UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 5 June, 2017 1:30 pm to 4:30 pm

PAPER 205

MODERN STATISTICAL METHODS

*Attempt no more than **FOUR** questions.*

*There are **SIX** 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

Given an input space $\mathcal{X}$, what does it mean for k to be a *positive definite kernel*? We will henceforth refer to a positive definite kernel as simply a kernel for brevity, and all kernels will be on the input space $\mathcal{X}$.

What is a *reproducing kernel Hilbert space* (RKHS)? [You need not define what a Hilbert space is.]

Show that if $k_1, \ldots, k_p$ are kernels, then $k = \sum_{j=1}^p k_j$ is also a kernel.

Let the RKHS $\mathcal{H}$ associated with a kernel k have norm denoted by $\|\cdot\|_{\mathcal{H}}$. Let $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be a loss function, $Y \in \mathbb{R}^n$ a vector of responses and $x_1, \ldots, x_n \in \mathcal{X}$ a collection of inputs. Let $K \in \mathbb{R}^{n \times n}$ be the kernel matrix with entries $K_{ij} = k(x_i, x_j)$ and let $\lambda > 0$. Prove that $\hat{f}$ minimises

$$Q_1(f) = c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2$$

over $f \in \mathcal{H}$ if and only if $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^T \in \mathbb{R}^n$ minimises

$$M(\alpha) = c(Y, x_1, \ldots, x_n, K\alpha) + \lambda \alpha^T K \alpha$$

over $\alpha \in \mathbb{R}^n$.

The final part of the question uses the following facts which you need not prove. Suppose $k = \sum_{j=1}^p k_j$ where $k_1, \ldots, k_p$ are kernels with associated RKHS's $\mathcal{H}_1, \ldots, \mathcal{H}_p$ having corresponding norms $\|\cdot\|_{\mathcal{H}_1}, \ldots, \|\cdot\|_{\mathcal{H}_p}$. Then the RKHS $\mathcal{H}$ associated with k satisfies

$$\mathcal{H} = \left\{ \sum_{j=1}^p f_j : f_j \in \mathcal{H}_j \text{ for all } j = 1, \ldots, p \right\}$$

with squared norm

$$\|f\|_{\mathcal{H}}^2 = \inf \left\{ \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}^2 : f = \sum_{j=1}^p f_j,\ f_j \in \mathcal{H}_j \text{ for all } j \right\}.$$

It can be shown that the infimum is achieved uniquely, so given $f \in \mathcal{H}$ there exists a unique $(f_1, \ldots, f_p) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$ such that $\sum_{j=1}^p f_j = f$ and $\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}^2$.

Suppose now that $(\hat{f}_1, \ldots, \hat{f}_p)$ minimises

$$Q_2(f_1, \ldots, f_p) = c\left(Y, x_1, \ldots, x_n, \sum_{j=1}^p f_j(x_1), \ldots, \sum_{j=1}^p f_j(x_n)\right) + \lambda \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}^2$$

over $(f_1, \ldots, f_p) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$. Show that then $Q_2(\hat{f}_1, \ldots, \hat{f}_p) = Q_1(\hat{f})$ where $\hat{f} = \sum_{j=1}^p \hat{f}_j$.

Show furthermore that $\hat{f}$ minimises Q_1. Finally prove that $\hat{f}_j(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k_j(\cdot, x_i)$ for all j, where $\hat{\alpha} \in \mathbb{R}^n$ minimises M.

[*Throughout your answer to this question, you may use properties of RKHS's derived or stated in lectures, without proof.*]

2

Let $Y \in \mathbb{R}^n$ be a vector of responses and $X \in \mathbb{R}^{n\times p}$ a matrix of predictors. Suppose that the columns of X have been centred and scaled to have ℓ_2-norm $\sqrt{n}$, and that Y is also centred. Consider the linear model (after centring),

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1},$$

where $\mathbf{1}$ is an n-vector of 1's and $\bar{\varepsilon} = \mathbf{1}^T\varepsilon/n$. Let $S = \{j : \beta_j^0 \neq 0\}$ and $s = |S|$. Define the *Lasso estimator* $\hat{\beta}$ of β^0 with regularisation parameter $\lambda > 0$ (here and throughout we suppress the dependence of the Lasso solution on λ).

What does it mean for a random variable W to be *sub-Gaussian* with parameter σ? Suppose that ε has independent mean-zero sub-Gaussian components with common parameter $\sigma > 0$. Prove that when $\lambda = 2\sigma A\sqrt{\log(p)/n}$ with $A > 0$, the event

$$\Omega = \{2\|X^T\varepsilon\|_\infty/n \leqslant \lambda\}$$

has probability at least $1 - 2p^{-(A^2/2-1)}$.

