Exercise sheet 8

The support of a distribution P on $\mathbb{R}^d$ (or any Polish space) is the set of points x such that $P(B_\epsilon(x)) > 0$ for all $\epsilon > 0$, where $B_\epsilon(x) = \{y : \|y - x\| < \epsilon\}$. You may use without proof that $\mathbb{P}(X \in \text{supp}P_X) = 1$, where $\text{supp}P_X$ is the support of the distribution of X. The proof of this is given below, but is **not** examinable.

For each $x \in \mathbb{R}^d$ let $r(x) := \sup\{r \geq 0 : P_X(B_r(x)) = 0\}$, with $r(x) = 0$ for $x \in \text{supp}(P_X)$ and $r(x) > 0$ otherwise. For each $x \notin \text{supp}(P_X)$ there exists $x' \in \mathbb{Q}^d$ with $\|x - x'\| \leq r(x)/4$. This satisfies $P_X(B_{r(x)/2}(x')) \leq P_X(B_{3r(x)/4}(x)) = 0$ and so $r(x') \geq r(x)/2$ and $\|x - x'\| \leq r(x')/2$. Hence

$$\mathbb{P}(X_0 \notin \text{supp}(P_X)) \leq P_X\left(\bigcup_{x' \in \mathbb{Q}^d \setminus \text{supp}(P_X)} B_{r(x')/2}(x')\right) \leq \sum_{x' \in \mathbb{Q}^d \setminus \text{supp}(P_X)} P_X(B_{r(x')/2}(x')) = 0$$

as required.

Exercise 1 Here we give an alternative proof that $\overline{X}_n$ is admissible in a Gaussian model with squared loss. Let δ have $R(\theta, \delta) \leq 1/n$ for all θ, with strict inequality for some θ_0. We wish to obtain a contradiction. By continuity of $\theta \mapsto R(\theta, \delta)$ we can find $\epsilon > 0$ and $\theta_1 > \theta_0$ such that $R(\theta, \delta) < 1/n - \epsilon$ for all $\theta \in (\theta_0, \theta_1)$.

For $\tau > 0$ consider the prior $\pi_\tau = N(0, \tau^2)$.

1. Show that for the π_τ-Bayes estimator δ_τ,

$$\frac{\frac{1}{n} - r(\pi_\tau, \delta)}{\frac{1}{n} - r(\pi_\tau, \delta_\tau)} = \frac{\int (\frac{1}{n} - R(\theta, \delta)) \frac{1}{\sqrt{2\pi}\tau} \exp(-\theta^2/2\tau^2) d\theta}{\frac{1}{n} - \frac{1}{n + \tau^{-2}}}$$

2. Show that as $\tau \to \infty$, this fraction converges to ∞ and deduce a contraction.

Solution 1

1. The numerator is obvious from the definition of Bayes risk as an integral of the risk function. The denominator follows from the formula for the Bayes risk of a Bayes estimator from a previous exercise.

2. Since the integrand is nonnegative and $\geq \epsilon$ on $[\theta_0, \theta_1]$, it is bounded below by

$$\epsilon \frac{1}{\tau\sqrt{2\pi}} \int_{\theta_0}^{\theta_1} \exp(-\theta^2/2\tau^2) d\theta$$

and as $\tau \to \infty$ the integral converges to the finite value $\theta_1 - \theta_0$. Therefore, for τ sufficiently large

$$\frac{\frac{1}{n} - r(\pi_\tau, \delta)}{\frac{1}{n} - r(\pi_\tau, \delta_\tau)} \geq \frac{n\tau^2(n + \tau^{-2})}{\tau\sqrt{2\pi}} 2\epsilon(\theta_1 - \theta_0) \to \infty, \qquad \tau \to \infty.$$

Therefore, for τ large we have $r(\pi_\tau, \delta_\tau) > r(\pi_\tau, \delta)$, but this is impossible since δ_τ is a Bayes estimator and thus its Bayes risk cannot exceed the Bayes risk of any other estimator δ.

Exercise 2 This problem considers minimaxity in nonparametric classes of distributions with squared loss.

1. Let $\mathcal{F}$ be the class of distributions with variance bounded by 1. Suppose we are interested in the mean $\mu = \mu(F)$. Show that $\overline{X}_n$ is minimax for the estimation of μ.

2. Let $\mathcal{F}$ be the class of all distributions on $[0,1]$. Find a minimax estimator for the mean $\mu = \mu(F)$. *Hint: we have a candidate from the previous exercise set. Show that it is indeed minimax. Write .*

Solution 2

1. The risk of $\overline{X}_n$ is

$$R(F, \overline{X}_n) = \mathbb{E}_F(\overline{X}_n - \mu(F))^2 = \mathrm{var}_F(\overline{X}_n) = \tfrac{1}{n}\mathrm{var}_F(X_1)$$

whose supremum over $\mathcal{F}$ is $1/n$.

