Infotheory for Statistics and Learning

Lecture 4

- Binary hypothesis testing [PW:14],[CT:11.7]
- The Neyman–Pearson lemma [PW:14]
- General theory [PW:28]
- Bayes and minimax [PW:28.3]
- The minimax theorem [PW:28.3]

Binary Hypothesis Testing

Consider P and Q on $(\Omega, \mathcal{A})$

One of P and Q is the correct measure, i.e. the probability space is either $(\Omega, \mathcal{A}, P)$ or $(\Omega, \mathcal{A}, Q)$

Based on observation ω we wish to decide P or Q, hypotheses $H_0 : P$ and $H_1 : Q$

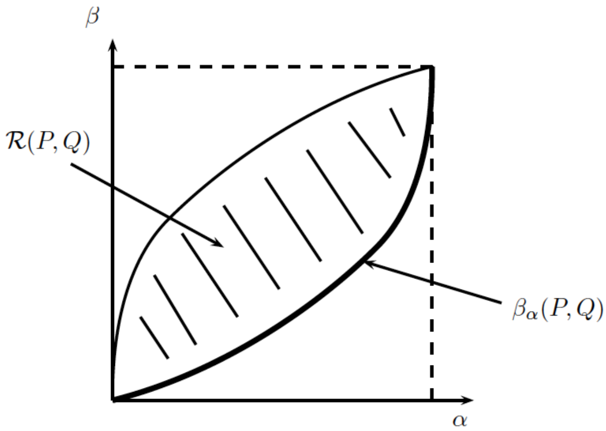
A decision kernel $P_{Z|\omega}$ for $Z \in \{0, 1\}$; $Z = 0 \to H_0$, $Z = 1 \to H_1$

Define $P_Z = P_{Z|\omega} \circ P$, $Q_Z = P_{Z|\omega} \circ Q$ and

$$\alpha = P_Z(\{0\}), \quad \beta = Q_Z(\{0\}), \quad \pi = Q_Z(\{1\})$$

Tradeoff between α (correct negative) and β (false negative)

$\pi = 1 - \beta$ power of the test (correct positive)



Define

$$\beta_\alpha(P,Q) = \inf_{P_{Z|\omega}: P_Z(\{0\}) \geq \alpha} Q_Z(\{0\})$$

and

$$\mathcal{R}(P,Q) = \bigcup_{P_{Z|\omega}} \{(\alpha,\beta)\}$$

Note that $(\alpha,\beta) \in \mathcal{R}(P,Q) \iff (1-\alpha, 1-\beta) \in \mathcal{R}(P,Q)$

Bounds on $\mathcal{R}(P,Q)$

Binary divergence for $0 \leq x \leq 1$, $0 \leq y \leq 1$,

$$d(x\|y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$$

Then if $(\alpha,\beta) \in \mathcal{R}(P,Q)$

$$d(\alpha\|\beta) \leq D(P\|Q), \quad d(\beta\|\alpha) \leq D(Q\|P)$$

Also, for all $\gamma > 0$ and $(\alpha,\beta) \in \mathcal{R}(P,Q)$

$$\alpha - \gamma\beta \leq P\left(\left\{\log \frac{dP}{dQ} > \log \gamma\right\}\right)$$
$$\beta - \frac{\alpha}{\gamma} \leq Q\left(\left\{\log \frac{dP}{dQ} < \log \gamma\right\}\right)$$

Neyman–Pearson Lemma

Define the log-likelihood ratio (LLR),

$$L(\omega) = \log \frac{dP}{dQ}(\omega)$$

For any α, $\beta_\alpha(P, Q)$ is achieved by the LLR test

$$P_{Z|\omega}(\{0\}|\omega) = \begin{cases} 1 & L(\omega) > \tau \\ \lambda & L(\omega) = \tau \\ 0 & L(\omega) < \tau \end{cases}$$

where τ and $\lambda \in [0, 1]$ solve

$$\alpha = P(\{L > \tau\}) + \lambda P(\{L = \tau\})$$

$\Rightarrow L(\omega)$ is a sufficient statistic for $\{H_i\}$

$\Rightarrow \mathcal{R}(P, Q)$ is closed and convex, and

$$\mathcal{R}(P, Q) = \{(\alpha, \beta) : \beta_\alpha(P, Q) \le \beta \le 1 - \beta_{1-\alpha}(P, Q)\}$$

We have implicitly assumed $P \ll Q$ (and $Q \ll P$), if this is not the case we can define $F = \cup\{A \in \mathcal{A} : Q(A) = 0 \text{ while } P(A) > 0\}$