Write down the KKT conditions for the Lasso. Let $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$ and set $\hat{s} = |\hat{S}|$. Show that on the event Ω, for any non-empty subset B of $\hat{S}$, we have

$$\frac{1}{n}\mathrm{sgn}(\hat{\beta}_B)^T X_B^T X(\beta^0 - \hat{\beta}) \geqslant \frac{\lambda|B|}{2}. \qquad (*)$$

State a sufficient condition based on X and S such that on Ω, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leqslant \frac{16\lambda^2 s}{\phi^2} \qquad (**)$$

where $\phi^2 > 0$.

Let κ_m^2 be the maximum eigenvalue of $X_M^T X_M/n$ over all $M \subset \{1,\ldots,p\}$ with $|M| = m$. Let

$$m^* = \min\{m \geqslant 1 : m > 64\kappa_m^2 s/\phi^2\},$$

with $m^* = \infty$ if there does not exist any m satisfying the condition defining the set above. Suppose that $(**)$ holds on Ω. Prove that on Ω, we have $\hat{s} < m^*$. [*Hint: First try to obtain an upper bound on the LHS of $(*)$ involving $\kappa_{|B|}$*].

By considering the minimality of m^*, show furthermore that on Ω, we have $\hat{s} \leqslant 64\kappa_{m^*}^2 s/\phi^2$.

3

Suppose we have null hypotheses $H_1, \ldots, H_m$ and associated p-values $p_1, \ldots, p_m$. What is the *family-wise error rate* (FWER)? What is the *false discovery rate* (FDR)?

Describe the *closed testing procedure* and state a result concerning its FWER.

Suppose I_0 is the set of indices of true null hypotheses and $m_0 = |I_0|$. In all that follows we will assume that p_i, $i \in I_0$ are independent and independent of $\{p_i : i \notin I_0\}$. Furthermore, we will assume that with probability 1, the p-values $p_1, \ldots, p_m$ are distinct. Let $p_{(1)}, \ldots, p_{(m)}$ be the order statistics of the p-values, so (i) is the index of the ith smallest p-value. Describe the *Benjamini–Hochberg procedure* and prove that it controls the FDR at a given level α.

For any non-empty $I \subseteq \{1, \ldots, m\}$, let $p_{(i,I)}$, $i = 1, \ldots, |I|$ be the order statistics of the p-values $\{p_i : i \in I\}$, so for example $(1, I)$ gives the index of the smallest p-value in $\{p_i : i \in I\}$. Show that

$$\mathbb{P}(\min_i p_{(i,I_0)}/i \leqslant \alpha/|I_0|) \leqslant \alpha$$

for all $\alpha \in [0, 1]$. [*Hint: Consider the Benjamini–Hochberg procedure in the setting where* $I_0^c = \emptyset$.]

Consider the procedure that rejects $H_{(i)}$ if there exists $j \geqslant i$ with

$$p_{(j)} \leqslant \frac{\alpha}{m - j + 1}.$$

Show that this procedure controls the FWER at level α. You may use the fact (which you need not prove) that given i

$$\text{if there exists } j \geqslant i \text{ with } p_{(j)} \leqslant \frac{\alpha}{m - j + 1},$$
$$\text{then } \min_k p_{(k,I)}/k \leqslant \alpha/|I| \text{ for all } I \text{ such that } (i) \in I. \qquad (*)$$

[*Hint: Consider a closed testing procedure.*]

4

What does it mean for a distribution to satisfy the *global Markov property* with respect to a DAG $\mathcal{G}$? What does it mean for two DAGs to be *Markov equivalent*? State a result relating Markov equivalence to the structure of DAGs. What does it mean for a distribution to be *faithful* to a DAG $\mathcal{G}$?

Describe, in detail, the population version of the *PC algorithm* applied to a distribution P. Prove that if P is faithful to a DAG $\mathcal{G}^0$, then the output of the PC algorithm will identify the Markov equivalence class of $\mathcal{G}^0$. [You need not prove the existence of topological orderings.]

Suppose now that $Z \in \mathbb{R}^4$ has a distribution P that is faithful to a DAG $\mathcal{G}^0$. The only independencies or conditional independencies satisfied by the components of Z are given below:

$$Z_2 \perp\!\!\!\perp Z_4$$
$$Z_1 \perp\!\!\!\perp Z_3 \,|\, (Z_2, Z_4).$$

Find the DAG $\mathcal{G}^0$, briefly justifying your answer.

