TOPICS IN STATISTICAL THEORY Part III.

Example Sheet 3 (of 3) YZ/Lent 2020

Comments and corrections to zemel@statslab.cam.ac.uk.

The sheet seems long, but that is because the questions are full with hints. Starred will be marked for the example class (I am willing to mark three questions).*

1. Let $g^* : \mathbb{R}^d \to \{0,1\}$ be the Bayes decision rule. Prove that

$$\text{(i)}\quad \mathbb{P}(g^*(X) \neq Y) = \mathbb{E}\left\{\min(\eta(X), 1-\eta(X))\right\}.$$

Now, for any decision $g : \mathbb{R}^d \to \{0,1\}$, show that

$$\text{(ii)}\quad \mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y).$$

Let $\tilde{\eta}$ be a random function that is independent of (X,Y) and consider the plug-in classifier $\tilde{g}(x) = 1$ if $\tilde{\eta}(x) \geq 1/2$. Show that

$$\text{(iii)}\quad \mathbb{P}(\tilde{g}(X) \neq Y) - \mathbb{P}(g^*(X) \neq Y) \leq 2\mathbb{E}|\eta(X) - \tilde{\eta}(X)|,$$

where the expectation and the probabilities are with respect to (X,Y) and $\tilde{\eta}$.

2. Denote the probability measure for X by P_X. Let $S_{x,\epsilon}$ be the closed ball centred at x of radius $\epsilon > 0$. The collection of all x with $P_X(S_{x,\epsilon}) > 0$ for all $\epsilon > 0$ is called the support of X or μ, denoted as $\text{supp}(P_X)$. Fix $x \in \text{supp}(P_X) \in \mathbb{R}^d$ and reorder the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to increasing values of $||X_i - x||$. The reordered data sequence is denoted by

$$(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x)).$$

Suppose that $\lim_{n\to\infty} k/n = 0$. Show that that $\|X_{(k)}(x) - x\| \to 0$ with probability one.

Show that if X_0 is independent of the data and has probability measure P_X, then $\|X_{(k)}(X_0) - X_0\| \to 0$ with probability one. [You do *not* need to prove that $\mathbb{P}(X_0 \in \text{supp}(P_X)) = 1$.]

3. Show that if $X_0, X_1, \ldots, X_n$ are one dimensional i.i.d. random variables and each has a continuous density f, then for all $u > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(n|X_{(1)}(X_0) - X_0| > u|X_0\right) = e^{-2f(X_0)u} \quad a.s.$$

4. Let P, Q be two probability measures, and let ν be a σ-finite measure that dominates μ and ν (we can always take $\nu = P + Q$). Let p and q be the densities of P and Q with respect to ν. Define

- (total variation) $TV(P,Q) = 1 - \int \min(p,q)\, d\nu$.

- (Hellinger) $h^2(P,Q) = \int (\sqrt{p} - \sqrt{q})^2 \, d\nu$
- (Kullback–Leibler) $KL(P,Q) = \int p \log \frac{p}{q} \, d\nu$.

By definition, for the product measures we have $KL(P^n, Q^n) = nKL(P,Q)$ (but not with the Hellinger or total variation distance). Show that

$$\text{(i)} \quad TV(P,Q) \le h(P,Q) \le \sqrt{KL(P,Q)}.$$

Also check that

$$\text{(ii)} \quad KL(N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

5. Let $X_1, \dots, X_n$ be an i.i.d. sample from $N(\mu, b^2)$ where $b > 0$ is a known constant. Prove using Le Cam's two-points lemma that there exists a positive C such that

$$\sup_{\mu \in \mathbb{R}} \mathbb{E}_\mu |\tilde{\mu} - \mu| \ge \frac{C}{\sqrt{n}},$$

for any estimator $\tilde{\mu}$ and all n.

6*. Let $X_1, \dots, X_n$ be an i.i.d. sample from $f \in \mathcal{F}$ where $\mathcal{F}$ denotes the set of twice continuously differentiable densities on $[0,1]$ with second derivative bounded by M. Prove that for an interior point $x_0 \in (0,1)$ there exists a constant C such that

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \left(\tilde{f}(x_0) - f(x_0) \right)^2 \ge C n^{-4/5}$$

for any density estimator $\tilde{f}$. [Hint: construct $f_0(x) = 1$ and $f_1(x) = 1 + c_n \left(K(\frac{x - x_0}{h_n}) - K(\frac{x - \tilde{x}_0}{h_n}) \right)$ where $\tilde{x}_0$ is taken to be a point in $[0,1]$ such that $|x_0 - \tilde{x}_0| \ge 1/3$ and K is the same function used in lectures, that is, $K(u) = \exp(-1/(1-u^2)) \mathbf{1}\{|u| < 1\}$. Choose c_n and h_n appropriately.]