Then set $P_{Z|\omega}(\{0\}|\omega) = 1$ on F and use the LLR test on F^c

In the extreme $P(F) = 1$ we can set $P_{Z|\omega}(\{0\}|\omega) = 1$ on F and $P_{Z|\omega}(\{0\}|\omega) = 0$ on F^c, to get

$$\alpha = P(F) = 1 \quad \text{and} \quad \beta = Q(F) = 0$$

the test is singular, $P \perp Q$

Proof of optimality

Let $g(\omega) = P_{Z|\omega}(\{0\}|\omega)$ for any $P_{Z|\omega}$ such that $E_P[g(\omega)] \geq \alpha$

Let

$$f(\omega) = \begin{cases} 1 & L(\omega) > \tau \\ \lambda & L(\omega) = \tau \\ 0 & L(\omega) < \tau \end{cases}$$

and $t = \exp(\tau)$, where τ and λ are chosen so that $\alpha = E_P[f(\omega)]$

Note that

$$(f(\omega) - g(\omega))\left(\frac{dP}{dQ}(\omega) - t\right) \geq 0$$

Hence

$$t\int (f-g)dQ \leq \int (f-g)dP \leq 0$$

$\Rightarrow E_Q[g(\omega)] \geq E_Q[f(\omega)]$

With probabilities on $\{H_i\}$: $\Pr(H_1 \text{ true}) = p$, $\Pr(H_0 \text{ true}) = 1 - p$

Let $g(\omega) = P_{Z|\omega}(\{0\}|\omega)$, then the average probability of error

$$P_e = (1-p)\left(1 - \int g(\omega)dP\right) + p\int g(\omega)dQ$$
$$= \int g(\omega)\left(p - (1-p)\frac{dP}{dQ}(\omega)\right)dQ + 1 - p$$

Thus the LLR test is optimal also for minimizing P_e, with

$$\tau = \log \frac{p}{1-p}$$

and with $\lambda \in [0,1]$ arbitrary (e.g. $\lambda = 1 - p$)

For the total variation between P and Q, we have

$$\mathsf{TV}(P,Q) = \sup_{E\in\mathcal{A}} (P(E) - Q(E))$$
$$= \sup_{E\in\mathcal{A}} \left\{ \int_E \left(\frac{dP}{dQ}(\omega) - 1 \right) dQ \right\}$$

achieved by $E = \{\omega : L(\omega) > 0\}$ (if $P \ll Q$)

Thus for the LLR test that minimizes P_e with $p = 1/2 \Rightarrow \tau = 0$ (and using $\lambda = 0$),

$$\mathsf{TV}(P,Q) = P(\{L(\omega) > 0\}) - Q(\{L(\omega) > 0\})$$
$$= \alpha - \beta_\alpha(P,Q) = 1 - 2P_e$$

$\Rightarrow P_e = (1 - \mathsf{TV}(P,Q))/2$

For $P \perp Q$, $E = F = \cup\{A \in \mathcal{A} : Q(A) = 0$ while $P(A) > 0\}$,

$$\mathsf{TV}(P,Q) = P(F) - Q(F) = 1 \quad \text{and} \quad P_e = 0$$

General Decision Theory

Given $(\Omega, \mathcal{A}, P)$ and assume $(E, \mathcal{E})$ is a standard Borel space (i.e., there is a topology $\mathcal{T}$ on E, $(E, \mathcal{T})$ is Polish, and $\mathcal{E} = \sigma(\mathcal{T})$)

$X : \Omega \to E$ is measurable if $\{\omega : f(\omega) \in F\} \in \mathcal{A}$ for all $F \in \mathcal{E}$

A measurable X is a random

- variable if $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$
- vector if $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}^n)$
- sequence if $(E, \mathcal{E}) = (\mathbb{R}^\infty, \mathcal{B}^\infty)$
- object in general

Let T be arbitrary, but typically $T = \mathbb{R}$

Denote $E^T = \{$functions from T to $E\}$, then X is a random

- process if $(E, \mathcal{E}) = (\mathbb{R}^T, \mathcal{B}^T)$

Given $(\Omega, \mathcal{A}, P)$, $(E, \mathcal{E})$ and $X : \Omega \to E$ measurable

For a general parameter set Θ let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a set of possible distributions for X on $(E, \mathcal{E})$

Assume we observe $X \sim P_\theta$ (i.e. P_θ is the correct distribution), and we are interested in knowing $T(\theta)$, for some $T : \Theta \to F$

