Tsinghua-Berkeley Shenzhen Institute
INFORMATION THEORY AND STATISTICAL LEARNING
Fall 2020

Problem Set 3

Issued: Monday 19th October, 2020 **Due:** Friday 30th October, 2020

Notations: We use $\mathsf{x}, \mathsf{y}, \mathsf{w}$ and $\underline{\mathsf{x}}, \underline{\mathsf{y}}, \underline{\mathsf{w}}$ to denote random variables and random vectors. We use $\mathrm{Bern}(p)$ to denote the Bernoulli distribution with the parameter p, and use $\mathrm{Binom}(n, p)$ to denote the binomial distribution with parameters n and p.

3.1. (a) $P_{\mathsf{xy}}(x, y)$ is a joint distribution of discrete random variables x and y. Assume $x_0 \in \mathcal{X}$ is a value of x, prove that

$$I(\mathsf{x};\mathsf{y}) = \sum_{x\in\mathcal{X}} P_{\mathsf{x}}(x) D(P_{\mathsf{y}|\mathsf{x}=x}||P_{\mathsf{y}|\mathsf{x}=x_0}) - D(P_{\mathsf{y}}||P_{\mathsf{y}|\mathsf{x}=x_0})$$

(b) Let $\{P_{\mathsf{y}|\mathsf{x}=x}, x \in \mathcal{X}\}$ be a set of distributions. Prove that

$$\sup_{P_{\mathsf{x}}} I(\mathsf{x};\mathsf{y}) \leq \sup_{x,x'\in\mathcal{X}} D(P_{\mathsf{y}|\mathsf{x}=x}\|P_{\mathsf{y}|\mathsf{x}=x'}).$$

This is the information-theoretic version of "radius $\leq$ diameter".

3.2. (a) For discrete random variables x, y, z, prove

$$2H(\mathsf{x},\mathsf{y},\mathsf{z}) \leq H(\mathsf{x},\mathsf{y}) + H(\mathsf{y},\mathsf{z}) + H(\mathsf{z},\mathsf{x}).$$

(b) Use the above inequality to prove *Shearer's lemma*: Place n points in $\mathbb{R}^3$ arbitrarily. Let n_1, n_2, n_3 denote the number of distinct points projected onto the xy, xz and yz-plane, respectively. Then:

$$n_1 n_2 n_3 \geq n^2.$$

3.3. Recall that $d(p||q) = D(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q))$ denotes the binary divergence function:

$$d(p||q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \tag{1}$$

(a) Prove for all $p, q \in [0, 1]$

$$d(p||q) \geq 2(p-q)^2 \log e \tag{2}$$

(b) Apply data processing inequality (Chain Rule for K-L divergence) to prove the *Pinsker-Csiszaŕr inequality*:

$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2\log e} D(P\|Q)} \tag{3}$$

where $\mathrm{TV}(P, Q)$ is the *total variation* distance between probability distribution P and Q:

$$\mathrm{TV}(P, Q) \triangleq \sup_{E\in\mathcal{F}}(P(E) - Q(E)), \tag{4}$$

with the supremum taken over all events E.

3.4. Let y be a continuous random variable distributed over the closed interval $[0,1]$. Under the null hypothesis H_0, y is uniform:

$$p_{\mathsf{y}|\mathsf{H}}(y|H_0) = \begin{cases} 1, & 0 \le y \le 1 \\ 0, & \text{o.w.} \end{cases}$$

Under the alternative hypothesis H_1, the conditional pdf of y is as follows:

$$p_{\mathsf{y}|\mathsf{H}}(y|H_1) = \begin{cases} 2y, & 0 \le y \le 1 \\ 0, & \text{o.w.} \end{cases}$$

The *a-priori* probability that y is uniformly distributed is p.

(a) Find the decision rule that minimizes the expected error.

(b) Find the closed form expression for the operating characteristic of the LRT, i.e., $P_{\text{D}} \triangleq \mathbb{P}(\hat{\mathsf{H}} = H_1 | \mathsf{H} = H_1)$ as a function of $P_{\text{F}} \triangleq \mathbb{P}(\hat{\mathsf{H}} = H_1 | \mathsf{H} = H_0)$ for the likelihood ratio test.

(c) Suppose we require that P_{D} is at least $(1+\varepsilon)P_{\text{F}}$, where $\epsilon > 0$ is a fixed constant.

i. Find $P_{\text{D}}^{\max}(\varepsilon)$, the maximal value of P_{D} that is achievable under this constraint.

ii. Find the range of values of ε that lead to non-trivial performance, i.e. $P_{\text{D}}^{\max}(\varepsilon) > 0$.

iii. When using the decision rule from part a, what values of p guarantee that $P_{\text{D}} \ge (1+\varepsilon)P_{\text{F}}$?

3.5. A 3-dimensional random vector $\underline{\mathsf{y}}$ is observed, and we know that one of the three hypotheses is true:

$$\begin{aligned} H_1:&\quad \underline{\mathsf{y}} = \underline{m}_1 + \underline{\mathsf{w}} \\ H_2:&\quad \underline{\mathsf{y}} = \underline{m}_2 + \underline{\mathsf{w}} \\ H_3:&\quad \underline{\mathsf{y}} = \underline{m}_3 + \underline{\mathsf{w}}, \end{aligned}$$

where

$$\underline{\mathsf{y}} = \begin{bmatrix} \mathsf{y}_1 \\ \mathsf{y}_2 \\ \mathsf{y}_3 \end{bmatrix}, \quad \underline{m}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \underline{m}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \underline{m}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and $\underline{\mathsf{w}}$ is a zero-mean Gaussian vector with covariance matrix $\sigma^2\mathbf{I}$.