7 (continuation). Let $X_1, \dots, X_n$ be an i.i.d. sample from $f \in \mathcal{F}$ where $\mathcal{F}$ denotes the set of twice continuously differentiable densities on $[0,1]$. Prove that for an interior point $x_0 \in (0,1)$ there exists a constant $c > 0$ such that

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \left(\tilde{f}(x_0) - f(x_0) \right)^2 \ge c$$

for any density estimator $\tilde{f}$. Thus, without a uniform bound on the second derivative, no estimator is even uniformly consistent on this function class!

8*. Here we give an alternative argument that $\mathbb{P}(\|X_{(k)}(X) - X\| > \delta) \to 0$ for all $\delta > 0$ for the k-nearest neighbour classifier when $k/n \to \infty$ and $k \to \infty$. Let $U_{(k)}$ be the k-th order statistic of independent $U_1, \dots, U_n \sim [0,1]$. Using that $U_{(k)}$ has mean $k/(n+1)$ and variance $k(n-k+1)/[(n+1)^2(n+2)]$, show that

$$\mathbb{P}\left(U_{(k)} > \frac{2k}{n} \right) \to 0.$$

For $x \in \text{supp}(P_X)$ define $F_x(t) = \mathbb{P}(\|X_1 - x\| \le t)$. Let F_x^{-1} denote the corresponding quantile function. Show that $\lim_{s \searrow 0} F_x^{-1}(s) = 0$. Deduce that $\mathbb{P}(\|X_{(k)}(x) - x\| > \delta) \to 0$ for all $\delta > 0$. Deduce further that $\mathbb{P}(\|X_{(k)}(X) - X\| > \delta) \to 0$, where X is independent of the sequence $X_1, \ldots$ and has the same distribution as X_1.

9. In this exercise we show that the infimum defining the Wasserstein distance is attained. Let P and Q be probability measures on $\mathbb{R}^d$. Show that the set of couplings of P and Q is *tight*: for any $\epsilon > 0$ there exists $M < \infty$ such that any random vector (X, Y) with $X \sim P$ and $Y \sim Q$ satisfies

$$\mathbb{P}(\|(X, Y)^t\| \ge M) < \epsilon.$$

Now let $p \ge 1$ and let (X_n, Y_n) be random vectors such that $X_n \sim P$, $Y_n \sim Q$, and $\mathbb{E}[\|X_n - Y_n\|^p] \le W_p^p(P, Q) + 1/n$. Using *Prokhorov's theorem*, deduce that there exists a subsequence $n_k \to \infty$ such that (X_{n_k}, Y_{n_k}) converge in distribution to some random vector (X, Y). Show that $X \sim P$, $Y \sim Q$ and that

$$\mathbb{E}[\|X - Y\|^p] \le \liminf_{k \to \infty} \mathbb{E}[\|X_{n_k} - Y_{n_k}\|^p].$$

[*Hint:* use the portmanteau lemma and the fact that the function $f_L(x, y) = \min(L, \|x - y\|^p)$ is continuous and bounded for all $L > 0$. Then take $L \to \infty$.]

Conclude that

$$\mathbb{E}[\|X - Y\|^p] = W_p^p(P, Q).$$

10*. Let $p \ge 1$, and let P_X denote the probability distribution of a random vector X taking values in $\mathbb{R}^d$.

a) Let X and Y be random vectors in $\mathbb{R}^d$, and let $a \in \mathbb{R}$. Show that

$$W_p(P_{aX}, P_{aY}) = |a| W_p(P_X, P_Y).$$

b) Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^d$ be a random vector. Show that

$$W_p(P_{X+Y}, P_X) \le (\mathbb{E}[\|Y\|^p])^{1/p}.$$

[In class we proved this under the assumption that X and Y are independent.]

c) Show that $W_p(P_Y, P_0) = (\mathbb{E}[\|Y\|^p])^{1/p}$.

d) Show that $W_p(P_{X+v}, P_{Y+v}) = W_p(P_X, P_Y)$ for a deterministic vector $v \in \mathbb{R}^d$.

e) Show that $W_p(P_{X+v}, P_X) \le \|v\|$ for a deterministic vector $v \in \mathbb{R}^d$.

11. Let $(X_1, Y_1) \in \mathbb{R}^d \times \mathbb{R}^d$ have joint distribution P_{XY} and let (Y_2, Z_2) have distribution P_{YZ}. Suppose that Y_1 and Y_2 have the same distribution. Assume that the random variables are discrete: there exist $(x_i), (y_j), (z_k), (p_i), (q_j), (r_k)$ such that

$$\mathbb{P}(X_1 = x_i) = p_i > 0, \quad \mathbb{P}(Y_1 = y_j) = q_j > 0, \quad \mathbb{P}(Z_2 = z_k) = r_k > 0, \quad \sum_i p_i = \sum_j q_j = \sum_k r_k = 1.$$

