Group Assignment 2

Introduction to Data Science H23

Elise Hammarström Theodora Moldovan Ella Schmidtobreick
Georgios Tsouderos Finn Vaughankraska

December 13, 2023

All group members attempted the proofs/exercises individually before meeting. After discussing, we finalized the problems and each of us chose a problem to write up in LaTeX.

1 Proof Corollary 3.7

Corollary 3.7. Let $(\Omega, \mathcal{F}, P)$ be a probability triple and let $X_1, \dots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a,b]) = 1$. Then for any $\epsilon > 0$, we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{\frac{-2n\epsilon^2}{(b-a)^2}},$$

furthermore

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}.$$

Proof. Let $S_n = \sum_{i=1}^n X_i$. Let $s, t > 0$ be positive numbers to be chosen, then using Theorem 3.1 (Markov's inequality), we get:

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &= \mathbb{P}(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st}) && (1)\\
&\leq e^{-st}\mathbb{E}[e^{s(S_n - \mathbb{E}[S_n])}] && (2)\\
&\overset{\text{i.i.d.}}{=} e^{-st}\prod_{i=1}^n \mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}]. && (3)\\
&\leq e^{-st}\prod_{i=1}^n e^{\frac{s^2(b_i-a_i)^2 n}{8}} && (4)\\
&= e^{-st+\sum_{i=1}^n \frac{s^2(b_i-a_i)^2 n}{8}} && (5)
\end{aligned}$$

where in equation 3 we used the independence of X_i. In equation 4 we used Hoeffding's Lemma (with $\lambda = s$ for each term in the product) and from that we arrive at equation 5 which is quadratic in s, for which we wish to find the minimum.
Notice that in Hoeffding's theorem, 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

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

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

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

Assembling all equations we get

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

Replacing $t = n\epsilon$ we get

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

Similarly, we can prove

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{\frac{-2n\epsilon^2}{(b-a)^2}},$$

by rearranging the inequality in the probability:

$$\begin{aligned}
&\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) && (6)\\
&= \mathbb{P}(-(\overline{X}_n - \mathbb{E}[\overline{X}_n]) \geq \epsilon) && (7)\\
&= \mathbb{P}(e^{s(S_n - \mathbb{E}(S_n)} \geq e^{st}) && (8)\\
&\leq e^{-st}\mathbb{E}(e^{s(S_n - \mathbb{E}[S_n]}) && (9)\\
& && (10)
\end{aligned}$$

In equation 7, assuming $\overline{X}_n<0 \Rightarrow\ \mathbb{E}(\overline{X}_n) < 0$ gives the equivalence to equation 8.
Now we have arrived at the same step as in equation 2 above. Developing the R.H.S. in the same way, we arrive at the equation to be proven,

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

We combine the two inequalities such that Hoeffding's inequality works as an upper and lower bound at the same time. The steps can be followed from the equation transformation 6:

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) = \mathbb{P}(-\epsilon \geq \overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) = 2 * \mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$$

□

2 Proof Lemma 3.15, properties 1-4

Lemma 1. *The following properties hold:*

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. $\mathbb{P}(X \in [a, b]) = 1$, then X is sub-Gaussian with parameter $(b - a)/2$. Specifically a Bernoulli RV is sub-Gaussian with parameter $1/2$.*

Proof. Proof of Property 1: Given that X is sub-Gaussian RV with parameter λ, by definition 3.11 the following holds:

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

For αX we get:

$$\mathbb{E}[e^{s(\alpha X-\mathbb{E}[\alpha X])}] = \mathbb{E}[e^{s(\alpha X-\alpha\mathbb{E}[X])}] = \mathbb{E}[e^{s\alpha(X-\mathbb{E}[X])}] \leq e^{\frac{s^2\alpha^2\lambda^2}{2}} = e^{\frac{s^2(\lambda')^2}{2}}, \quad \text{where } \lambda' = |\alpha|\lambda \text{ for } \lambda > 0.$$

The parameter $|\alpha|$ is absolute as it has to be positive. This shows that αX is sub-Gaussian with parameter $|\alpha|\lambda$.

