ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Foundations of Data Science	Assignment date: Thursday, November 14th, 2024, 17:15
Fall 2024	Due date: Thursday, November 14th, 2024, 18:45

Midterm Exam – INF1

This exam is open book. No electronic devices of any kind are allowed. There are three problems. Good luck!

Only answers given on this handout count.

Name: ____________________________

SCIPER: ____________________________

Problem 1	/ 10
Problem 2	/ 8
Problem 3	/ 10
Total	/28

Problem 1 (Gaussian bandits with unknown mean and variance — 10 pts)**.** Consider the standard setup of the bandit problem we discussed in the course. We assumed that the K arms have unknown means and are all 1-subgaussian. For the *upper confidence bound* (UCB) algorithm, after sampling each arm once in the beginning, in each subsequent round we decided on the arm according to

$$\text{argmax}_k \quad \underbrace{\hat{\mu}_{k,t-1} + \sqrt{\frac{2}{T_k(t-1)} \ln \frac{1}{\delta_t}}}_{\text{decision metric}} . \tag{1}$$

Here, $T_k(t-1)$ denotes the number of times we chose arm k up to and including time $t-1$, $\hat{\mu}_{k,t-1}$ is the empirical mean of arm k at time $t-1$, using the relevant $T_k(t-1)$ samples, and δ_t denotes the confidence we want to have at time t.

a) [2 pts] Write down the decision metric if arm k, $1 \leq k \leq K$, is σ_k^2-subgaussian and the parameters $\{\sigma_k^2\}$ are known?

In general the parameters $\{\sigma_k^2\}_{k=1}^K$ are unknown. However, we can estimate their values from the samples. More precisely, if we are given iid samples $W_1, \cdots, W_n$ let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n W_n$ and let $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (W_i - \hat{\mu}_n)^2$ be the sample mean and the sample variance, respectively. It is then tempting to conjecture that we can construct an UCB algorithm by replacing the variance terms σ_k^2 in point (a) above with their empirical values $\hat{\sigma}_{k,t-1}^2$ and possibly change the involved constants.

We will now confirm this conjecture for the concrete case where all arms have a Gaussian distribution. More precisely, we assume that arm k, $1 \leq k \leq K$, is distributed according to $\mathcal{N}(\mu_k, \sigma_k^2)$. The parameters $\{(\mu_k, \sigma_k^2)\}_{k=1}^K$ are unknown.

b) [5 pts] Write down the decision metric for this case.

HINT: If we are given iid samples $W_1, \cdots, W_n$ from a Gaussian distribution with parameters (μ, σ^2), and $(\hat{\mu}_n, \hat{\sigma}_n^2)$ are the respective empirical quantities, then

$$S_n = \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n / \sqrt{n}}$$

is distributed according to the so-called student t-distribution with $n-1$ degrees of freedom. Note that the distribution of S_n does not depend on (μ, σ^2) and is symmetric around 0. Further, the confidence values for this distribution can be looked up in tables or can be computed. I.e., you can asssume that for any $\delta \in [0, 1]$ the real numbers $\alpha_{\delta,n}$, so that

$$\mathrm{P}\{S_n > \alpha_{\delta,n}\} = \delta$$

are known.

c) [3 pts] Show that the distribution of S_n indeed does not depend on (μ, σ^2).

HINT: The distribution of $(\hat{\mu}_n - \mu)$ is Gaussian with mean zero and variance σ^2/n. Further, $(n-1)\hat{\sigma}_n^2/\sigma^2$ follows the so-called χ^2 distribution with $n-1$ degrees of freedom. Note that the χ^2 distribution does not depend on (μ, σ^2). Further, $(\hat{\mu}_n - \mu)$ and $(n-1)\hat{\sigma}_n^2/\sigma^2$ are independent.

Solution 1.
a) In this case we would decide on the next arm according to

$$\text{argmax}_k \quad \underbrace{\hat{\mu}_{k,t-1} + \sigma_k \sqrt{\frac{2}{T_k(t-1)} \ln \frac{1}{\delta_t}}}_{\text{decision metric}}.$$

b) For a given value $\delta > 0$ (the confidence value) and $n \in \mathbb{N}$ (the degrees of freedom), let $\alpha_{\delta,n}$ be the value so that $\mathrm{P}\{S_n > \alpha_{\delta,n}\} = \delta$ according to the student t-distribution with n degrees of freedom. This value can be looked up in a table or computed numerically.

