PRINCIPLES OF STATISTICS Part II
Example Sheet 4 (of 4) RDS/Michaelmas 2024

1. Consider a classification setting where $(X, Y) \in \mathbb{R}^p \times \{0, 1\}$ is a random input–output pair. Let f_j be the conditional density of X given $Y = j$ and let $\pi_j = \mathbb{P}(Y = j)$ for $j \in \{0, 1\}$. Show that δ_π given by

$$\delta_\pi(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)\pi_1}{f_0(x)\pi_0} > 1 \\ 0 & \text{otherwise} \end{cases}$$

is a Bayes classifier. Show moreover that if

$$\mathbb{P}\left(\frac{f_1(X)\pi_1}{f_0(X)\pi_0} = 1\right) = 0,$$

then any Bayes classifier δ satisfies $\mathbb{P}(\delta(X) = \delta_\pi(X)) = 1$.

2. In each of the parts below, we consider the classification setting in Question 1.

(a) Consider first the special case in which $X \,|\, Y = j \sim N_p(\mu_j, \Sigma)$ where Σ is a known positive definite matrix and the means μ_0, μ_1 are known with $\mu_0 \neq \mu_1$. Show that a minimax classifier δ, that is one where

$$\max_{y \in \{0,1\}} \mathbb{P}(\delta(X) \neq y \,|\, Y = y) = \inf_{\delta'} \max_{y \in \{0,1\}} \mathbb{P}(\delta'(X) \neq y \,|\, Y = y),$$

is obtained by selecting $\delta(X) = 1$ whenever

$$D := \frac{1}{2}(\mu_0 + \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1) + X^\top \Sigma^{-1}(\mu_1 - \mu_0) > 0,$$

and 0 otherwise. [*Hint: First argue that* $D \sim N(\Delta^2/2, \Delta^2)$ *when* $X \sim N_p(\mu_1, \Sigma)$ *and* $D \sim N(-\Delta^2/2, \Delta^2)$ *when* $X \sim N_p(\mu_0, \Sigma)$, *where* $\Delta^2 := (\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 - \mu_0)$.]

(b) We now return to a more general setting where the conditional distributions of $X \,|\, Y = j$ are not necessarily Gaussian. Suppose we have i.i.d. copies $(X_i, Y_i)_{i=1}^n$ of (X, Y). Consider a sample version of linear discriminant analysis involving estimates

$$\widehat{\mu}_j := \frac{1}{n_j} \sum_{i:Y_i=j} X_i \qquad \text{and} \qquad \widehat{\Sigma} := \frac{1}{n-2} \sum_{j=0,1} \sum_{i:Y_i=j}^{n} (X_i - \widehat{\mu}_j)(X_i - \widehat{\mu}_j)^\top$$

where $n_j := \sum_{i=1}^n \mathbb{1}_{\{Y_i=j\}}$, for $j \in \{0, 1\}$.

(i) Writing $\Sigma_j := \mathrm{Var}(X \,|\, Y = j)$ for $j \in \{0, 1\}$ and $\pi := \mathbb{P}(Y = 1)$, show that as $n \to \infty$,

$$\widehat{\Sigma} \xrightarrow{p} \Sigma := \pi\Sigma_1 + (1 - \pi)\Sigma_0.$$

(ii) Suppose that Σ is positive definite and $\pi \in (0, 1)$. Show that the vector $\widehat{\beta} := \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0)$ satisfies $\widehat{\beta} \xrightarrow{p} \beta^*$ as $n \to \infty$, where β^* maximises

$$\frac{\mathrm{Var}(\mathbb{E}(\beta^\top X \,|\, Y))}{\mathbb{E}(\mathrm{Var}(\beta^\top X \,|\, Y))}$$

over $\beta \in \mathbb{R}^p$, $\beta \neq 0$. (Thus β^* has the interpretation of being a direction upon which the projection of X has the maximal ratio of the "between class variance" to the "within class variance".)

3. Let (X_i, Y_i) be i.i.d. copies of a random pair $(X, Y) \in \mathbb{R} \times \mathbb{R}$. Let $\gamma := \mathrm{Cov}(X, Y)$, $\sigma_1 := \sqrt{\mathrm{Var}(X)}$, $\sigma_2 := \sqrt{\mathrm{Var}(Y)}$ and let $v := \mathrm{Var}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$, with all of these quantities assumed to be finite and non-zero.

