Midterm review

18.466 review session, p. 0

Let $\Theta$ be a parameter space and $A$ an action space. Let $L(\theta, a)$ be a loss function on $\Theta \times A$ (jointly measurable). Let $X$ be a sample space. A statistical decision rule is a measurable function $a(\cdot)$ from $X$ into $A$. Its <u>risk</u> for a given $\theta$ is

$$r(\theta, a(\cdot)) = E_\theta L(\theta, a(X)) = \int L(\theta, a(x))\, dP_\theta(x).$$

$a(\cdot)$ is <u>inadmissible</u> if there is a decision rule $b(\cdot)$ such that $r(\theta, b(\cdot)) \le r(\theta, a(\cdot))$ for all $\theta$ and $r(\theta, b(\cdot)) < r(\theta, a(\cdot))$ for some $\theta$. Otherwise, $a(\cdot)$ is <u>admissible</u>.

If there is a prior $\pi(\cdot)$ given on $\Theta$, the decision rule $a(\cdot)$ is <u>Bayes</u> for $\pi$ if the risk

$$r(\pi, a(\cdot)) := \int r(\theta, a(\cdot))\, d\pi(\theta)$$

is finite and achieves the minimum among all decision rules,

$$r(\pi, a(\cdot)) \le r(\pi, b(\cdot))$$

for all decision rules $b(\cdot)$.

18.466 Review session, Monday, March 31, 2003

Priorities: know definitions and statements of main facts. Then, possibly, other facts and something about proofs.

Given two laws $P$ and $Q$ on the same sample space $(X, B)$, a likelihood ratio $R_{Q/P}(x)$ is always defined with $0 \le R_{Q/P}(x) \le +\infty$. There is always a measure $\nu$ such that $P$ and $Q$ both have densities with respect to $\nu$, e.g. $\nu = P + Q$, with $f = dP/d\nu$, $g = dQ/d\nu$, $0 \le f < \infty$, $0 \le g < \infty$, $R_{Q/P}(x) = g(x)/f(x)$. $R_{Q/P}$ doesn't depend on the choice of $\nu$ (Appendix A).

1.1.3 <u>Neyman-Pearson Lemma</u> For any such $P \ne Q$, a test of $P$ vs. $Q$ is inadmissible unless there is some $c$ with $0 \le c \le \infty$ such that the test chooses $Q$ if $R_{Q/P}(x) > c$ or $R_{Q/P}(x) = +\infty$, and chooses $P$ if $R_{Q/P}(x) < c$ or $R_{Q/P}(x) = 0$.

Let $L_{\mu\rho}$ be the loss when $\mu$ is true and $\rho$ is chosen. If losses and priors are given, with $L_{PP} = L_{QQ} = 0$, $L_{PQ} > 0$, $L_{QP} > 0$, $0 < \pi(P) = 1 - \pi(Q) < 1$, then (<u>Theorem 1.1.8</u>) for a Bayes test,

$$c = \frac{\pi(P) L_{PQ}}{\pi(Q) L_{QP}}.$$

If a family $\{P_\theta, \theta \in \Theta\}$ of laws on a sample space is dominated by a $\sigma$-finite measure $\nu$, we have a likelihood function $f(\theta, x) = \frac{dP_\theta}{d\nu}(x)$.

If $\pi$ is a prior distribution on $\Theta$ and $x$ is observed, then a <u>posterior distribution</u> $\pi_x$ exists if and only if $0 < \int f(\theta, x)\, d\pi(\theta) < \infty$, and then it has a density with respect to the prior $\pi$ given by
$$\frac{d\pi_x}{d\pi}(\varphi) = \frac{f(\varphi, x)}{\int f(\theta, x)\, d\pi(\theta)}.$$

<u>Theorem 1.3.7</u> says that (if $f$ is jointly measurable) then for $\pi$-almost all $\theta$ and $P_\theta$-almost all $x$, the posterior $\pi_x$ exists.

---

§§1.5–1.7 Given two laws $P \neq Q$ on $(X, \mathcal{B})$, let $X_1, X_2, \ldots,$ be i.i.d. $\mu = P$ or $Q$. Let $f(x) = R_{Q/P}(x)$, so $R_{Q^n/P^n}(X_1, \ldots, X_n) \equiv r_n \equiv \prod_{j=1}^n f(X_j)$. If $0 \leq A \leq B \leq +\infty$, SPRT$(A, B)$ is the test which makes a choice between $P$ or $Q$ after $N$ observations, the least $n$ such that $r_n \leq A$ or $r_n \geq B$, and then chooses $P$ if $r_n \leq A$, or $Q$ if $r_n \geq B$ and $r_n > A$. Usually $0 < A < 1 < B < +\infty$. We define $r_0 := 1$.

Theorem 1.5.1. Let $0 < A < 1 < B < +\infty$ and $\psi = SPRT(A, B)$. Let $\phi$ be any other sequential test. Suppose that the error probabilities $\alpha(P, \phi) \le \alpha(P, \psi)$ and $\alpha(Q, \phi) \le \alpha(Q, \psi)$. Then $E_{P,\psi} N \le E_{P,\phi} N$ and $E_{Q,\psi} N \le E_{Q,\phi} N$.

The theorem doesn't involve losses $L_{\mu\rho}$, costs $c$ per observation, or priors $p = \pi(P)$, but the proof involves all these quantities, which can be chosen for convenience.

Theorem 1.7.7 For any given losses $> 0$, cost $c > 0$, and prior $0 < p < 1$, some SPRT gives a Bayes test (how to choose $A$ and $B$ is given there but need not be memorized).