According to the hint, $\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}}$ has a student t-distribution with $n-1$ degrees of freedom. Hence

$$\mathrm{P}\{\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} > \alpha_{\delta,n-1}\} = \delta.$$

By the symmetry of this distribution we have

$$\mathrm{P}\{\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} \geq -\alpha_{\delta,n-1}\} = 1 - \delta.$$

This can be rewritten as

$$\mathrm{P}\{\mu \leq \hat{\mu}_n + \frac{\alpha_{\delta,n-1}\hat{\sigma}_n}{\sqrt{n}}\} = 1 - \delta.$$

In turn this is equivalent to

$$\mathrm{P}\{\mu > \hat{\mu}_n + \frac{\alpha_{\delta,n-1}\hat{\sigma}_n}{\sqrt{n}}\} = \delta.$$

Hence we see that the decision should be taken according to

$$\text{argmax}_k \quad \hat{\mu}_{k,t-1} + \frac{\alpha_{\delta,T_k(t-1)-1}\hat{\sigma}_{k,t-1}}{\sqrt{T_k(t-1)}}.$$

This is of the same form as our original decision metric, just with a different constant and we swapped out the true standard deviaiton for the empirical standard deviation.

c) We have

$$S_n = \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} = \frac{(\hat{\mu}_n - \mu)/\sigma}{\hat{\sigma}_n/(\sigma\sqrt{n})}.$$

Now note that by the hint, the numerator and the denominator are independent random variables and that the distributions of those random variables do not depend on (μ, σ^2). Hence, S_n has a distribution that does not depend on (μ, σ^2).

Problem 2 (KL Divergence between mixtures — 8 pts)**.** Mixture distributions are a key modeling tool and appear in many guises in Data Science. In this problem, we derive a bound on the KL divergence between mixture distributions.

a) [5 pts] Consider two mixture distributions of K components, given as

$$P_Y(y) = \sum_{i=1}^{K} \mu_i P_i(y) \quad \text{and} \quad Q_Y(y) = \sum_{i=1}^{K} \nu_i Q_i(y), \tag{2}$$

where $0 \leq \mu_i \leq 1$ and $0 \leq \nu_i \leq 1$ and $\sum_i \mu_i = \sum_i \nu_i = 1$. Here, $P_i(y)$ and $Q_i(y)$ are distributions over an alphabet $\mathcal{Y}$. Prove that

$$D(P_Y \| Q_Y) \leq D(\mu \| \nu) + \sum_{i=1}^{K} \mu_i D(P_i \| Q_i), \tag{3}$$

where $D(\mu\|\nu)$ denotes the KL divergence between the distributions $(\mu_1, \mu_2, \ldots, \mu_K)$ and $(\nu_1, \nu_2, \ldots, \nu_K)$. *Hint: Recall conditional KL divergence. Also, it may be helpful to introduce a random variable X distributed over the set $\mathcal{X} = \{1, 2, \ldots, K\}$ and rewrite $P_Y(y)$ in the form $P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x)$, for appropriately chosen $P_X(x)$ and $P_{Y|X}(y|x)$.*

b) [3 pts] Give examples where the bound is good and where the bound is bad. The more extreme your examples, the more points you get. The less trivial your examples, the more points you get. *Hint: Try $K = 2$.*

Solution 2. (a) To connect to the class, let us change notation. Specifically, let us introduce a random variable X distributed over the alphabet $\{1, 2, \ldots, K\}$. Define

$$P_X(x = i) = \mu_i \quad \text{and} \quad P_{Y|X}(y|x = i) = P_i(y) \tag{4}$$

and then, of course, $P_{X,Y}(x, y) = P_X(x) P_{Y|X}(y|x)$. The resulting marginal distribution of Y is then

$$P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x) = \sum_{i=1}^{K} \mu_i P_i(y), \tag{5}$$

exactly as in the problem statement. By the same token, define

$$Q(x = i) = \nu_i \quad \text{and} \quad Q(y|x = i) = Q_i(y) \tag{6}$$

and $Q_{X,Y}(x, y) = Q_X(x) Q_{Y|X}(y|x)$. The resulting marginal distribution of Y is then

$$Q_Y(y) = \sum_{x \in \mathcal{X}} Q_X(x) Q_{Y|X}(y|x) = \sum_{i=1}^{K} \nu_i Q_i(y), \tag{7}$$

exactly as in the problem statement.