We have seen that the supremum risk of any other estimator δ on the smaller class of normal distributions with unit variance is at least $1/n$. Therefore the supremum risk of δ on the whole class $\mathcal{F}$ is at least $1/n$, which is the maximal risk of $\overline{X}_n$. Thus $\overline{X}_n$ Is minimax.

2. We have seen that

$$\delta(\vec{X}) = \tfrac{2\sqrt{n}\overline{X}_n + 1}{2 + 2\sqrt{n}}$$

is minimax under the smaller class of binomial distributions. Let us see that the supremum risk is not larger when considered over the whole class $\mathcal{F}$. We have

$$R(F, \delta) = \mathbb{E}_F[\delta(\vec{X}) - \mu(F)]^2 = \mathrm{var}_F(\delta(\vec{X})) + bias_F^2(\delta(\vec{X})) = \tfrac{1}{(1+\sqrt{n})^2}\mathrm{var}_F(X_1) + \left(\tfrac{1-2\mu(F)}{2+2\sqrt{n}}\right)^2$$

$$= \tfrac{1}{(2+2\sqrt{n})^2}(4\mathbb{E}_F X_1^2 - 4\mu^2(F) + 1 - 4\mu(F) + 4\mu^2(F)) \leq \tfrac{1}{(2+2\sqrt{n})^2}(4\mu(F) + 1 - 4\mu(F)) = \tfrac{1}{(2+2\sqrt{n})^2}$$

where the inequality follows from $\mathbb{E}_F X_1^2 \leq \mathbb{E}_F X_1$, which is itself a consequence of $X_1 \in [0,1]$. The upper bound is the supremum risk of δ over the subclass of binomial distributions, where we know δ is minimax. Therefore it is minimax over the whole class $\mathcal{F}$.

Exercise 3 Let $g^* : \mathbb{R}^d \to \{0,1\}$ be the Bayes classifier.

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

2. Show that for any classifier $g : \mathbb{R}^d \to \{0,1\}$,
$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y).$$

3. For $\tilde{\eta}(x)$ and $\tilde{g}(x) = 1$ if $\tilde{\eta}(x) \geq 1/2$, prove that
$$\mathbb{P}(\tilde{g}(X) \neq Y) - \mathbb{P}(g^*(X) \neq Y) \leq 2\mathbb{E}|\eta(X) - \tilde{\eta}(X)|.$$

Solution 3

1. Denote P_X the marginal distribution of X and observe that
$$\begin{aligned}
\mathbb{P}\{g(X) \neq Y\} &= \int_{\mathbb{R}^d} \mathbb{P}\{g(x) \neq Y | X = x\}\, dP_X(x) = \int_{\mathbb{R}^d} 1_{\{g(x)=0\}}\eta(x) + 1_{\{g(x)=1\}}\{1-\eta(x)\}\, dP_X(x) \\
&= \int_{\mathbb{R}^d} \eta(x)\, dP_X(x) + \int_{\mathbb{R}^d} 1_{\{g(x)=1\}}\{1-2\eta(x)\}\, dP_X(x) \\
&\geq \int_{\mathbb{R}^d} \eta(x)\, dP_X(x) + \int_{\mathbb{R}^d} 1_{\{\eta(x)\geq 1/2\}}\{1-2\eta(x)\}\, dP_X(x) \\
&= \int_{\mathbb{R}^d} \min\{\eta(x), 1-\eta(x)\}\, dP_X(x) = \mathbb{P}\{g^*(X) \neq Y\}.
\end{aligned}$$

2. Since we need to minimise $\int 1_{g(x)=1}(1-2\eta(x))dP_X(x)$ over g, we can minimise the integrand pointwise in x. If $\eta(x) < 1/2$, the contribution of x can only be nonnegative, so it is best to choose $g(x) = 0$ so the indicator function eliminates the contribution of x. If $\eta(x) > 1/2$, the best is to choose $g(x) = 1$. For $\eta(x) = 1/2$ it does not matter what $g(x)$ is, and by convention we can choose it to be 1. Thus g^*, the Bayes classifier, is optimal, and any other optimal classifier is equal to g^* on the set $\{x : \eta(x) \neq 1/2\}$ P_X-almost surely.

3. Now let $\tilde{g}(x) = 1_{\{\tilde{\eta}(x)\geq 1/2\}}$. Then
$$\begin{aligned}
\mathbb{P}\{\tilde{g}(X) \neq Y\} - \mathbb{P}\{g^*(X) \neq Y\} &= \int_{\mathbb{R}^d} \{1_{\{\tilde{\eta}(x)\geq 1/2\}} - 1_{\{\eta(x)\geq 1/2\}}\}\{1-2\eta(x)\}\, dP_X(x) \\
&= \int_{\mathbb{R}^d} \{1_{\{\tilde{\eta}(x)\geq 1/2\}}1_{\{\eta(x)< 1/2\}} - 1_{\{\tilde{\eta}(x)< 1/2\}}1_{\{\eta(x)\geq 1/2\}}\}\{1-2\eta(x)\}\, dP_X(x) \\
&\leq 2\int_{\mathbb{R}^d} 1_{\{\tilde{\eta}(x)\geq 1/2\}}1_{\{\eta(x)< 1/2\}}\{\tilde{\eta}(x) - \eta(x)\} \\
&\qquad + 1_{\{\tilde{\eta}(x)< 1/2\}}1_{\{\eta(x)\geq 1/2\}}\}\{\eta(x) - \tilde{\eta}(x)\}\, dP_X(x) \\
&\leq 2\int_{\mathbb{R}^d} |\tilde{\eta}(x) - \eta(x)|\, dP_X(x).
\end{aligned}$$

Exercise 4 Denote the probability measure for X by P_X. Fix $x \in \text{supp}(P_X) \in \mathbb{R}^d$ 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 $k/n \to 0$.

