Probability Inequalities, Information Theory and Beyond

Hoang Nguyen, Huy Nguyen

March 19, 2020

In God we trust, all others must bring data.

W.Edwards Deming

1 Probability Inequalities

The Gaussian Tail Inquality

Let $X \sim N(0,1)$. Then

$$\mathbb{P}[|X| > \epsilon] \leqslant \frac{2e^{-\frac{\epsilon^2}{2}}}{\epsilon}$$

If $X_1, X_2, .., X_n \sim N(0,1)$ then

$$\mathbb{P}[|\bar{X_n}| > \epsilon] \leqslant \frac{2}{\sqrt{n}\epsilon} e^{-n\epsilon^2/2} \leqslant e^{-n\epsilon^2/2} \text{ (large n)}$$

Proof. The pdf of X is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Hence,

$$\mathbb{P}[X > \epsilon] = \int_\epsilon^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \leqslant \frac{1}{\epsilon} \int_\epsilon^\infty \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = -\frac{1}{\epsilon} \int_\epsilon^\infty f'(x) dx = \frac{f(\epsilon)}{\epsilon} \leqslant \frac{e^{-\frac{\epsilon^2}{2}}}{\epsilon}$$

By symmetry, we have $\mathbb{P}[|X| > \epsilon] \leqslant \frac{2e^{-\frac{\epsilon^2}{2}}}{\epsilon}$. Now let $X_1, X_2, .., X_n \sim N(0,1)$. Then $\bar{X_n} = \frac{1}{n}\sum_{i=1}^n X_i \sim N(0, \frac{1}{n})$. Hence, $\bar{X_n} = \frac{1}{\sqrt{n}} Z$ where $Z \sim N(0,1)$, then

$$\mathbb{P}[|\bar{X_n}| > \epsilon] = \mathbb{P}[|Z| > \sqrt{n}\epsilon] \leqslant \frac{2}{\sqrt{n}\epsilon} e^{-n\epsilon^2/2}$$

Markov's Inequality

Let X be a non-negative random variable and suppose that $\mathbb{E}[X]$ exists. For any $t > 0$,

$$\mathbb{P}[X > t] \leqslant \frac{\mathbb{E}[X]}{t}$$

Proof. Since $X > 0$,

$$\mathbb{P}[X > t] = \int_t^\infty f(x) dx \leqslant \frac{1}{t} \int_t^\infty x f(x) dx \leqslant \frac{1}{t} \int_0^\infty x f(x) dx = \frac{\mathbb{E}[X]}{t}$$

Chebyshev's Inequality

Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = Var(X)$. Then,

$$\mathbb{P}[|X - \mu| \geqslant t] \leqslant \frac{\sigma^2}{t^2}$$

Proof. Using Markov's inequality to show that,

$$\mathbb{P}[|X - \mu| \geqslant t] = \mathbb{P}[(X - \mu)^2 \geqslant t^2] \leqslant \frac{E[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

Hoeffding's Inequality

Hoeffding's Lemma: Suppose that $a \leqslant X \leqslant b$. Then

$$\mathbb{E}[e^{tX}] \leqslant e^{t\mu} e^{\frac{t^2(b-a)^2}{8}}$$

Where $\mu = \mathbb{E}[X]$.

Proof. Firstly, consider $\mu = 0$. Since $a \leqslant X \leqslant b$, we can write X as a convex combination of a and b, namely, $X = \alpha b + (1 - \alpha)a$ where $\alpha = \frac{X-a}{b-a}$. By the convexity of the function $y \to e^{ty}$, we have

$$e^{tX} \leqslant \alpha e^{tb} + (1 - \alpha)e^{ta} = \frac{X - a}{b - a} e^{tb} + \frac{b - X}{b - a} e^{ta}$$

Take expectations of both sides and use the fact that $\mathbb{E}[X] = \mu = 0$ to get

$$\mathbb{E}[e^{tX}] \leqslant -\frac{a}{b - a} e^{tb} + \frac{b}{b - a} e^{ta} = e^{g(u)}$$

where $u = t(b - a), g(u) = -\gamma u + log(1 - \gamma + \gamma e^u)$ and $\gamma = -\frac{a}{b-a}$. Because $g(0) = g'(0) = 0$ and $g''(u) \leqslant \frac{1}{4}$ for all $u > 0$. Using Taylor's theorem, there has to be a $\xi \in (0, u)$ such that

