Data Science. Group Assignment 2

Ekaterina Toropova, Marina Dushkina,Haijia Chen,
Tse-An Hsu, Siyuan Lai

November 2022

1 Corollary 3.7

Let (Ω, F, P) be a probability triple and let $X_1, X_2, \ldots, X_n \overset{IID}{\sim}$ F R-valued RVs such that $P(X_i \in [a,b]) = 1$, then for any $\epsilon > 0$ we get for $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$P(\bar{X_n} - \mathbb{E}[\bar{X_n}] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

furthermore

$$P(|\bar{X_n} - \mathbb{E}[\bar{X_n}]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

Proof:

$$P(\bar{X_n} - \mathbb{E}[\bar{X_n}] \leq -\epsilon) = P(\mathbb{E}[\bar{X_n}] - \bar{X_n} \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

Let $S_n = \sum_{i=1}^{n} X_i$. Let $s, t > 0$ be positive numbers to be chosen, then using Theorem 3.1 we get

$$P(\mathbb{E}[S_n] - S_n \geq t) = P(e^{s(\mathbb{E}[S_n]-S_n)} \geq e^{st}) \qquad (1.1)$$

$$\leq e^{-st}\,\mathbb{E}[e^{s(\mathbb{E}[S_n]-S_n}] \qquad (1.2)$$

$$= e^{-st}\Pi_{i=1}^{n}\,\mathbb{E}[e^{s(\mathbb{E}[X_i]-X_i)}] \qquad (1.3)$$

$$= e^{-st}\Pi_{i=1}^{n}\,\mathbb{E}[e^{-s(X_i-\mathbb{E}[X_i])}]$$

where in (1.3) we used the independence of $X_1, X_2, \ldots, X_n$ together with Theorem 2.45. Now using Lemma 3.5 with $\lambda = -s$ for each term in the product, we get

$$e^{-st}\Pi_{i=1}^{n}\,\mathbb{E}[e^{-s(X_i-\mathbb{E}[X_i])}] \leq e^{-st}e^{(-s)^2(b-a)^2n/8} = e^{-st}e^{s^2(b-a)^2n/8} \qquad (1.4)$$

Notice now, that the value s was arbitrarily chosen and we can choose it to make the right hand side as small as possible. That is we want to minimize

$$h(x) = s^2 \frac{(b-a)^2 n}{8}. \qquad (1.5)$$

This function is minimized at $s^* = \frac{4t}{n(b-a)^2}$, plugging that in we get

$$h(s^*) = s^2 \frac{(b-a)^2 n}{8} = -\frac{2t^2}{n(b-a)^2} \qquad (1.6)$$

Assembling (1.1) and (1.4)-(1.6) we get

$$P(\mathbb{E}[S_n] - S_n \geq t) \leq e^{-\frac{2t^2}{n(b-a)^2}}$$

Replacing $t = n\epsilon$, we get

$$P(\mathbb{E}[\bar{X_n}] - \bar{X_n} \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Mutiplying (-1) to $\mathbb{E}[\bar{X_n}] - \bar{X_n} \geq \epsilon$, we get

$$P(\bar{X_n} - \mathbb{E}[\bar{X_n}] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Combining with Theorem 3.6 $P(\bar{X_n} - \mathbb{E}[\bar{X_n}] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$, we get

$$P(\bar{X_n} - \mathbb{E}[\bar{X_n}] \leq -\epsilon) + P(\bar{X_n} - \mathbb{E}[\bar{X_n}] \geq \epsilon)$$

$$= P(|\bar{X_n} - \mathbb{E}[\bar{X_n}]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

which proves the corollary.

2 Lemma 3.15, properties 1-4

1. Let X be a sub-Gaussian RV with parameter λ, then αX is sub-Gaussian with parameter $|\alpha|\lambda$.

2. Let X be a sub-exponential RV with parameter λ, then αX is sub-exponential with parameter $|\alpha|\lambda$.

3. A sub-Gaussian RV X with parameter λ is sub-Exponential with parameter λ.

4. A bounded RV X, i.e. $P(X \in [a, b]) = 1$, then X is sub-Gaussian with parameter $(b-a)/2$. Specifically a Bernoulli RV is sub-Gaussian with parameter $1/2$.