(a) Let

$$\underline{\pi}(\underline{y}) = \begin{bmatrix} \mathbb{P}(\mathsf{H} = H_1 | \underline{\mathsf{y}} = \underline{y}) \\ \mathbb{P}(\mathsf{H} = H_2 | \underline{\mathsf{y}} = \underline{y}) \\ \mathbb{P}(\mathsf{H} = H_3 | \underline{\mathsf{y}} = \underline{y}) \end{bmatrix} = \begin{bmatrix} \pi_1(\underline{y}) \\ \pi_2(\underline{y}) \\ \pi_3(\underline{y}) \end{bmatrix},$$

and suppose that the Bayes costs are

$$C_{11} = C_{22} = C_{33} = 0, \quad C_{12} = C_{21} = 1, \quad C_{13} = C_{31} = C_{23} = C_{32} = 2.$$

i. Specify the optimum decision rule in terms of $\pi_1(\underline{y}), \pi_2(\underline{y})$ and $\pi_3(\underline{y})$.

ii. Recalling that $\pi_1 + \pi_2 + \pi_3 = 1$, express this rule completely in terms of π_1 and π_2, and sketch the decision regions in the (π_1, π_2) plane.

(b) Suppose that the three hypotheses are equally likely a priori and that the Bayes costs are

$$C_{ij} = 1 - \delta_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases}.$$

Show that the optimum decision rule can be specified in terms of the pair of sufficient statistics

$$\ell_2(\underline{\mathsf{y}}) = \mathsf{y}_2 - \mathsf{y}_1,$$
$$\ell_3(\underline{\mathsf{y}}) = \mathsf{y}_3 - \mathsf{y}_1.$$

Hint: To begin, see if you can specify the optimum decision rules in terms of

$$L_i(\underline{y}) = \frac{p_{\underline{\mathsf{y}}|\mathsf{H}}(\underline{y}|H_i)}{p_{\underline{\mathsf{y}}|\mathsf{H}}(\underline{y}|H_1)}, \quad \text{for } i = 2, 3.$$

3.6. A binary random variable x with prior $p_{\mathsf{x}}(\cdot)$ takes values in $\{-1, 1\}$. It is observed via n separate sensors; y_i denotes the observation at sensor i. The $\mathsf{y}_1, \cdots, \mathsf{y}_n$ are conditionally independent given x, i.e.,

$$p_{\mathsf{y}_1,\cdots,\mathsf{y}_n|\mathsf{x}}(y_1, \cdots, y_n|x) = \prod_{i=1}^{n} p_{\mathsf{y}_i|\mathsf{x}}(y_i|x).$$

A *local* decision $\hat{x}_i(y_i) \in \{-1, 1\}$ about the value of x is made at each sensor.

(a) In this part of the problem, each sensor sends its local decision to a fusion center. The fusion center combines the local decisions from all sensors to produce a global decision $\hat{x}(\hat{x}_1, \cdots, \hat{x}_n)$. Consider the special case in which:

- $P_{\mathsf{x}}(1) = P_{\mathsf{x}}(-1) = 1/2$;
- $\mathsf{y}_i = \mathsf{x} + \mathsf{w}_i$, where $\mathsf{w}_1, \cdots, \mathsf{w}_n$ are independent and each uniformly distributed over the interval $[-2, 2]$;
- the local decision rule is a simple thresholding of the observation, i.e.,

$$y_i \underset{\hat{x}_i(y_i)=-1}{\overset{\hat{x}_i(y_i)=1}{\gtrless}} 0.$$

Determine the minimum probability of error decision $\hat{x}(\cdot, \ldots, \cdot)$, at the fusion center.

In the remainder of the problem, there is no fusion center. The prior $P_{\mathsf{x}}(\cdot)$, observation model $p_{\mathsf{y}_i|\mathsf{x}}(\cdot|x), i = 1, 2$, and local decision rules $\hat{x}_i$, are no longer restricted as in part (a). However, we limit our attention to the two-sensor case ($n = 2$).

Consider local decisions $\hat{x}_i(y_i), i = 1, 2$, that minimize the expected cost, where the cost is defined for the two local rules jointly. Specifically, $C(\hat{x}_1, \hat{x}_2, x)$ is the cost of deciding $\hat{x}_1$ at sensor 1 and deciding $\hat{x}_2$ at sensor 2 when the true value of x is x. The cost C strictly increases with the number of errors made by the two sensors but is not necessarily symmetric.

(b) First, assume $\hat{x}_2(\cdot)$ is given. Show that the choice $\hat{x}_1^*(\cdot)$ for $\hat{x}_1(\cdot)$ that minimizes the expected (joint) cost is a likelihood ratio test of the form

$$\frac{p_{\mathsf{y}_1|\mathsf{x}}(y_1|1)}{p_{\mathsf{y}_1|\mathsf{x}}(y_1|-1)} \underset{\hat{x}_1(y_1)=-1}{\overset{\hat{x}_1(y_1)=1}{\gtrless}} \gamma_1.$$

where γ_1 is a threshold that depends on the rule $\hat{x}_2(\cdot)$. Determine the threshold γ_1.

(c) Assuming, instead, that $\hat{x}_1(\cdot)$ is given, determine the choice $\hat{x}_2^*(\cdot)$ for $\hat{x}_2(\cdot)$ that minimizes the expected joint cost.

(d) Consider a joint cost function $C(\hat{x}_1, \hat{x}_2, x)$ such that the cost is: 0 if both sensors making correct decisions; 1 if exactly one sensor makes a mistake; and L if both sensors make an error. Determine the value of L such that the optimal local decision rules at the two sensors are decoupled, i.e., the optimal threshold γ_1 does not depend on $\hat{x}_2^*(\cdot)$, and *vice versa*.