Next, let us rewrite the claim that needs to be proved in terms of the new notation. To this end, we recall Homework 2, Problem 3, which introduced Conditional KL Divergence. From that homework problem, we can write:

$$\begin{aligned} D(P_{Y|X}\|Q_{Y|X}|P_X) &= \sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x)) && (8) \\ &= \sum_{i=1}^{K} \mu_i D(P_i\|Q_i). && (9) \end{aligned}$$

Moreover, we can write

$$D(P_X\|Q_X) = D(\mu\|\nu). \tag{10}$$

Combining terms, the statement to be proved can thus be rewritten as

$$D(P_Y\|Q_Y) \leq D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X). \tag{11}$$

Note that this is *almost* the same statement as what you did in Homework 2, Problem 3, Part (c), but not exactly. In fact, in Homework 2, Problem 3, Part (c), we considered the case $P_X = Q_X$. The proof here proceeds along the same lines. Namely, we can use, exactly as in Homework 2, Problem 3, Part (a), the fact that

$$D(P_{X,Y}\|Q_{X,Y}) = D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X). \tag{12}$$

The remaining part is to show that indeed, $D(P_Y\|Q_Y) \leq D(P_{X,Y}\|Q_{X,Y})$. This is, of course, a direct application of the data processing inequality for KL divergence. In fact, the proof technique from Homework 2, Problem 3, Part (c) works without any changes. Namely, define the kernel

$$W(\tilde{y}|x,y) = \begin{cases} 1, & \text{if } \tilde{y} = y, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Then we have

$$P_{\tilde{Y}}(\tilde{y}) = \sum_{x,y} P_{XY}(x,y) W(\tilde{y}|x,y) = P_Y(\tilde{y}) \tag{14}$$

and

$$Q_{\tilde{Y}}(\tilde{y}) = \sum_{x,y} Q_{XY}(x,y) W(\tilde{y}|x,y) = Q_Y(\tilde{y}) \tag{15}$$

Hence, we have, by the data processing inequality for KL divergence,

$$D(P_{XY}\|Q_{XY}) \geq D(P_{\tilde{Y}}\|Q_{\tilde{Y}}) = D(P_Y\|Q_Y). \tag{16}$$

This completes the proof.

(b) Give examples where the bound is good and where the bound is bad.

Where it is good:

- Let $P_i \equiv P_1$ for all i. Let $Q_i \equiv Q_1$ for all i. Then, the bound reads
$$D\left(P_1 \| Q_1\right) \leq D(\mu\|\nu) + D(P_1\|Q_1). \tag{17}$$
We can see that in this case, the bound is tight if and only if $\mu_i = \nu_i$ for all i.

- Let $P_i = Q_i$ for all i. Then, the bound reads
$$D\left(\sum_{i=1}^{K} \mu_i P_i(y) \middle\| \sum_{i=1}^{K} \nu_i P_i(y)\right) \leq D(\mu\|\nu). \tag{18}$$
If we additionally assume $\mu_i = \nu_i$ for all i, then both the LHS as well as the RHS are zero, and thus, the bound is tight.

Where it is bad: The main insight is that in our bound, it can happen that the LHS is finite but the RHS is infinite:

- Let $P_i \equiv P_1$ for all i. Let $Q_i \equiv Q_1$ for all i. Then, the bound reads
$$D\left(P_1 \| Q_1\right) \leq D(\mu\|\nu) + D(P_1\|Q_1). \tag{19}$$
Now, if there is a single i for which $\mu_i > 0$ but $\nu_i = 0$, then $D(\mu\|\nu) = \infty$. So in this case, the bound is as loose as it gets.

- Alternatively, select P_1 and Q_1 such that there exists a value of y such that $P_1(y) > 0$ but $Q_1(y) = 0$. Then, $D(P_1\|Q_1) = \infty$, and thus, choosing $\mu_1 > 0$, this is enough to make the RHS of our bound infinite. To complete the example, we now select Q_2 such that the LHS of our bound is finite. Namely, select it such that $Q_2(y) > 0$ for all y in the alphabet of Y and select $\nu_2 > 0$. This is enough to ensure that the marginal distribution $Q_Y(y) > 0$ for all y, and thus, the LHS of our bound is finite.