Proof of Property 2: Given that X is sub-exponential RV with parameter λ, by definition 3.12, the following holds:

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

For αX we get:

$$\mathbb{E}[e^{s(\alpha X-\mathbb{E}[\alpha X])}] = \mathbb{E}[e^{s(\alpha X-\alpha\mathbb{E}[X])}] = \mathbb{E}[e^{s\alpha(X-\mathbb{E}[X])}] \leq e^{\frac{s^2\alpha^2\lambda^2}{2}} = e^{\frac{s^2(|\alpha|\lambda)^2}{2}}, \quad \forall |s| \leq \frac{1}{|\alpha|\lambda}.$$

The parameter $|\alpha|$ is absolute as it has to be positive. Thus, αX is sub-exponential with parameter $|\alpha|\lambda$.

Proof of Property 3: Let X be a sub-Gaussian random variable with parameter λ. By definition 3.11 the following holds:

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

Therefore the following holds immediately,

$$\mathbb{E}[e^{s(X-E(X))}] \leq \exp(\frac{s^2\lambda^2}{2}), \quad \forall |s| \leq \frac{1}{\lambda} \tag{11}$$

since $\{s \,|\, |s| \leq \lambda\} \subset \mathbb{R}$. Equation 11 is the definition for a variable X being sub-exponential. Therefore, the given property is proven.

Proof of Property 4: A $\mathbb{R}$-valued random variable X is said to be sub-Gaussian with parameter λ if

$$\mathbb{E}\left[e^{s(X-\mathbb{E}[X])}\right] \leq e^{\frac{s^2\lambda^2}{2}}, \quad \forall s \in \mathbb{R}.$$

Take

$$\lambda = \frac{b-a}{2}.$$

Then

$$e^{\frac{s^2\lambda^2}{2}} = e^{\frac{s^2\left(\frac{b-a}{2}\right)^2}{2}} = e^{\frac{s^2(b-a)^2}{8}}.$$

Lemma 3.5 says that for a bounded RV X with $\mathbb{P}(X \in [a,b]) = 1$,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Hence, X is sub-Gaussian with parameter $\frac{b-a}{2}$.
Specifically, if X is a Bernoulli RV, then

$$e^{\frac{s^2\lambda^2}{2}} = e^{\frac{s^2\left(\frac{1}{2}\right)^2}{2}} = e^{\frac{s^2(1)^2}{8}}.$$

thus X is sub-Gaussian with parameter $1/2$.

□

3 Solution Exercise 3.16

Exercise 3.16. For the Poisson distribution, we have

$$\mathbb{E}[e^{sX}] = e^{\lambda(e^s-1)}$$

is this sub-Gaussian, sub-exponential, or neither?

Solution
Recall the definition of sub-Gaussian and sub-Exponential.
Sub-Gaussian function with parameter $\lambda > 0$:

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

Sub-Exponential function with parameter $\lambda > 0$:

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

Using the Poisson distribution $\mathbb{E}[e^{sX}] = e^{\mu(e^s-1)}$, the following holds:

$$\begin{aligned}
\mathbb{E}[e^{s(X-\mathbb{E}[X])}] &\leq e^{\frac{s^2\lambda^2}{2}} \\
\mathbb{E}\left[e^{s(X-\mu)}\right] &\leq e^{\frac{s^2\lambda^2}{2}} \\
\mathbb{E}\left[\frac{e^{sX}}{e^{s\mu}}\right] &\leq e^{\frac{s^2\lambda^2}{2}} \\
e^{-\mu s}\mathbb{E}\left[e^{sX}\right] &\leq e^{\frac{s^2\lambda^2}{2}} \quad \text{then} \\
\frac{e^{\mu(e^2-1)}}{e^{s\mu}} &\leq e^{\frac{s^2\lambda^2}{2}} \\
e^{\mu(e^s-1)-s\mu} &\leq e^{\frac{s^2\lambda^2}{2}}
\end{aligned}$$