Solution 4 Fix $\epsilon > 0$. Since $x \in \text{supp}(P_X)$, we have $P_X(B_\epsilon(x)) > 0$. If $\|X_{(k)}(x) - x\| > \epsilon$ then

$$\tfrac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \in B_\epsilon(x)\}} \leq k/n.$$

The event Ω, that the left-hand side converges to $P_X(B_\epsilon(x))$ for all $\epsilon = 1/m$ and m integer, has probability one. Since $k/n \to 0$, on Ω it holds that for all $m \in \mathbb{N}$,

$$\tfrac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \in B_{1/m}(x)\}} - k/n \to P_X(B_{1/m}(x)) - 0 > 0.$$

Therefore almost surely, for all m and all $n > N_m$, $\|X_{(k)}(x) - x\| \leq 1/m$. Hence $\|X_{(k)}(x) - x\| \to 0$ with probability one.

Now suppose $X_0 \sim P_X$. We have that $\mathbb{P}\{X_0 \in \text{supp}(P_X)\} = 1$. Now

$$\mathbb{P}(\|X_{(k)}(X_0) - X_0\| \to 0) = \mathbb{E}_{X_0}[\mathbb{P}(\|X_{(k)}(X_0) - X_0\| \to 0)|X_0] = 1$$

by the first part of the question, since the conditional expectation is equal to 1 for $X_0 \in \text{supp}(P_X)$, namely P_{X_0}-almost surely.

Exercise 5 Here we give an alternative argument that $\mathbb{P}(\|X_{(k)}(X) - X\| > \delta) \to 0$ for all $\delta > 0$ for the k-nearest neighbour classifier when $k/n \to 0$ and $k \to \infty$. Let $U_{(k)}$ be the k-th order statistic of independent $U_1, \ldots, U_n \sim [0,1]$. Using that $U_{(k)}$ has mean $k/(n+1)$ and variance $k(n-k+1)/[(n+1)^2(n+2)]$, show that

$$\mathbb{P}\left(U_{(k)} > \tfrac{2k}{n}\right) \to 0.$$

For $x \in \operatorname{supp}(P_X)$ define $F_x(t) = \mathbb{P}(\|X_1 - x\| \leq t)$. Let F_x^{-1} denote the corresponding quantile function. Show that $\lim_{s \searrow 0} F_x^{-1}(s) = 0$. Deduce that $\mathbb{P}(\|X_{(k)}(x) - x\| > \delta) \to 0$ for all $\delta > 0$. Deduce further that $\mathbb{P}(\|X_{(k)}(X) - X\| > \delta) \to 0$, where X is independent of the sequence $X_1, \ldots$ and has the same distribution as X_1.

Solution 5 By Chebychev's inequality

$$\mathbb{P}\left(U_{(k)} > \tfrac{2k}{n}\right) \leq \mathbb{P}\left(U_{(k)} - \tfrac{k}{n+1} > \tfrac{k}{n}\right) \leq \tfrac{n^2 k(n-k+1)}{k^2(n+1)^2(n+2)} \leq \tfrac{1}{k} \to 0.$$

Since $x \in \operatorname{supp}(P_X)$ we have $F_x(t) > 0$ for all $t > 0$. Let $s_m = F_x(1/m) > 0$. Then $F_x^{-1}(s_m) \leq 1/m \to 0$. Thus $F^{-1}(s) \to 0$ as $s \searrow 0$.

By the probability transform, $\|X_{(k)}(x) - x\|$ has the same distribution as $F_x^{-1}(U_{(k)})$. For n large $2k/n < F_x(\delta)$ and therefore

$$\mathbb{P}(\|X_{(k)}(x) - x\| > \delta) = \mathbb{P}(F_x^{-1}(U_{(k)}) > \delta) = \mathbb{P}(U_{(k)} > F_x(\delta)) \leq \mathbb{P}\left(U_{(k)} > \tfrac{2k}{n}\right) \to 0.$$

Taking expectation over X now gives $\mathbb{P}(\|X_{(k)}(X) - X\| > \delta) \to 0$ by the dominated convergence theorem, as the sequence of functions $x \mapsto \mathbb{P}(\|X_{(k)}(x) - x\| > \delta)$ converges to 0 P_X-almost surely and is bounded in absolute value by one.