PRINCIPLES OF STATISTICS – EXAMPLES 4/4

Part II, Michaelmas 2015, RN (email: r.nickl@statslab.cam.ac.uk)

1. Consider classifying an observation of a random vector X in $\mathbb{R}^p$ into either a $N(\mu_1, \Sigma)$ or a $N(\mu_2, \Sigma)$ population, where Σ is a known nonsingular covariance matrix and where $\mu_1 \neq \mu_2$ are two distinct known mean vectors.

a) For a prior π assigning probability q to μ_1 and $1-q$ to μ_2, show that the Bayes classifier is unique and assigns X to $N(\mu_1, \Sigma)$ whenever

$$U \equiv D - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$$

exceeds $\log((1-q)/q)$, where $D = X^T\Sigma^{-1}(\mu_1 - \mu_2)$ is the discriminant function.

b) Show that $U \sim N(\Delta^2/2, \Delta^2)$ whenever $X \sim N(\mu_1, \Sigma)$, and that $U \sim N(-\Delta^2/2, \Delta^2)$ whenever $X \sim N(\mu_2, \Sigma)$, where Δ is the *Mahalanobis distance* between μ_1 and μ_2 given by

$$\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2).$$

c) Show that a minimax classifier is obtained from selecting $N(\mu_1, \Sigma)$ whenever $U \geq 0$.

2. Consider classification of an observation X into a population described by a probability density equal to either f_1 or f_2. Assume $P_{f_i}(f_1(X)/f_2(X) = k) = 0$ for all $k \in [0, \infty], i \in \{1, 2\}$. Show that any admissible classification rule is a Bayes classification rule for some prior π.

3. Based on an i.i.d. sample $X_1, \ldots, X_n$, consider an estimator $T_n = T(X_1, \ldots, X_n)$ of a parameter $\theta \in \mathbb{R}$. Suppose the bias function $B_n(\theta) = ET_n - \theta$ can be approximated as

$$B_n(\theta) = \frac{a}{n} + \frac{b}{n^2} + O(n^{-3})$$

for some real numbers a, b. Show that the jackknife bias corrected estimate $\tilde{T}_n$ of θ based on T_n satisfies

$$E\tilde{T}_n - \theta = O(n^{-2}).$$

4. For $F : \mathbb{R} \to [0, 1]$ a probability distribution function, define its generalised inverse $F^-(u) = \inf\{x : F(x) \geq u\},\ x \in [0, 1]$. If U is a uniform $U[0, 1]$ random variable, show that the random variable $F^-(U)$ has distribution function F.

5. 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 from a uniform $U[0, 1]$ distribution. 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.

6. Let U_1, U_2 be i.i.d. uniform $U[0, 1]$ 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 are i.i.d. $N(0, 1)$.

7. Let $X_1, \ldots, X_n$ be drawn i.i.d. random variables from distribution P with unknown mean μ and variance σ^2. Write $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for the sample mean, and let $\bar{X}_n^b = (1/n)\sum_{i=1}^n X_{ni}^b$ be the mean of a bootstrap sample $(X_{ni}^b : i = 1, \ldots, n) \sim^{i.i.d.} \mathbb{P}_n$ generated from the X_i's. Choosing roots R_n such that

$$\mathbb{P}_n\left(|\bar{X}_n^b - \bar{X}_n| \leq \frac{R_n}{\sqrt{n}}\right) = 1 - \alpha$$

for some $0 < \alpha < 1$, let

$$C_n^b = \left\{v \in \mathbb{R} : |\bar{X}_n - v| \leq \frac{R_n}{\sqrt{n}}\right\}$$

be the corresponding bootstrap confidence interval. Show that R_n converges to a constant in $P^{\mathbb{N}}$-probability and deduce further that C_n^b is an exact asymptotic level $1 - \alpha$ confidence set, that is, show that, as $n \to \infty$,

$$P^{\mathbb{N}}(\mu \in C_n^b) \to 1 - \alpha.$$

8. Let $X_1, \ldots, X_n$ be drawn i.i.d. from a continuous distribution function $F : \mathbb{R} \to [0, 1]$, and let $F_n(t) = (1/n)\sum_{i=1}^n 1_{(-\infty,t]}(X_i)$ be the empirical distribution function. Use the Kolmogorov-Smirnov theorem to construct a confidence band for the unknown function F of the form

$$\{C_n(x) = [F_n(x) - R_n, F_n(x) + R_n] : x \in \mathbb{R}\}$$

that satisfies $P_F^{\mathbb{N}}(F(x) \in C_n(x)\ \forall x \in \mathbb{R}) \to 1 - \alpha$ as $n \to \infty$, and where $R_n = R/\sqrt{n}$ for some fixed quantile constant $R > 0$.

9. Let $X_1, \ldots, X_n$ be drawn i.i.d. from a differentiable probability density $f : \mathbb{R} \to [0, \infty)$, and assume that $\sup_{x\in\mathbb{R}}(|f(x)| + |f'(x)|) \leq 1$. Define the density estimator

$$\hat{f}_n(x) = \frac{1}{n^{2/3}}\sum_{i=1}^n \mathbf{1}\{-1/2 \leq n^{1/3}(x - X_i) \leq 1/2\},\ x \in \mathbb{R}.$$

Show that, for every $x \in \mathbb{R}$ and every $n \in \mathbb{N}$,

$$E|\hat{f}_n(x) - f(x)| \leq \frac{2}{n^{1/3}}.$$