UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 1 June, 2017 1:30 pm to 3:30 pm

PAPER 210

TOPICS IN STATISTICAL THEORY

*Attempt no more than **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1

Let $X_1, X_2, \ldots, X_n$ $(n \geqslant 2)$ be independent and identically distributed random variables with density function f.

Define the *univariate kernel density estimator* $\hat{f}_h$, with kernel K and bandwidth h. Show that

$$\mathbb{E}\{\hat{f}_h(x)\} = (K_h * f)(x),$$

where $(g * f)$ denotes the convolution between g and f, and K_h is a function which you should specify.

Let the kernel be $K(z) = 1_{\{|z| \leqslant 1/2\}}$. Suppose that f is twice differentiable with bounded second derivative. Show that, for all $n \geqslant 2$, $h > 0$ and $x \in \mathbb{R}$,

$$\big|\mathbb{E}\{\hat{f}_h(x)\} - f(x)\big| \leqslant \frac{h^2}{24} \sup_{z \in \mathbb{R}} |f''(z)|.$$

Show that, for $t > 0$,

$$\mathbb{P}\Big[\big|\hat{f}_h(x) - \mathbb{E}\{\hat{f}_h(x)\}\big| \geqslant t\Big] \leqslant 2\exp\Big(-2nh^2t^2\Big).$$

By integrating this bound, deduce that, for all $n \geqslant 2$, $h > 0$ and $x \in \mathbb{R}$, we have

$$\mathrm{Var}\{\hat{f}_h(x)\} \leqslant \frac{1 + \log 2}{2nh^2}.$$

2

Let (X, Y) be a random pair taking values in $\mathbb{R}^d \times \{0,1\}$. Let $\eta(x) := \mathbb{P}(Y = 1|X = x)$, and let P_X denote the marginal distribution of X. Define the *Bayes classifier* C^{Bayes} and find its risk $\mathbb{P}\{C^{\text{Bayes}}(X) \neq Y\}$.

Now let $(X_1, U_1), \ldots, (X_n, U_n)$ be independent pairs, with $X_i \sim P_X$, $U_i \sim U[0,1]$, with X_i and U_i independent, for $i = 1, \ldots, n$. Let $Y_i := \mathbb{1}_{\{U_i \leqslant \eta(X_i)\}}$. Show that the pair (X_1, Y_1) has the same joint distribution as (X, Y).

For $k \in \{1, \ldots, n\}$, define the k-nearest neighbour classifier, denoted by $\hat{C}_n^{k\text{nn}}$, with training data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Consider the case $k = 1$. Given $x \in \mathbb{R}^d$, let $Y_i' = Y_i'(x) := \mathbb{1}_{\{U_i \leqslant \eta(x)\}}$, and let $(X_{(1)}(x), U_{(1)}(x)), \ldots, (X_{(n)}(x), U_{(n)}(x))$ denote a reordering of the pairs $(X_1, U_1), \ldots, (X_n, U_n)$, such that

$$\|X_{(1)}(x) - x\| \leqslant \|X_{(2)}(x) - x\| \leqslant \ldots \leqslant \|X_{(n)}(x) - x\|.$$

Let $\tilde{C}_n^{1\text{nn}}$ denote the 1-nearest neighbour classifier trained with the pairs $(X_1, Y_1'), \ldots, (X_n, Y_n')$. Show that, for each $x \in \mathbb{R}^d$,

$$\mathbb{P}\{\tilde{C}_n^{1\text{nn}}(x) \neq \hat{C}_n^{1\text{nn}}(x)\} = \mathbb{E}\{|\eta(X_{(1)}(x)) - \eta(x)|\}.$$

Write

$$L(C) := \mathbb{P}\big\{C(X) \neq Y | (X_1, Y_1, U_1), \ldots, (X_n, Y_n, U_n)\big\}.$$

Deduce that

$$\lim_{n\to\infty} \mathbb{E}\{L(\hat{C}_n^{1\text{nn}})\} = \lim_{n\to\infty} \mathbb{E}\{L(\tilde{C}_n^{1\text{nn}})\} = \mathbb{E}[2\eta(X)\{1 - \eta(X)\}].$$

[You may use the fact that $\mathbb{E}\{|\eta(X_{(1)}(X)) - \eta(X)|\} \to 0$ as $n \to \infty$ without proof.]

Deduce further that

$$\mathbb{P}\{C^{\text{Bayes}}(X) \neq Y\} \leqslant \lim_{n\to\infty} \mathbb{E}\{L(\hat{C}_n^{1\text{nn}})\} \leqslant 2\mathbb{P}\{C^{\text{Bayes}}(X) \neq Y\}.$$

3

Let P, Q be two probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, and let ν be a $\sigma-$finite measure on $(\mathcal{X}, \mathcal{A})$. Suppose that P and Q are mutually absolutely continuous with respect to ν, and dominated by ν. Define the *Kullback–Leibler* KL(P,Q), *Total Variation* $TV(P, Q)$ and *Hellinger* $h(P, Q)$ distances between P and Q. Show that

$$TV(P, Q) \leqslant h(P, Q) \leqslant \sqrt{KL(P, Q)}.$$

[*Hint: You may use the fact that* $\log(1 + x) \leqslant x$ *for* $x > -1$ *without proof.*]

State and prove *Le Cam's two points lemma.*

Let $X_1, \ldots, X_n$ be an independent and identically distributed sample from $N(\mu, \sigma^2)$ where σ is a known constant. Show that there exists $c > 0$ such that

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

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

4

Consider a fixed design homoscedastic regression model

$$Y_i = m(x_i) + \sigma\epsilon_i, \quad \text{for} \quad i = 1, 2, \dots, n,$$

where $a < x_1 < \ldots < x_n < b$ and ϵ_i are independent and identically distributed with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = 1$.

Define a *cubic spline* on $[a, b]$ with knots at $x_1, \dots, x_n$. When is a cubic spline a *natural cubic spline*? Define the *natural cubic spline interpolant* to $\mathbf{g} = (g_1, \dots, g_n)^T$ at $x_1, \dots, x_n$.

Let g denote the natural cubic spline interpolant to $\mathbf{g} = (g_1, \dots, g_n)^T$ at $x_1, \dots, x_n$. Show that for any twice continuously differentiable function $\tilde{g}$ on $[a, b]$ satisfying $\tilde{g}(x_i) = g_i$, for $i = 1, \dots, n$, we have

$$\int_a^b g''(x)^2 \, dx \leqslant \int_a^b \tilde{g}''(x)^2 \, dx,$$

with equality if and only if $\tilde{g} = g$.

Deduce that, for each $\lambda \in (0, \infty)$, there exists a unique minimiser $\hat{g}_\lambda$, which you should specify, of

$$S_\lambda(\tilde{g}) := \sum_{i=1}^n \{Y_i - \tilde{g}(x_i)\}^2 + \lambda \int_a^b \tilde{g}''(x)^2 \, dx$$

over $\tilde{g} \in S_2[a, b]$, the set of twice continuously differentiable functions on $[a, b]$.

[In this question you may use the fact that the natural cubic spline interpolant to $(g_1, \dots, g_n)^T$ at $x_1, \dots, x_n$ is unique, and that there exists a nonnegative definite matrix Γ, such that $\int_a^b g''(x)^2 \, dx = \mathbf{g}^T \Gamma \mathbf{g}$.]

END OF PAPER