- Finally, the challenge is if we can make the LHS equal to zero and the RHS infinite — the most extreme case. This is actually not hard at all. For example, as follows. Let $K = 2$. Let the alphabet of Y be $\mathcal{Y} = \{1, 2, \ldots, M\}$, where M is even. Let $P_1(y) = \frac{2}{M}$ for $y = 1, 2, \ldots M/2$, and zero otherwise. Let $P_2(y) = \frac{2}{M}$ for $y = M/2+1, M/2+2, \ldots M$, and zero otherwise. Let $Q_1(y) = P_2(y)$ and $Q_2(y) = P_1(y)$. Let $\mu_1 = \mu_2 = \nu_1 = \nu_2 = \frac{1}{2}$. With this, we find that $P(y)$ and $Q(y)$ are both the uniform distribution, making the LHS of our bound zero. But $D(P_1\|Q_1) = \infty$, making the RHS of our bound infinite.

Problem 3 (Estimation — 10 pts)**.** Let $S \in [0,1]$ be distributed with a Beta distribution with parameters $(1/2, 1/2)$, which, as we have seen in class, is $p(s) = \frac{1}{\pi} s^{-\frac{1}{2}}(1-s)^{-\frac{1}{2}}$. We make n observations $X_1, X_2, \ldots, X_n$ that are (conditionally) independent Bernoulli(S) random variables.

a) [5 pts] Calculate the conditional distribution $p(s|x_1, x_2, \ldots, x_n)$. Express it in terms of the integer t, which is the number of '1's in the sample $(x_1, x_2, \ldots, x_n)$.

Hint: For $a, b \in \mathbb{R}^+$, we have $\int_0^1 y^{a-1}(1-y)^{b-1}dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, where $\Gamma(\cdot)$ denotes the Gamma function.

b) [5 pts] We would like to estimate S from $X_1, X_2, \ldots, X_n$ such as to minimize the mean-squared error $\mathbb{E}[(S - \hat{S}(X_1, X_2, \ldots, X_n)^2]$. Find the optimum estimate $\hat{S}(X_1, X_2, \ldots, X_n)$. Simplify your result as much as possible.

Hint: The Gamma function satisfies the property, for $c \in \mathbb{R}^+$, that $\Gamma(c+1) = c\Gamma(c)$.

Solution 3. The MMSE estimator is the conditional expectation. Let t denote the number of ones in the sample $(x_1, x_2, \ldots, x_n)$.

Let us first find the conditional distribution $p(s|x_1, x_2, \ldots, x_n)$.

$$\begin{aligned} p(s, x_1, x_2, \ldots, x_n) &= p(s)p(x_1, x_2, \ldots, x_n|s) && (20) \\ &= \frac{s^{-\frac{1}{2}}(1-s)^{-\frac{1}{2}}}{\pi} s^t (1-s)^{n-t} && (21) \\ &= \frac{1}{\pi} s^{t-\frac{1}{2}}(1-s)^{n-t-\frac{1}{2}} && (22) \end{aligned}$$

and thus,

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= \frac{1}{\pi}\int_0^1 s^{t-\frac{1}{2}}(1-s)^{n-t-\frac{1}{2}} ds && (23) \\ &= \frac{1}{\pi}\frac{\Gamma(t+\frac{1}{2})\Gamma(n-t+\frac{1}{2})}{\Gamma(n+1)} && (24) \end{aligned}$$

Thus,

$$\begin{aligned} p(s|x_1, x_2, \ldots, x_n) &= \frac{p(s)p(x_1, x_2, \ldots, x_n|s)}{p(x_1, x_2, \ldots, x_n)} && (25) \\ &= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} s^{t-1/2}(1-s)^{n-t-1/2} && (26) \end{aligned}$$

To calculate the conditional mean, we now proceed as follows:

$$
\begin{aligned}
&\mathbb{E}[S|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \\
&= \int_0^1 sp(s|x_1, x_2, \ldots, x_n)ds && (27) \\
&= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} \int_0^1 s \cdot s^{t-1/2}(1-s)^{n-t-1/2}ds && (28) \\
&= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} \int_0^1 s^{t+1/2}(1-s)^{n-t-1/2}ds && (29) \\
&= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} \cdot \frac{\Gamma(t+3/2)\Gamma(n-t+1/2)}{\Gamma(n+2)} && (30) \\
&= \frac{\Gamma(n+1)}{\Gamma(n+2)} \cdot \frac{\Gamma(t+3/2)}{\Gamma(t+1/2)} && (31) \\
&= \frac{t+1/2}{n+1}, && (32)
\end{aligned}
$$

which, intriguingly, is exactly the “add-1/2” estimator that we have studied (from a different perspective) in the chapter on Distribution Estimation...