Since the exponential function is a monotonically strictly growing function, we can only look at the exponents on both sides of the inequality:

$$\begin{aligned}
\mu e^s - \mu - s\mu &\leq \frac{s^2\lambda^2}{2} \\
\mu(e^s - s - 1) &\leq \frac{s^2\lambda^2}{2}
\end{aligned}$$

Substitute $s = \frac{1}{\lambda}$, since $|s| \leq \frac{1}{\lambda}$:

$$\begin{aligned}
\mu(e^{\frac{1}{\lambda}} - \frac{1}{\lambda} - 1) &\leq \frac{\lambda^2}{2\lambda^2} \\
\mu(e^{\frac{1}{\lambda}} - \frac{1}{\lambda} - 1) &\leq \frac{1}{2} \\
\text{Now take limit of} \quad \lambda &\to \infty \\
\lim_{\lambda \to \infty} \mu(e^{\frac{1}{\lambda}} - \frac{1}{\lambda} - 1) &\to +0 \leq \frac{1}{2}
\end{aligned}$$

If you choose λ big enough, the above property holds. Therefore, there exists a parameter λ such that the Poisson distribution is sub-exponential.

To evaluate if the Poisson distribution is sub-Gaussian, we have a look at the exponents once again:

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

Since for sub-Gaussian the above equation needs to hold for all $s \in \mathbb{R}$, let us take the limit of $s \to \infty$. By using L'Hospital you can evaluate, that the left-hand side grows faster than the right-hand side. Therefore, the inequality does not hold for all s such that the Poission distribution is not sub-Gaussian.

Therefore the answer to the exercise is clearly that it is sub-Exponential but not sub-Gaussian.

4 Solution Exercise 4.7

Exercise 4.7. What is a reasonable statistical model for the Pattern Recognition problem?
The classification problem is in modern times very often associated with the prototypical example of classification of images of dogs and cats. In that example, the images is X and the class is given by Y. The risk above has a natural interpretation, given the "decision rule" g_λ, the risk $R(\lambda)$ is the probability of an incorrect classification by the rule g_λ,

$$E\left[L(Y, g_\lambda(X))\right] = \mathbb{P}\{Y \neq g_\lambda(X)\}.$$

Solution:
A reasonable statistical model for the Pattern Recognition (classification) problem is Naive Bayes. Recall Bayes' theorem as:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Where X is a vector of independent features and Y is the class of the variable of interest. The Naive Bayes model makes an important assumption that the features that make up X are independent. This allows the above equation to be rewritten as:

$$P(Y = k|X_1, X_2, \ldots, X_n) = \frac{P(Y = k) \cdot P(X_1|Y = k) \cdot P(X_2|Y = k) \cdot \ldots \cdot P(X_n|Y = k)}{P(X_1) \cdot P(X_2) \cdot \ldots \cdot P(X_n)}$$

$$= \frac{P(Y = k) \prod_{i=1}^{n} P(X_i|Y = k)}{P(X_1) \cdot P(X_2) \cdot \ldots \cdot P(X_n)}$$

The decision rule is then:

$$Class = \arg\max_k \left(\frac{P(Y = k) \prod_{i=1}^{n} P(X_i|Y = k)}{P(X_1) \cdot P(X_2) \cdot \ldots \cdot P(X_n)} \right)$$

Which can be interpreted as the class k that has the maximum likelihood given input vector X.

In the context of cat and dog photos, our training data might include many many image entities with X pixels labeled as classes: cat, dog, and neither. Obviously this application of the model would fail due to auto-correlation between pixels (covariances). More generally, with any classification model we are assuming that the underlying true process F_{xy} is categorical. It follows that the Naive Bayes Model also yields categorical outputs from its decision rule.

5 Proof Theorem 4.9 with all details, basically referring to all the properties of the indicator function used, the monotonicity of measures etc.