   (i) Show that the sample covariance

   $$\widehat{\gamma} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

   satisfies $\sqrt{n}(\widehat{\gamma} - \gamma) \xrightarrow{d} N(0, v)$.

   (ii) Now let ρ be the correlation af X and Y. Find the distributional limit of $\sqrt{n}(\widehat{\rho} - \rho)$ where $\widehat{\rho}$ is the sample correlation, in the case where X and Y are independent.

4. Let $F : \mathbb{R} \to [0, 1]$ be a probability distribution function and let $F^{-1} : (0, 1) \to \mathbb{R}$ be the quantile function $F^{-1}(p) := \inf\{t : F(t) \geq p\}$.

   (a) Show that for $p \in (0, 1)$ and $t \in \mathbb{R}$,

   $$F^{-1}(p) \leq t \Longleftrightarrow p \leq F(t).$$

   Conclude that if $U \sim U[0, 1]$, then $F^{-1}(U) \sim F$.
   [*Hint: F is always right continuous, that is $F(t + a_n) \downarrow F(t)$ for all $a_n \downarrow 0$.*]

   (b) Now suppose F is continuous and strictly increasing, and F_n for $n \in \mathbb{N}$ are probability distribution functions such that $F_n(t) \to F(t)$ for all $t \in \mathbb{R}$. Show that then $F_n^{-1}(p) \to F^{-1}(p)$ for all $p \in (0, 1)$. [*Hint: Consider (for example) $F(F_n^{-1}(p))$.*]

5. Suppose $X_1, X_2, \ldots$ are i.i.d. and $\widehat{\theta}_n := T_n(X_1, \ldots, X_n)$ is an estimate of a parameter $\theta \in \mathbb{R}$. Denoting the true parameter by θ_0, suppose $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} F$ where F is some continuous and strictly increasing distribution function. Suppose we have an estimate $\widehat{F}_n$ of F, e.g. coming from the bootstrap, satisfying $\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)| \xrightarrow{a.s.} 0$. Given $\alpha \in (0, 1)$, let $\widehat{l}_n := \widehat{F}_n^{-1}(\alpha/2)$ and $\widehat{u}_n := \widehat{F}_n^{-1}(1 - \alpha/2)$. Show that the confidence interval

   $$\widehat{C}_n := \{\theta : \widehat{l}_n \leq \sqrt{n}(\widehat{\theta}_n - \theta) \leq \widehat{u}_n\}$$

   satisfies

   $$\mathbb{P}(\theta_0 \in \widehat{C}_n) \to 1 - \alpha.$$

   [*Hint: Recall that $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$.*]

6. Let $f, g : \mathbb{R} \to [0, \infty)$ be bounded probability density functions such that $f(x) \leq Mg(x)$ for all $x \in \mathbb{R}$ and some constant $M > 0$. Suppose you can simulate a random variable X of density g and a random variable $U \sim U[0, 1]$. Consider the following 'accept–reject' algorithm:

   `Step 1.  Draw` $X \sim g, U \sim U[0, 1]$.

   `Step 2.  Accept` $Y = X$ `if` $U \leq f(X)/(Mg(X))$`, and return to Step 1 otherwise.`

   Show that Y has density f.

7. Let $U_1, U_2 \overset{\text{i.i.d.}}{\sim} U[0, 1]$ and define

   $$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \;\; X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2).$$

   Show that $X_1, X_2 \overset{\text{i.i.d.}}{\sim} N(0, 1)$.

8. Consider observations $X_1, \dots, X_n$ from a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}, \Theta = \mathbb{R}^p, p \in \mathbb{N}$, and denote by $\Pi(\cdot|X_1, \dots, X_n)$ the posterior distribution arising from a $N_p(0, I)$ prior π on Θ. The Markov chain $(\vartheta_m : m \in \mathbb{N})$ is started at arbitrary $\vartheta_0 \in \mathbb{R}^p$ and generated as follows:

   `Step 1.` `For` $m \in \mathbb{N} \cup \{0\}, \delta \in (0, 1/2)$ `and given` ϑ_m, `generate` $\xi \sim \pi = N_p(0, I)$ `and set`
   $$s_m = \sqrt{1 - 2\delta}\vartheta_m + \sqrt{2\delta}\xi.$$