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2} g''(\xi) = \frac{u^2}{2} g''(\xi) \leqslant \frac{u^2}{8} = \frac{t^2(b - a)^2}{8}$$

Hence, $\mathbb{E}[e^{tX}] \leqslant e^{g(u)} \leqslant e^{\frac{t^2(b-a)^2}{8}}$.
If $\mu \neq 0$. Let $Z = X - \mu$, hence $\mathbb{E}[Z] = 0$. Using result above, we get

$$\mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t(X-\mu)}] \leqslant e^{\frac{t^2(b-a)^2}{8}} \Leftrightarrow \mathbb{E}[e^{tX}] \leqslant e^{t\mu} e^{\frac{t^2(b-a)^2}{8}}$$

Hoeffding's Inequality Let $X_1, X_2, .., X_n$ be iid observations such that $\mathbb{E}[X_i] = \mu$ and $a \leqslant X_i \leqslant b$. Then, for any $\epsilon > 0$,

$$\mathbb{P}[|\bar{X_n} - \mu| \geqslant \epsilon] \leqslant 2e^{-2n\epsilon^2/(b-a)^2}$$

Proof. Without loss of generality, we assume that $\mu = 0$. First we have

$$\mathbb{P}[|\bar{X_n}| \geqslant \epsilon] = \mathbb{P}[\bar{X_n} \geqslant \epsilon] + \mathbb{P}[-\bar{X_n} \geqslant \epsilon]$$

From Markov's inequality,

$$\mathbb{P}[\bar{X_n} \geqslant \epsilon] = \mathbb{P}[\sum_{i=1}^{n} X_i \geqslant n\epsilon] = \mathbb{P}[e^{\sum_{i=1}^n X_i} \geqslant e^{n\epsilon}] = \mathbb{P}[e^{t\sum_{i=1}^n X_i} \geqslant e^{tn\epsilon}]$$

$$\leqslant e^{-tn\epsilon} \mathbb{E}[e^{t\sum_{i=1}^n X_i}] = e^{-tn\epsilon} \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] = e^{-tn\epsilon} (\mathbb{E}[e^{tX_i}])^n$$

From Hoeffding's Lemma, $\mathbb{E}[e^{tX_i}] \leqslant e^{t^2(b-a)^2/8}$. So,

$$\mathbb{P}[\bar{X_n} \geqslant \epsilon] \leqslant e^{-tn\epsilon} e^{t^2 n(b-a)^2/8}$$

Consider function $f(t) = -tn\epsilon + t^2 n(b - a)^2/8$. This function is minimized at $t = \frac{4\epsilon}{(b-a)^2}$, giving

$$\mathbb{P}[\bar{X_n} \geqslant \epsilon] \leqslant e^{-2n\epsilon^2/(b-a)^2}$$

Similarly, applying the same argument to $\mathbb{P}[-\bar{X_n} \geqslant \epsilon]$ yields the result.

Mill's inequality

Let $Z \sim N(0,1)$. Then,

$$\mathbb{P}[|Z| > t] \leqslant \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}$$

Proof. We can obviously see $\mathbb{P}[|Z| > t] = 2\mathbb{P}[Z > t]$
Using the fact

$$\frac{1}{\sqrt{2\pi}} \int_t^\infty x e^{-\frac{x^2}{2}} \geq \frac{1}{\sqrt{2\pi}} t \int_t^\infty e^{-\frac{x^2}{2}} = t\mathbb{P}[Z > t]$$

We have

$$2\mathbb{P}[Z > t] \leq \frac{2}{t\sqrt{2\pi}} \int_t^\infty x e^{-\frac{x^2}{2}} = \frac{1}{t}\sqrt{\frac{2}{\pi}}(-e^{-\frac{x^2}{2}})|_t^\infty = \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2}}}{t}$$

Jensen's Inequality

If g is convex, then

$$\mathbb{E}[g(X)] \geqslant g(\mathbb{E}[X])$$

If g is concave, then

$$\mathbb{E}[g(X)] \leqslant g(\mathbb{E}[X])$$

Proof. Let $L(x) = ax + b$ be a line, tangent to $g(x)$ at the point $\mathbb{E}[X]$. Since g is convex, it lies above the line $L(x)$. So,

$$\mathbb{E}[g(X)] \geqslant \mathbb{E}[L(X)] = \mathbb{E}[ax + b] = a\mathbb{E}[X] + b = L(\mathbb{E}[X]) = g(E[X])$$

Cauchy-Schwart Inequality

If X and Y have finite variances then,

$$\mathbb{E}[|XY|] \leqslant \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Proof. This is directly implied from following inequality $|\sum_{i=1}^n u_i v_i|^2 \leqslant \sum_{j=1}^n |u_j|^2 \sum_{k=1}^n |v_k|^2$.