[Each of the sums could be finite or countably infinite, the (x_i)'s are distinct, the (y_j)'s are distinct and the (z_k)'s are distinct.] Show that there exists a random vector (X, Y, Z) such that $(X, Y) \sim P_{XY}$ and $(Y, Z) \sim P_{YZ}$. [*Hint: make X and Z conditionally independent given Y.*]

12. Let $X_n \to X$ in distribution, where X_n and X are random vectors taking values in $\mathbb{R}^d$. Suppose that $\mathbb{E}[\|X_n\|^p] \to \mathbb{E}[\|X\|^p] < \infty$ for some $p \geq 1$. Show that for any δ there exists $1 \leq R_\delta < \infty$ such that for all $R \geq R_\delta$,

$$\limsup_{n\to\infty} \mathbb{E}\|X_n\|^p 1(\|X_n\|^p > R) \leq (1 + 2^p)\delta.$$

[Hint: This holds with X_n replaced by X, without the factor $1 + 2^p$. The function

$$f_R(x) = \begin{cases} 0 & \|x\| \leq R \\ (1+R)^p(\|x\| - R) & R \leq \|x\| \leq R+1 \\ \|x\|^p & \|x\| > R+1 \end{cases}$$

is such that $f_R(x) - \|x\|^p$ is continuous and bounded and

$$\|x\|^p 1(\|x\| \geq R+1) \leq f_R(x) \leq \|x\|^p 1(\|x\| \geq R+1) + (R+1)^p 1(R < \|x\| \leq R+1).$$

Use the portmanteau lemma.]

Conclude that for all $R \geq R_\delta$, for $Z = X1(\|X\| \leq R)$ and $Z_n = X_n 1(\|X_n\| \leq R)$ we have

$$W_p^p(P_X, P_Z) \leq (1 + 2^p)\delta \quad \text{and} \quad \limsup_{n\to\infty} W_p^p(X_n, Z_n) \leq (1 + 2^p)\delta.$$

13. Consider the motivating problem of the boostrap. For any distribution F on $\mathbb{R}^d$ with finite covariance matrix and $X \sim F$ let $\mu(F) = \mathbb{E}X \in \mathbb{R}^d$ and $\Sigma(F) = \mathbb{E}(X - \mu(F))(X - \mu(F))^t$ be the mean and covariance associated with F. Let $X_1, \ldots, X_n \sim F$ be independent with empirical distribution function F_n. Let $X_1^*, \ldots, X_m^* \sim F_n$ (conditionally upon $X_1, \ldots, X_n$) be independent (a bootstrap sample). Consider the sample mean $\mu_n = n^{-1}\sum_{i=1}^n X_i$ and its bootstrapped version $\mu_m^* = m^{-1}\sum_{i=1}^m X_i^*$. Assume that the $d \times d$ covariance matrix $\Sigma(F) = mathbbE(X_1 - \mu(F))(X_1 - \mu(F))^t$ exists. The empirical mean and empirical covariance matrix associated to $X_1, \ldots, X_n$ are $\mu(F_n) = \mu_n$ and

$$\Sigma(F_n) = \mathbb{E}(X_1^* - \mu_n)(X_1^* - \mu_n)^t = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_n)(X_i - \mu_n)^t.$$

Let $\mathcal{P}$ be the set of distributions P on $\mathbb{R}^d$ satisfying $\mathbb{E}[\|X\|^2] < \infty$ for $X \sim P$. Define $T : \mathcal{P} \times \mathbb{N} \to \mathcal{P}$ by

$$T(P, m) = P_{m^{-1/2}\sum_{i=1}^m [X_i(P) - \mathbb{E}X_i(P)]}, \qquad X_i(P) \sim P \text{ independent.}$$

Show that the range of T is included in $\mathcal{P}$ and that for all $m \geq 1$ and all $P, Q \in \mathcal{P}$,

$$W_2(T(P, m), T(Q, m)) \leq W_2(T(P, 1), T(Q, 1)) \leq W_2(P, Q) + \|\mu(P) - \mu(Q)\|.$$

Let $G_n = n^{-1/2} \sum_{i=1}^{n} [X_i - \mathbb{E}X_i]$ and $G_m^* = m^{-1/2} \sum_{i=1}^{m} [X_i^* - \mathbb{E}X_i^*]$. Show that almost surely with respect to the infinite sequence $X_1, \ldots$

$$\sup_m W_2(P_{G_m^*}, P_{G_m}) \to 0.$$

[Hint: For a random vector X with mean zero and covariance matrix Σ, $\mathbb{E}[\|X\|^2]$ is the trace of Σ (you don't need to prove this).]

Condclude that as $m, n \to \infty$ (with no relation as to which of them goes faster), almost surely

$$W_2(P_{G_m^*}, N_d(0, \Sigma(F)) \to 0$$

conditionally on the infinite sequence $X_1, \ldots$. Thus, the limiting distribution of the bootstrapped mean μ_m^* is the same as that of the original statistic μ_n.