Proof:

1. According to the definition of the sub-Gaussian random variable, X is a sub-Gaussian RV with parameter λ satisfies the inequality:

$$\mathbb{E}[\exp^{s(X-\mathbb{E}[X])}] \leq \exp^{\frac{s^2\lambda^2}{2}}, \forall s$$

Consider a random variable αX. Now write down this RV according to the definition of the sub-Gaussianity:

$$\mathbb{E}[\exp^{s(\alpha X-\mathbb{E}[\alpha X])}] = \mathbb{E}[\exp^{\alpha s(X-\mathbb{E}[X])}] \leq \exp^{\frac{\alpha^2 s^2\lambda^2}{2}} = \exp^{\frac{(|\alpha|s)^2\lambda^2}{2}}, \forall s$$

Thus αX is sub-Gaussian with parameter $|\alpha|\lambda$.

2. According to the definition of the sub-exponential random variable, X is a sub-exponential RV with parameter λ satisfies the inequality:

$$\mathbb{E}[\exp^{s(X-\mathbb{E}[X])}] \leq \exp^{\frac{s^2\lambda^2}{2}}, \forall |s| \leq \frac{1}{\lambda}$$

Consider a random variable αX. Now write down this RV according to the definition of the sub-exponentiality:

$$\mathbb{E}[\exp^{s(\alpha X-\mathbb{E}[\alpha X])}] = \mathbb{E}[\exp^{\alpha s(X-\mathbb{E}[X])}] \leq \exp^{\frac{\alpha^2 s^2\lambda^2}{2}} = \exp^{\frac{(|\alpha|s)^2\lambda^2}{2}}.$$

This inequality holds $\forall |\alpha|s \leq \frac{1}{\lambda}$ and thus, according to the definition of sub-Gaussian random variable, αX is sub-Gaussian with parameter $|\alpha|\lambda$.

3 According to the definition of the sub-Gaussian random variable, X is a sub-Gaussian RV with parameter λ satisfies the inequality:

$$\mathbb{E}[\exp^{s(X-\mathbb{E}[X])}] \leq \exp^{\frac{s^2\lambda^2}{2}}.$$

This inequality holds holds for all s and thus holds for the subset of all possible s including $|s| \leq \frac{1}{\lambda}$.
This satisfies the definition of the sub-exponential.

4. Consider Hoeffdings lemma (3.5) which states that if $P(X \in [a,b]) = 1$ then $\forall s \in R$:

$$\mathbb{E}[\exp^{s(X-\mathbb{E}[X])}] \leq \exp^{\frac{s^2(b-a)^2}{8}}$$

This inequality completely corresponds to the definition of the sub-Gaussian random variable where $\lambda = \frac{b-a}{2}$. If we have a Bernoully RV then $X \in [0,1]$ and thus $\lambda = \frac{1-0}{2} = \frac{1}{2}$.

3 Exercise 3.16

X has a Poisson distribution with parameter λ:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,\dots$$

Recall Def. 3.11 & 3.12, and that

$$E[X]=\sum_{k=0}^{+\infty}k\cdot P(X=k)=\lambda\cdot\sum_{k=0}^{+\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda}=\lambda$$

Then

$$\begin{aligned}
&E\left[e^{s(X-E[X])}\right]\\
=&E\left[e^{sX-sE[X]}\right]\\
=&E\left[e^{sX}\cdot e^{-s\lambda}\right]\\
=&e^{-s\lambda}\cdot E\left[e^{sX}\right]\\
=&e^{-s\lambda}\cdot e^{\lambda(e^s-1)}\\
=&e^{\lambda(e^s-s-1)}
\end{aligned}$$

The following explains it is sub-exponential, but not sub-Gaussian:

(1) Assume $\forall s\in\mathbb{R}$, $\exists\lambda_0$ such that

$$E\left[e^{s(X-E[X])}\right]\le e^{\frac{s^2\lambda_0^2}{2}}$$

That is,

$$\lambda(e^s-s-1)\le\frac{s^2\lambda_0^2}{2}$$

holds for $\forall s\in\mathbb{R}$. Since

$$\lim_{s\to+\infty}\frac{e^s-s-1}{s^2}=+\infty$$

it follows that no such λ_0, therefore it is not sub-Gaussian.

