TOPICS IN STATISTICAL THEORY Part III.

Example Sheet 3 (of 3) TBB/Mich 2017

Comments and corrections to t.berrett@statslab.cam.ac.uk.

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

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

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

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

Also, for $\tilde{\eta}(x)$ which approximates η using the plug-in rule $\tilde{g}(x) = 1$ if $\tilde{\eta}(x) \geq 1/2$, prove that

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

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 $\mathrm{supp}(P_X)$. Fix $x \in \mathrm{supp}(P_X) \in \mathbb{R}^d$ 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

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

If $\lim_{n\to\infty} k/n = 0$, then prove 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 whenever $k/n \to 0$.

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 on $(\chi, \mathcal{A})$, and let ν be the $\sigma-$finite measure on $(\chi, \mathcal{A})$. Suppose that P and Q are mutually absolutely continuous, and dominated with respect to ν (we can always take $\nu = P + Q$). Let p and q be the densities of P and Q with respect to ν. Define the distance functions

- (Hellinger) $h^2(P,Q) = \int(\sqrt{dP} - \sqrt{dQ})^2 = \int(\sqrt{p} - \sqrt{q})^2\, d\nu$
- (Total Variance) $TV(P,Q) = 1 - \int \min(dP, dQ) = 1 - \int \min(p,q)\, d\nu.$
- (Kullback Leibler) $KL(P,Q) = \int \log \frac{dP}{dQ} dP = \int p \log \frac{p}{q}\, d\nu.$

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

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

Also check that

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

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

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

for any estimator $\tilde{\mu}$.

6.* Let $X_1, \ldots, X_n$ 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 such that

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

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

7. Use the fact that $\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n)$ to state and prove the extremal types theorem for minima.

8. Use integration by parts to prove the *Mills ratio*:

$$\left(\frac{1}{y} - \frac{1}{y^3}\right)\phi(y) < 1 - \Phi(y) < \frac{1}{y}\phi(y) \quad \text{for } y > 0.$$

Let (X_n) be independent $N(0,1)$ random variables, and let $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Show that there exist $a_n > 0$ and b_n such that $\mathbb{P}\left(\frac{X_{(n)} - b_n}{a_n} \leq x\right) \xrightarrow{d} G_3(x)$. Prove that for $x > 0$,

$$\frac{1}{x^2}\phi(x) - 3\int_x^\infty \frac{1}{y^3}\phi(y)\,dy < \int_x^\infty \{1 - \Phi(y)\}\,dy < \frac{1}{x^2}\phi(x) - 2\int_x^\infty \frac{1}{y^3}\phi(y)\,dy,$$

and deduce that $R(x) = 1/x + O(1/x^3)$ as $x \to \infty$. Use the Mills ratio again to show that $b_n = (2\log n)^{1/2} + o\{(2\log n)^{1/2}\}$, and deduce that we may replace a_n with $\alpha_n = (2\log n)^{-1/2}$. Finally, deduce that we may replace b_n with

$$\beta_n = (2\log n)^{1/2} - \frac{\frac{1}{2}(\log\log n + \log 4\pi)}{(2\log n)^{1/2}}.$$

9. Let (X_n) be independent with distribution function F, and let $X_{(n)} = \max_{1\le i\le n} X_i$. In each case below, where F is the distribution function corresponding to the given distribution, find constants $a_n > 0$, b_n and a nondegenerate distribution function G such that $\mathbb{P}\big(\frac{X_{(n)}-b_n}{a_n} \le x\big) \xrightarrow{d} G(x)$. Further, find constants $\alpha_n > 0$ and β_n, in terms of standard elementary functions, such that we may replace a_n with α_n and b_n with β_n: (i) $U(a,b)$; (ii) Weibull(α) (hint: look up Karamata's theorem on integrals involving slowly varying functions); (iii) Lognormal; (iv) Pareto(α); (v) Cauchy.

10. Let (X_n) be independent Bernoulli$(1/2)$ random variables, and let $X_{(n)} = \max_{1\le i\le n} X_i$. Let (x_n) be an arbitrary sequence of real numbers. By considering separately the the two cases where $x_n < 1$ infinitely often, and where $x_n \ge 1$ eventually, show that if $\mathbb{P}(X_{(n)} \le x_n) \to \rho$, then $\rho = 0$ or $\rho = 1$. Deduce that there do not exist constants $a_n > 0$, b_n and a nondegenerate distribution function G such that $\mathbb{P}\big(\frac{X_{(n)}-b_n}{a_n} \le x\big) \xrightarrow{d} G(x)$.

Generalise this argument to any distribution function F such that $x_+ = \inf\{y : F(y) \ge 1\}$ is finite and such that F has a jump at x_+.

11. (a) Let (X_n) be independent with distribution function F, and let $X_{(n)} = \max_{1\le i\le n} X_i$. Let $\tau \in [0,\infty]$ and (u_n) be a sequence of real numbers. By first considering $\tau \in [0,\infty)$, show that $\mathbb{P}(X_{(n)} \le u_n) \to e^{-\tau}$ as $n \to \infty$ if and only if $n\{1 - F(u_n)\} \to \tau$ as $n \to \infty$.

(b) Now let $X_{(1)} = \min_{1\le i\le n} X_i$, and suppose (v_n) is a sequence such that $nF(v_n) \to \eta$, for some $\eta \in [0,\infty]$. If also $n\{1 - F(u_n)\} \to \tau \in [0,\infty]$, show that

$$\mathbb{P}(X_{(1)} > v_n, X_{(n)} \le u_n) \to e^{-(\tau+\eta)}.$$

Deduce that if there exist constants $a_n > 0$, b_n and $\alpha_n > 0$, β_n and nondegenerate distribution functions G and H such that

$$\mathbb{P}\left(\frac{X_{(n)} - b_n}{a_n} \le x\right) \xrightarrow{d} G(x) \quad \text{and} \quad \mathbb{P}\left(\frac{X_{(1)} - \beta_n}{\alpha_n} \le x\right) \xrightarrow{d} H(x),$$

then

$$\mathbb{P}\left(\frac{X_{(n)} - b_n}{a_n} \le x, \frac{X_{(1)} - \beta_n}{\alpha_n} \le y\right) \xrightarrow{d} G(x)H(y).$$