,in relation with the rank of Gram Matrices from a linear kernel,for $x, z \in \mathbb{R}^d$, on the input space $\mathbb{R}^d$.

$$k(x, z) = x^T z$$

Let $x_1, ..., x_m \in \mathbb{R}^d$. Construct the matrix X by letting the vector x_i by the i^{th} column of X.
By the definition of matrix multiplication, this means that for all $i, j \in 1, ..., m$

$$(X^T X)_{ij} = x_i^T x_j = k(x_i, x_j) \tag{14}$$

from the definition we know that the Gram matrix with respect to $x_1, ..., x_m$, $(X^T X)_{ij} = G_{ij}$ which means that $X^T X = G$ from (13), we have

$$rank(G) = rank(X^T X) = rank(X)$$

let X has $d \times m$ matrix, then let us consider rank of matrix X as rank(X),

$$rank(X) \leq \min(d, m) \tag{15}$$

so, for the kernel k as above on the input space $\mathbb{R}^d$, then for all $x_1, ..., x_m \in \mathbb{R}^d$, the rank of the Gram matrix G is bounded by

$$rank(G) \leq \min(d, m) \tag{16}$$

A square Gram matrix of G,$m \times m$ matrix, will have upper bound of m,

$$rank(G) \leq m \tag{17}$$