(2) For all $|s|\le\frac{1}{\lambda_0}$, there exists

$$\lambda(e^s-s-1)\le\frac{s^2\lambda_0^2}{2}$$

or equivalently,

$$\frac{2(e^s-s-1)}{s^2}\le\frac{\lambda_0^2}{\lambda} \tag{*}$$

(*) holds when $\forall |s| \leq \frac{1}{\lambda_0}, s \neq 0$. Since

$$\frac{2(e^s - s - 1)}{s^2} \longrightarrow 1 (s \to 0)$$

thus, it follows the definition of limit that $\exists \delta > 0$ such that $\frac{2(e^s-s-1)}{s^2} \leq 1 + 1 = 2$ when $|s| < \delta$.

Pick an arbitrary $\lambda_0 \geq \max\left\{\frac{1}{\lambda}, \sqrt{2\lambda}\right\}$, we obtain $|s| \leq \frac{1}{\lambda_0} < \delta$.

Therefore,

$$\frac{2(e^s - s - 1)}{s^2} \leq 2 \leq \frac{\lambda_0^2}{\lambda}$$

Thus, (*) is proved. That is, there exists parameter λ_0 which makes it sub-exponential.

4 Exercise 4.7

What is a reasonable statistical model for the Pattern Recognition problem?
A reasonable statistical model would be logistic regression model.
Suppose we want to model the binary outcome y= 0,1, the function is given as follow:

$$p(y = 0|x) = \frac{1}{1 + e^{\beta x}}$$

$$p(y = 1|x) = 1 - p(y = 0|x) = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

where p is the probability that y will either equal 0 or 1, $x \in R^n$ is the input data, $\beta \in R^n$ are estimated parameters.
The classification rule will be that if $p(y = 0|x) > p(y = 1|x)$,then y would be assigned to class 0,or if $p(y = 0|x) < p(y = 1|x)$, then y would be assigned to class 1.
The logistic regression model can also be generalized to multi-nominal logistic regression model,for $y \in R$ {1,2,...,K},the model becomes:

$$p(y = k|x) = \frac{e^{\beta_k x}}{1 + \sum_{k=2}^{K} e^{\beta_k x}}$$

The classification rule will be that y would be assigned to the class k with the highest probability $p(y = k|x)$.

5 Theorem 4.9

For any decision function $g(x)$ taking values in $\{0, 1\}$, we have

$$R(h*) \leq R(g)$$

Proof:(from the Lecture Notes)
First, we can rewrite

$$R(g) = \mathbb{E}[L\left(Y, g(X)\right)] = \mathbb{E}[\mathbb{E}[L\left(Y, g(X)\right) \mid X]].$$

That holds by the tower property of conditional expectation.

Part one:

Now we work with inner part.

$$\mathbb{E}[L\left(Y, g(X)\right) \mid X]$$

$$\underset{by(1)}{=} 1 - \mathbb{E}[\mathbb{1}_{y=g(x)} \mid X = x]$$

$$\underset{by(2)}{=} 1 - \mathbb{E}[\mathbb{1}_{1=g(x)}\mathbb{1}_{y=1} + \mathbb{1}_{1=g(x)}\mathbb{1}_{y=0} \mid X = x]$$

$$\underset{by(3)}{=} 1 - \mathbb{1}_{1=g(x)}\mathbb{E}[\mathbb{1}_{y=1} \mid X = x] - \mathbb{1}_{0=g(x)}\mathbb{E}[\mathbb{1}_{y=0} \mid X = x]$$

$$\underset{by(4)}{=} 1 - \mathbb{1}_{1=g(x)}r(x) - \mathbb{1}_{0=g(x)}(1 - r(x))$$