Lemma 1.7.10 For given $0 < A < 1 < B < \infty$ and $\varepsilon > 0$, there exist $c > 0$, $L_{PQ} > 0$ and $L_{QP} > 0$ such that SPRT$(A, B)$ is a Bayes test for a $p$ with $0 < p < \varepsilon$; or for other $c$, $L_{PQ}$, $L_{QP}$, with $1 - \varepsilon < p < 1$.

---

2.1 Sufficient statistics

Let $P_\theta$, $\theta \in \Theta$, be a family of laws. A statistic $T$ is *pairwise sufficient* for the family if for every $\varphi$ and $\theta$, $RP_\varphi / P_\theta$ can be written as a (measurable) function of $T$.

If $\{P_\theta, \theta \in \Theta\}$ is dominated by a $\sigma$-finite measure $\nu$, so that we have a likelihood function $f(\theta, x) = \frac{dP_\theta}{d\nu}(x)$, then we have the factorization theorem (Corollary 2.1.5, (b') $\Longleftrightarrow$ (c')): $T$ is pairwise sufficient if and only if we can write $f(\theta, x) = G(\theta, T(x))h(x)$ for some measurable function $h(x)$ not depending on $\theta$.

"If" is rather easy to prove since

$$R_{P_\varphi/P_\theta}(x) = \frac{f(\varphi, x)}{f(\theta, x)} = \frac{G(\varphi, T(x))}{G(\theta, T(x))}\frac{\cancel{h(x)}}{\cancel{h(x)}},$$

a function of $T(x)$, noting that $P_\psi(h(x) = 0) = P_\psi(h(x) = +\infty) = 0$ for all $\psi$.

---

2.2 Unbiased estimators: $E_\theta T(x) = g(\theta)$ for all $\theta$.

Can be inadmissible, for squared-error loss

2.2.2 Rao-Blackwell theorem For a convex loss function and a sufficient statistic $S$, a non-randomized decision rule $U$ with values in an action space $A$ which is a convex subset of some $\mathbb{R}^k$ with $E_\theta \|U\| < \infty$ for all $\theta$ can be replaced by $E(U|S) = E_\theta(U|S)$ for all $\theta$, without increasing the risk for any $\theta$.

§2.3 A Lehmann-Scheffé statistic $T$ is one such that if $E_\theta(f(T(x))) = E_\theta(h(T(x)))$ for all $\theta$, then $P_\theta(f(T(x)) = h(T(x))) = 1$ for all $\theta$. (Unbiased estimators which are functions of $T$ are a.s. unique.)

2.3.3 A LS sufficient stat. is minimal suff.

2.4.3 Information inequality Let $T$ be an unbiased estimator of $g(\theta)$, $\theta \in \Theta$, an open interval. For a likelihood function $f(\theta, x)$, the Fisher information is $I(\theta) := E_\theta\left(\left(\frac{\partial \log f(\theta,x)}{\partial\theta}\right)^2\right)$.

If $\frac{\partial f}{\partial \theta}(\theta, x)$ exists for $P_\theta$ - almost all $x$, $f(\theta,x) = \frac{dP_\theta}{d\nu}(x)$, assume $I(\theta) > 0$,

$$\int (|T(x)|+1)\left|\frac{\partial f(\theta,x)}{\partial\theta}\right| d\nu(x) < \infty,$$

$$\int \frac{\partial f(\theta,x)}{\partial\theta}\, d\nu(x) = 0,$$

and $g'(\theta) = \int T(x) \frac{\partial f(\theta,x)}{\partial\theta}\, d\nu(x)$.

Then $\operatorname{var}_\theta T \geq g'(\theta)^2 / I(\theta)$.

2.4.10. Let $I_1(\theta) = I(\theta)$ (one observation). Then for $n$ i.i.d. observations we have $I(\theta) = I_n(\theta) = n I_1(\theta)$.

2.4.12 If $E_\theta T(x) = g(\theta) + b(\theta)$, so $T$ has a bias $b(\theta)$ as an estimator of $g(\theta)$, then $\forall \theta$

$$E_\theta((T - g(\theta))^2) \geq \frac{(g+b)'(\theta)^2}{I(\theta)} + b(\theta)^2.$$

2.5 An exponential family has a likelihood function, for $\theta \in \Theta \subset \mathbb{R}^k$,

$$f(\theta, x) = C(\theta) h(x) e^{\theta \cdot T(x)}$$

where the natural parameter space $\Theta$ is convex and has non-empty interior and $T_1(x), \ldots, T_k(x)$ are affinely independent.

For any $n = 1, 2, \ldots,$ and i.i.d. observations
$x_i$ for $i = 1, \ldots, n,$
$\{\sum_{i=1}^n T_j(x_i)\}_{j=1}^k$ is a k-dimensional
sufficient statistic.

2.6. For $N(\mu, I)$ on $\mathbb{R}^3$, $x$ (the identity function) is an inadmissible estimator of $\mu$ for squared-error loss $\|T-\mu\|^2$; a better estimator is $T(x) := (1 - \frac{1}{|x|^2})x.$

3.1.1 For estimation of $g(\theta) \in \mathbb{R}^d$ with squared-error loss and a given prior $\pi(\theta)$, if there is an estimator with finite risk, then there is a Bayes estimator minimizing the risk, and it is

$$T(x) = \int g(\theta)\, d\pi_x(\theta)$$

(expectation of $g$ with respect to the posterior dist.).

3.2.2 In an exponential family, a maximum likelihood estimate in the interior of the natural parameter space is the same as a solution of the likelihood equation(s) and then is unique, but may not exist.

§3.3 Gave Solari's example of a situation where the likelihood equations have two solutions but neither gives a maximum.