   `Step 2.` `Define`
   $$\vartheta_{m+1} = \begin{cases} s_m, & \texttt{with probability } \rho(\vartheta_m, s_m) \\ \vartheta_m, & \texttt{with probability } 1 - \rho(\vartheta_m, s_m), \end{cases}$$
   `where the acceptance probabilities are given by`
   $$\rho(\vartheta_m, s_m) = \min\left\{e^{\ell(s_m) - \ell(\vartheta_m)}, 1\right\}, \qquad \ell(\theta) = \sum_{i=1}^{n} \log f(X_i, \theta).$$

   `Step 3.` `Repeat the above with` $m \mapsto m + 1$.

   Show that the posterior distribution $\Pi(\cdot|X_1, \dots, X_n)$ is an invariant distribution for $(\vartheta_m : m \in \mathbb{N})$.

   [*Hint: Show that the algorithm given is a special case of the Metropolis–Hastings algorithm.*]

9. Let $X_1, \dots, X_n$ be drawn i.i.d. from a continuous distribution function $F : \mathbb{R} \to [0, 1]$, and let $\widehat{F}_n(t) := (1/n) \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i)$ be the empirical distribution function. Use the Kolmogorov–Smirnov theorem to construct a confidence band for the unknown function F of the form
   $$\{C_n(x) := [\widehat{F}_n(x) - R_n, \widehat{F}_n(x) + R_n] : x \in \mathbb{R}\}$$
   that satisfies $\mathbb{P}(F(x) \in C_n(x)\ \forall x \in \mathbb{R}) \to 1 - \alpha$ as $n \to \infty$, and where $R_n = R/\sqrt{n}$ for some fixed $R > 0$.

10. Suppose for real-valued random variables $X, X_1, X_2, \dots$ we have $X_n \xrightarrow{d} X$ and the distribution function F of X is continuous. Show that
   $$\sup_t |F_n(t) - F(t)| \to 0.$$
   [*Hint: Argue similarly to the proof of the Glivenko–Cantelli theorem.*]

11. Let $X_1, X_2, \dots$ be i.i.d. and consider estimating some parameter $\theta \in \mathbb{R}$ using $\widehat{\theta}_n := T_n(X_1, \dots, X_n)$. We wish to use this to test the null hypothesis $\theta = \theta_0$. We assume that
   $$R_n := \sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} G$$
   for some unknown distribution G. Now let $m_n \in \mathbb{N}$ be such that $m_n \to \infty$ but $m_n/n \to 0$. Let $B_n := \lfloor n/m_n \rfloor$ and for $b = 1, \dots, B_n$, define
   $$R_n^{(b)} := \sqrt{m_n}\{T_{m_n}(X_{(b-1)m_n+1}, \dots, X_{bm_n}) - \theta_0\}.$$
   Finally, write $\widehat{G}_n$ for the empirical distribution function of $\{R_n^{(1)}, \dots, R_n^{(B_n)}\}$.

(a) Using the fact that for any $Z_1, \ldots, Z_k \overset{\text{i.i.d.}}{\sim} F$, their empirical distribution $\widehat{F}_k$ satisfies

$$\mathbb{P}\left(\sup_t |\widehat{F}_k(t) - F(t)| > \epsilon\right) \leq 2e^{-2k\epsilon^2},$$

show that $\sup_t |\widehat{G}_n(t) - G(t)| \xrightarrow{p} 0$.
[*Hint: Note that* $\sup_t |\widehat{G}_n(t) - G(t)| \leq \sup_t |\widehat{G}_n(t) - G_n(t)| + \sup_t |G_n(t) - G(t)|$ *where* G_n *is the distribution of* $R_n^{(1)}$.]

(b*) Argue that the test ϕ_n that rejects (i.e. $\phi_n = 1$) when

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) > \widehat{G}_n^{-1}(1 - \alpha)$$

has $\mathbb{P}(\phi_n = 1) \to \alpha$.
[*Hint: Use the fact that for any sequence* $Z_1, Z_2, \ldots$ *of random variables,* $Z_n \xrightarrow{p} Z$ *if and only if every subsequence of the* Z_n *contains a further subsequence* n_k *where* $Z_{n_k} \xrightarrow{a.s.} Z$ *as* $k \to \infty$.]