Loomis-Whitney Inequality

Let m be the measure of an open subset O of Euclidean $n-space$, and let $m1, .., m_n$ be the $(n-1)$-dimensional measures of the projections of O on the coordinate hyperplanes. Then

$$m^{n-1} \leqslant \prod_i^n m_i$$

Proof. We will see in Exercise section.

2 Information Theory and Entropy

Consider a set of of N identical objects that are to be divided amongst a set of bins, such that there are n_i objects in the i^{th} bin. It is clear that the total number of ways of allocating the N objects to the bins is

$$W = \frac{N!}{\prod_i} n_i$$

The *Entropy* is defined as

$$H = \frac{1}{N} lnW = \frac{1}{N} ln(N!) - \frac{1}{N} \sum_i ln(n_i!)$$

Now keep $\frac{n_i}{N}$ are held fixed, when $N \to \infty$, by Stirling's approximation

$$ln(N!) \approx NlnN - N$$

which gives

$$H = -lim_{N\to\infty} \sum_i (\frac{n_i}{N}) ln(\frac{n_i}{N}) = -\sum_i p_i ln(p_i)$$

where $\sum_i n_i = N$ and $p_i = lim_{N\to\infty}(\frac{n_i}{N})$ is the probability of an object being assigned to the i^{th} bin. We can interpret the bins as the states x_i of the discrete random variable X, where $\mathbb{P}[X = x_i] = p_i$, then the *entropy* of the random variable X is

$$H[p] = -\sum_i p(x_i) ln(p(x_i))$$

For the countinous random variable X with pdf $f(x)$, we have *entropy* of X is

$$H[x] = -\int f(x) ln(f(x)) dx$$

In Information theory, **Entropy** is a basic quantity in information theory associated to any random variable, which can be interpreted as the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes. For example, consider a random variable x having 8 possible states (000, 001, 010, 011, 100, 101, 110, 111), each of which is equally likely. In order to communicate the value of x to receiver, we need to $H[x] = -\sum_{i=1}^{8} \frac{1}{8} log_2 \frac{1}{8} = 3$ bits of length of transmitted message.
Next, we consider *joint entropy* to be a measure of the uncertainty associated with a set of variables,

$$H[X_1, ..., X_n] = \begin{cases} -\sum ... \sum P(x_1, ..., x_n) log P(x_1, ..., x_n) = \mathbb{E}[log \frac{1}{P(x_1,...,x_n)}] \text{ if X is dicrete} \\ -\int ... \int f(x_1, ..., x_n) log f(x_1, ..., x_n) dx_1 .. dx_n = \mathbb{E}[log \frac{1}{f(x_1,...,x_n)}] \text{ if X is continous} \end{cases}$$

and

$$H[X|Y] = \begin{cases} \mathbb{E}_{y\sim P_y}[H[P_{X|Y=y}]] = \mathbb{E}[log \frac{1}{P_{X|Y}}(X|Y)] \text{ if X is dicrete} \\ \mathbb{E}_{y\sim f(y)}[H[P_{X|Y=y}]] = \mathbb{E}[log \frac{1}{f_{X|Y}}(X|Y)] \text{ if X is continous} \end{cases}$$

i.e., the *conditional entropy* $H[X|Y]$ of X given Y. It is easily seen, using the product rule, that conditional entropy satisfies the relation

$$H[X, Y] = H[X|Y] + H[Y]$$

3 Inequalities in Entropy

Theorem.

$$H[X_1, ..., X_n] \leqslant \sum_{i=1}^{n} H[X_i]$$

Proof. This theorem is directly implied from definition of *joint entropy*.
Han's Inequality Let $X_1, ..., X_n$ be discrete random variables. Then

$$H[X_1, ..., X_n] \leqslant \frac{1}{n-1} \sum_{i=1}^{n} H[X_1, .., X_{i-1}, X_{i+1}, .., X_n]$$

Proof. Let $X_i^c = (X_1, .., X_{i-1}, X_{i+1}, .., X_n)$. We have

$$H[X_1, ..., X_n] = H[X_i|X_i^c] + H[X_i^c] \leqslant H[X_i|X_1, ..., X_{i-1}] + H[X_i^c]$$

Writing this inequality for each $i = 1, .., n$, we obtain

$$nH[X_1, ..., X_n] \leqslant \sum H[X_i^c] + \sum H[X_i|X_1, ..., X_{i-1}] = \sum H[X_i^c] + H[X_1, ..., X_n]$$

and subtract $H[X_1, ..., X_n]$ from both sides gives the result.