[*In no part of your answer to this question do you need to explain what a graph is or define any graph terminology such as d-separation or topological ordering.*]

5

Let $(x_i, Y_i) \in \mathcal{X} \times \{-1, 1\}$, $i = 1, \ldots, n$ and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Let $K \in \mathbb{R}^{n \times n}$ be the kernel matrix so $K_{ij} = k(x_i, x_j)$. Write down the optimisation problem solved by the *support vector machine* in terms of the intercept $\mu \in \mathbb{R}$, parameter $\alpha \in \mathbb{R}^n$, tuning parameter $\lambda > 0$, K and the data. Let $(\hat{\mu}, \hat{\alpha}) \in \mathbb{R} \times \mathbb{R}^n$ minimise the support vector machine objective. Write down the predicted response corresponding to an input vector $x \in \mathcal{X}$.

Now define the *subdifferential* $\partial f(x)$ of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$.

Suppose that $h : \mathbb{R} \to \mathbb{R}$ is convex and given a vector $c \in \mathbb{R}^n$ and $b \in \mathbb{R}$, let $f(x) = h(c^T x + b)$. Show that f is convex and that

$$\partial f(x) = c\, \partial h(u) = \{cv : v \in \partial h(u)\}$$

where $u = c^T x + b$. [*Hint: First show that if $v \in \partial h(u)$ then $cv \in \partial f(x)$. For the converse, it may help to consider the orthogonal projection $P = cc^T / \|c\|_2^2$.*]

Assume that K is invertible and let K_i denote the ith column of K. Prove that if $Y_i(K_i^T \hat{\alpha} + \hat{\mu}) > 1$ then $\hat{\alpha}_i = 0$. [*Hint: It may help to use the fact that* $\max(u, 0) = (|u| + u)/2$.]

Discuss very briefly the computational implications of the result above. Will there be much of a computational benefit when there are many points for which Y_i disagrees with the corresponding prediction? Briefly justify your answer.

[*You may use standard results concerning subdifferentials and convex functions given in lectures, without proof.*]

6

Let $Y \in \mathbb{R}^n$ be a vector of responses and $X \in \mathbb{R}^{n \times p}$ a matrix of predictors. Consider a normal linear model $Y = X\beta^0 + \varepsilon$ where $\varepsilon \sim N_n(0, \sigma^2 I)$. Define the *debiased Lasso* estimator $\hat{b}$ in terms of an approximate inverse $\hat{\Theta}$ and a Lasso estimate $\hat{\beta}$ of β^0.

For $j = 1, \ldots, p$, set

$$\hat{\gamma}^{(j)} = \text{argmin}_{\gamma \in \mathbb{R}^{p-1}} \{\|X_j - X_{-j}\gamma\|_2^2/(2n) + \lambda_j \|\gamma\|_1\},$$
$$\hat{\tau}_j^2 = \|X_j - X_{-j}\hat{\gamma}^{(j)}\|_2^2/n + \lambda_j \|\hat{\gamma}^{(j)}\|_1,$$

where $\lambda_j > 0$, $j = 1, \ldots, p$ are tuning parameters. Give, with detailed justification, a construction of the approximate inverse $\hat{\Theta}$ such that we have

$$\sqrt{n}(\hat{b} - \beta^0) = W + \Delta$$

where

$$W|X \sim N_p(0, \sigma^2 \hat{\Omega}),$$
$$\|\Delta\|_\infty \leqslant \sqrt{n}\|\beta^0 - \hat{\beta}\|_1 \max_j \frac{\lambda_j}{\hat{\tau}_j^2},$$

and $W, \hat{\Omega}$ and Δ should all be specified. [You need not derive KKT conditions for the Lasso but should state them clearly if you use them.]

Based on this, write down an expression for an approximate $(1-\alpha)$-level confidence interval for β_j^0 in terms of $\hat{b}$, σ^2, $\hat{\gamma}^{(j)}$ and $\hat{\tau}_j^2$.

Let $s = |\{j : \beta_j^0 \neq 0\}|$. Give a random design for X and conditions such that asymptotically as $n \to \infty$, there exist λ_j, tuning parameter λ for the Lasso estimate $\hat{\beta}$, and constant A with

$$\mathbb{P}(\|\Delta\|_\infty > As\log(p)/\sqrt{n}) \to 0.$$

Your conditions should allow for p and s to grow with n. [You need not prove that your conditions imply the above limiting result.]

END OF PAPER