(1) Note that by definition of L and the property of indicator function (Exercise 2.8 (1))

$$L\left(y, g(x)\right) = \mathbb{1}_{y\neq g(x)} = 1 - \mathbb{1}_{y=g(x)}.$$

(2) Since X and $g(Y)$ only take values in $\{0, 1\}$ and $X = g(Y)$, either $X = g(Y) = 1$ or $X = g(Y) = 0$, and these two event can't happen simultaneously. By properties of indicator function (Exercise 2.8 (2) and (3)),

$$\mathbb{1}_{X=g(Y)} = \mathbb{1}_{X=g(Y)=1} + \mathbb{1}_{X=g(Y)=0} = \mathbb{1}_{1=g(x)}\mathbb{1}_{y=1} + \mathbb{1}_{1=g(x)}\mathbb{1}_{y=0}$$

(3) Linearity of mathematical expectation. For a fixed x, $\mathbb{1}_{g(x)=1}$ and $\mathbb{1}_{g(x)=0}$ are non-random variables (i. e. constants).

(4) By definition, $r(x) = \mathbb{E}[\mathbb{1}_{y=1} \mid X = x] = \mathbb{P}[y = 1|X = x] =$
$= 1 - \mathbb{P}[y = 0|X = x]$, hence, $\mathbb{P}[y = 0|X = x] = 1 - r(x)$

Part 2:

$$\mathbb{E}[L(Y, g(X)) \mid X = x] - \mathbb{E}[L(Y, h^*(X)) \mid X = x]$$

$$= [1 - \mathbb{1}_{1=g(x)}r(x) - \mathbb{1}_{0=g(x)}(1 - r(x))] - [1 - \mathbb{1}_{1=h^*(x)}r(x) - \mathbb{1}_{0=h^*(x)}(1 - r(x))]$$

$$= -\mathbb{1}_{1=g(x)}r(x) - \mathbb{1}_{0=g(x)}(1 - r(x)) + \mathbb{1}_{1=h^*(x)}r(x) + \mathbb{1}_{0=h^*(x)}(1 - r(x))$$

$$= r(x)(\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)}) - (1 - r(x))(\mathbb{1}_{0=h^*(x)} - \mathbb{1}_{0=g(x)})$$

$$= r(x)(\mathbb{1}_{1=h^*(x)} + \mathbb{1}_{1=g(x)}) - (1 - r(x))((1 - \mathbb{1}_{0=h^*(x)}) - (1 - \mathbb{1}_{0=g(x)}))$$

$$= r(x)(\mathbb{1}_{1=h^*(x)} + \mathbb{1}_{1=g(x)}) - (1 - r(x))(\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)})$$

$$= (2r(x) - 1)(\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)})$$

If $2r(x) - 1 > 0$,
then $r(x) > \frac{1}{2}$ and $h^*(x) = 1$, so $\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)} \geq 0$.
In this case $(2r(x) - 1)(\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)}) \geq 0$.

If $2r(x) - 1 \leq 0$,
then $r(x) \leq \frac{1}{2}$ and $h^*(x) = 0$, so $\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)} \leq 0$.
Again $(2r(x) - 1)(\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)}) \geq 0$.

It follows that $(2r(x) - 1)(\mathbb{1}_{1=h^*(x)} - \mathbb{1}_{1=g(x)}) \geq 0$ for all x, so

$$\mathbb{E}[L(Y, g(X)) \mid X = x] - \mathbb{E}[L(Y, h^*(X)) \mid X = x] \geq 0$$

for all x, then

$$\mathbb{E}[\mathbb{E}[L(Y, g(X)) - L(Y, h^*(X)) \mid X]] \geq 0.$$

$$\mathbb{E}[L(Y, g(X))] - \mathbb{E}[L(Y, h^*(X))] \geq 0$$

$$\mathbb{E}[L(Y, g(X))] \geq \mathbb{E}[L(Y, h^*(X))]$$