4 Exercises

1. Let $X_1, .., X_n \sim Bernoulli(p)$. Let $\alpha > 0$ be fixed and define

$$\epsilon_n = \sqrt{\frac{1}{2n} log(\frac{2}{\alpha})}$$

Let $\hat{p_n} = n^{-1} \sum_{i=1}^{n} X_i$. Define $C_n = (\hat{p_n} - \epsilon_n, \hat{p_n} + \epsilon_n)$. Show that

$$\mathbb{P}[p \in C_n] \geqslant 1 - \alpha$$

2. Let $X_1, .., X_n \sim N(0, 1)$. Bound $\mathbb{P}[|\bar{X_n}| > t]$ using Mill's inequality, where $\bar{X_n} = n^{-1} \sum_{i=1}^{n} X_i$. Compare to the Chebyshev bound.

3. The goal of binary classification problem is to build a rule to predict $Y \in \{0, 1\}$ given X using only the data at hand. Such a rule is a function $h : X \longrightarrow \{0, 1\}$ called a *classifier*. Some classifiers are better than others and we will favor ones that have low *classifier error*

$$R(h) = \mathbb{P}[h(X) \neq Y]$$

Next, given n observations $(X_1, Y_1), ..., (X_n, Y_n)$, let *emporical risk* $\hat{R}_n(.)$ defined for any classifier h by

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(X_i) \neq Y_i)$$

Where $\mathbb{I}(.)$ is *indicator function*. Consider a finite set of *classifier* $H = \{h_1, .., h_M\}$, we define the empirical risk minimizer $\hat{h}^{erm}$ by

$$\hat{h}^{erm} \in argmin_{h \in H} \hat{R}_n(h)$$

and the true risk is defined by

$$\bar{h} \in argmin_{h \in H} R_n(h)$$

Show that

$$R(\hat{h}) \leqslant R(\bar{h}) + \sqrt{\frac{2log(2M/\xi)}{n}}$$

with probability at least $1 - \xi$. In expectation, it holds that

$$\mathbb{E}[R(\hat{h})] \leqslant R(\bar{h}) + \sqrt{\frac{2log(2M)}{n}}$$

Remark: These inequalities above tell us that although $\hat{h}$ cannot hope to do better than $\bar{h}$ in general, the difference between the true one and the estimated one should not to be too large as long as the sample size is not too small compared to M.

4. Prove Loomis-Whitney's inequality.

5. (IMO 1992) Let S be a finite set of points in three-dimensional space. Let S_x, S_y, S_z be the sets consisting of the orthogonal projections of the points of S onto the yz-plane, zx-plane, xy-plane, respectively. Prove that

$$|S|^2 \leqslant |S_x||S_y||S_z|$$

where $|A|$ denotes the number of elements in the finite set A. (Note: The orthogonal projection of a point onto a plane is the foot of the perpendicular from that point to the plane.)

5 Solutions

1. We have $\mathbb{P}[C_n \text{ contains } p] = \mathbb{P}[\hat{p_n} - \epsilon_n \leq p \leq \hat{p_n} + \epsilon_n] = \mathbb{P}[|\hat{p_n} - p| \leq \epsilon_n] = 1 - \mathbb{P}[|\hat{p_n} - p| \geq \epsilon_n]$ Using Hoeffding's inequality, we have

$$1 - \mathbb{P}[|\hat{p_n} - p| \geq \epsilon_n] \geq 1 - 2e^{-2n\epsilon_n^2}$$

Set $2e^{-2n\epsilon_n^2} = \alpha$, we get $\epsilon_n = \sqrt{\frac{1}{2n} log(\frac{2}{\alpha})}$

2. We know by CTL, $\bar{X_n} \sim N(0, \frac{1}{n})$. Thus, using Mill's inequality we know

$$\mathbb{P}[|\bar{X_n}| > t] = \mathbb{P}[|Z| > nt] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{n^2t^2}{2}}}{nt}$$