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

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

Proof. Note that we can write the following by using the power property for the second equal sign:

$$R(g) \stackrel{\text{def}}{=} \mathbb{E}[L(Y, g(X))] = \mathbb{E}[\mathbb{E}[L(Y, g(X))|X]]$$

We will now only work with the inner part:

$$\begin{aligned}
\mathbb{E}[L(Y, g(X))|X = x] &= 1 - \mathbb{E}[\mathbb{1}_{\{y=g(x)\}}|X = x] && (12)\\
&= 1 - \mathbb{E}[\mathbb{1}_{\{g(x)=1\}}\mathbb{1}_{\{y=1\}} + \mathbb{1}_{\{0=g(x)\}}\mathbb{1}_{\{y=0\}}|X = x] && (13)\\
&= 1 - \mathbb{1}_{\{1=g(x)\}}\mathbb{E}[\mathbb{1}_{\{y=1\}}|X = x] - \mathbb{1}_{\{0=g(x)\}}\mathbb{E}[\mathbb{1}_{\{y=0\}}|X = x] && (14)\\
&= 1 - \mathbb{1}_{\{1=g(x)\}}r(x) - \mathbb{1}_{\{0=g(x)\}}(1 - r(x)) && (15)
\end{aligned}$$

Let us explain the above equations in more detail. Equation 12 holds by the following property:

$$\mathbb{E}[L(Y, g(X))|X = x] = \mathbb{P}(\{Y \neq g(x)\}) \overset{Y\in\{0,1\}}{=} 1 - \mathbb{P}(\{Y = g(x)\}) = 1 - \mathbb{E}[\mathbb{1}_{\{y=g(x)\}}|X = x]$$

Equation 13 follows from Lemma 2.8(2) from the lecture notes as well as the definition of $Y \in \{0, 1\}$. In equation 13 the indicator function referring to $g(x)$ does not depend on y and therefore can be treated as a scalar in the expectation. From this property, equation 14 follows. Equation 15 proceeds from the definition of $r(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x\}$.
The transformation above is used to prove the property:

$$\begin{aligned}
&\mathbb{E}[L(Y, g(X))|X = x] - \mathbb{E}[L(Y, h^*(X))|X = x] \\
&= [1 - \mathbb{1}_{\{1=g(x)\}}r(x) - \mathbb{1}_{\{0=g(x)\}}(1 - r(x))] - [1 - \mathbb{1}_{\{1=h^*(x)\}}r(x) - \mathbb{1}_{\{0=h^*(x)\}}(1 - r(x))] \\
&= 1 - 1 + \mathbb{1}_{\{1=h^*(x)\}}r(x) - \mathbb{1}_{\{1=g(x)\}}r(x) + \mathbb{1}_{\{0=h^*(x)\}}(1 - r(x)) - \mathbb{1}_{\{0=g(x)\}}(1 - r(x)) \\
&= r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) + (1 - r(x))(\mathbb{1}_{\{0=h^*(x)\}} - \mathbb{1}_{\{0=g(x)\}}) \\
&= r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) + (1 - r(x))[(1 - \mathbb{1}_{\{1=h^*(x)\}}) - (1 - \mathbb{1}_{\{1=g(x)\}})] \qquad (16) \\
&= r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) + (1 - r(x))[(-1) \cdot (\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}})] \\
&= r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) - (1 - r(x))(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) \\
&= (2r(x) - 1)(\mathbb{1}_{\{h^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}) \geq 0 \qquad (17)
\end{aligned}$$

Equation 16 follows from Lemma 2.8(1). Inequality 17 follows by looking at the following two cases with $g(x) \in \{0, 1\}$ as given:

$$\text{Let } r(x) \leq \frac{1}{2}. \text{ By definition } h^*(x) = 0 \text{ follows.} \quad \Rightarrow \quad \underbrace{(2r(x) - 1)}_{\leq 0} \underbrace{(\mathbb{1}_{\{h^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}})}_{\in \{-1,0\}} \geq 0$$

$$\text{Let } r(x) > \frac{1}{2}. \text{ By definition } h^*(x) = 1 \text{ follows.} \quad \Rightarrow \quad \underbrace{(2r(x) - 1)}_{> 0} \underbrace{(\mathbb{1}_{\{h^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}})}_{\in \{0,1\}} \geq 0$$

By the definition of $R(g)$ and $R(h^*)$, the inequality immediately implies the statement of the theorem. □