Using Chebyshev's inequality

$$\mathbb{P}[|\bar{X_n}| > t] \leq \frac{\sigma^2}{t^2} = \frac{1}{n^2t^2}$$

By L'Hopital law we have

$$\frac{\frac{1}{n^2t^2}}{\sqrt{\frac{2}{\pi}} \frac{e^{-\frac{n^2t^2}{2}}}{nt}} = \frac{1}{\sqrt{\frac{2}{\pi}}} \frac{e^{\frac{n^2t^2}{2}}}{nt} \to \infty \text{ when n tends to } \infty$$

3. From the definition of $\hat{h}$, we have $\hat{R}_n(\hat{h}) \leqslant \hat{R}_n(\bar{h})$, which gives

$$R(\hat{h}) \leqslant R(\bar{h}) + [\hat{R}_n(\bar{h}) - R(\bar{h})] + [R(\hat{h}) - \hat{R}_n(\hat{h})]$$

On the other hand, by using Hoffding Theorem over all *classifier* h over H:

$$[\hat{R}_n(\bar{h}) - R(\bar{h})] + [R(\hat{h}) - \hat{R}_n(\hat{h})] \leqslant 2max_j|\hat{R}_n(h_j) - R(h_j)| \leqslant 2\sqrt{\frac{2log(2M/\xi)}{2n}}$$

with probability at least $1 - \xi$ which yields the result.
For the second part, let $\{Z_j\}_j$ be centered random variables, then

$$\mathbb{E}[max_j|Z_j|] = \frac{1}{s} log(e^{s\mathbb{E}[max_j|Z_j|]}) \leqslant \frac{1}{s} log\Big(\mathbb{E}[e^{s\ max_j|Z_j|}]\Big)$$

Where the last inequality comes from applying Jensen's inequality to the convex function $f(x) = e^x$, now bound the max by a sum, we get

$$\leqslant \frac{1}{s} log \sum_{j=1}^{2M} \mathbb{E}[e^{sZ_j}] \leqslant \frac{1}{s} log(2Me^{\frac{s^2}{8n}}) = \frac{log(2M)}{s} + \frac{s}{8n}$$

Where we used $Z_j = \hat{R}_n(h_j) - R(h_j)$, then applied Hoeffding's Lemma. Minimizing over s gives $s = 2\sqrt{2nlog(2M)}$ and plugging in produces

$$\mathbb{E}\Big[max_j|\hat{R}_n(h_j) - R(h_j)|\Big] \leqslant \sqrt{\frac{log(2M)}{2n}}$$

which yeilds the result. We have used the fact that applying Hoeffdings Theorem, for any classifier h, the bound (similarly to exercise 1)

$$|\hat{R}_n(h) - R(h)| \leqslant \sqrt{\frac{2/\xi}{2n}}$$

4. (Provided by Pro. Hung Q. Ngo) We pick a point $(X_1, ..., X_n)$ randomly from S containing m points on n-space with probability $\frac{1}{m}$. We have

$$H[X_1, ..., X_n] = log(m) \tag{1}$$

Next, from Han's inequality we have

$$(n-1)H[X_1, ..., X_n] \leqslant \sum_{i=1}^{n} H[X_1, .., X_{i-1}, X_{i+1}, .., X_n] \tag{2}$$

Because $H[X_1, .., X_{i-1}, X_{i+1}, .., X_n]$ is a random point on set S_i containing the projection of S on the hyperplanes, so it can not be greater thaan $log(m_i)$, hence

$$H[X_1, .., X_{i-1}, X_{i+1}, .., X_n] \leqslant log(m_i) \tag{3}$$

From $(3), (2)$ and (1) we can get Loomis-Whitney inequality.

5. Apply Loomis-Whitney inequality with $n = 3$ yields the result.

References

[1] All of statistic. Larry Wasserman, 2003.

[2] An inequality related to the isoperimetric inequality. L. H. Loomis and H. Whitney

[3] LECTURE NOTES ON INFORMATION THEORY. Yale University

[4] Tat ca la tai Entropy. NTZUNG.

[5] blogkhoahocmaytinh. Hung Ngo.

[6] Concentration Inequalities. Stanford Statistics 311.

[7] 33 rd International Mathematical Olympiad.

[8] Mathematics of Machine Learning. Prof. Philippe Rigollet

[9] Pattern Recognition and Machine Learning. Christopher M. Bishop.

[10] Three Proofs of the Sauer-Shelah Lemma. Hung Q. Ngo

[11] History of the Sauer-Shelah Lemma. Leon Bottou