A decision rule is a kernel $P_{\hat{T}|X=x}$ such that $P_{\hat{T}} = P_{\hat{T}|X} \circ P_X$ on $(\hat{F}, \hat{\mathcal{F}})$ (for $(\hat{F}, \hat{\mathcal{F}})$ standard Borel, typically $\hat{F} = F = \mathbb{R}$ and $\hat{\mathcal{F}} = \mathcal{B}$)

Define a loss function $\ell : F \times \hat{F} \to \mathbb{R}$ and the corresponding risk

$$R_\theta(\hat{T}) = \int \left\{ \int \ell(T(\theta), \hat{T}) dP_{\hat{T}|X=x} \right\} dP_\theta = E_\theta[\ell(T, \hat{T})]$$

Bayes Risk

Assume $\Theta = \mathbb{R}$ and $T(\theta) = \theta$ (for simplicity)

Postulate a prior distribution π for θ on $(\mathbb{R}, \mathcal{B})$

The average risk

$$R_\pi(\hat{\theta}) = \int R_\theta(\hat{\theta}) d\pi = \int \left\{ \int \ell(\theta, \hat{\theta}) d(P_{\hat{\theta}|X} \circ P_\theta) \right\} d\pi$$

and the Bayes risk

$$R^*_\pi = \inf_{P_{\hat{\theta}|X}} R_\pi(\hat{\theta})$$

achieved by the Bayes estimator $P^*_{\hat{\theta}|X=x}$

Define $P_{\theta|X}$ from $\pi = P_{\theta|X} \circ P_\theta$, then since $\theta \to X \to \hat{\theta}$

$$E_\pi \left[\int \left\{ \int \ell(\theta, \hat{\theta}) dP_{\hat{\theta}|X=x} \right\} dP_\theta \right]$$
$$= \int \left\{ \int \left\{ \int \ell(\theta, \hat{\theta}) dP_{\hat{\theta}|X=x} \right\} dP_{\theta|X=x} \right\} d(P_\theta \circ \pi)$$

Hence we can define $\hat{\theta}(x)$ via $\ell(\theta, \hat{\theta}(x)) = \int \ell(\theta, \hat{\theta}) dP_{\hat{\theta}|X=x}$ and for each $X = x$ minimize

$$\int \ell(\theta, \hat{\theta}(x)) dP_{\theta|X=x}$$

$\Rightarrow$ the Bayes estimator is always deterministic

- Thus we can always work with $\hat{\theta}(x)$ instead of $P_{\hat{\theta}|X}$
- Can also be proved more formally from the fact that $R_\pi(\hat{\theta})$ is linear in $P_{\hat{\theta}|X}$ and the set $\{P_{\hat{\theta}|X}\}$ is convex

Data processing inequality

Given a prior distribution π for θ, assume that

$$\theta \to X \to Y$$

and let $R_\pi^*(X)$ denote the Bayes risk based on observing X, and similarly $R_\pi^*(Y)$ based on Y. Then

$$R_\pi^*(X) \le R_\pi^*(Y)$$

Proof Define

$$f(x, u) = \sup\{v \in [0, 1] : P_{Y|X=x}([0, v]) < u\}$$

Let $U \sim \mathcal{U}([0, 1])$ and independent of X, then $f(x, U) \sim P_{Y|X=x}$ and

$$R_\pi^*(X) = \inf_{\hat{\theta}(\cdot)} E[\ell(\theta, \hat{\theta}(X))] \le \inf_{u \in [0,1]} E[\ell(\theta, \tilde{\theta}(f(X, u)))]$$
$$\le E[\ell(\theta, \tilde{\theta}(f(X, U)))] = E[\ell(\theta, \tilde{\theta}(Y))] = R_\pi^*(Y)$$

where $\tilde{\theta}(Y)$ is the Bayes estimator based on Y.

Minimax Risk

Let

$$R^* = \inf_{P_{\hat{\theta}|X}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta}) = \inf_{P_{\hat{\theta}|X}} \sup_{\theta \in \Theta} \int \left\{ \int \ell(\theta, \hat{\theta}) dP_{\hat{\theta}|X=x} \right\} dP_\theta$$

denote the minimax risk

The problem is convex, and we can write

$$R^* = \inf t \quad \text{s.t.} \quad E_\theta[\ell(\theta, \hat{\theta})] \le t \quad \text{for all } \theta \in \Theta$$

over $P_{\hat{\theta}|X}$ and t

Assuming Θ is finite for simplicity, we get the Lagrangian

$$L(P_{\hat{\theta}|X}, t, \{\lambda(\theta)\}) = t + \sum_\theta \lambda(\theta)(E_\theta[\ell(\theta, \hat{\theta})] - t)$$

and the dual function $g(\{\lambda(\theta)\}) = \inf_{P_{\hat{\theta}|X}, t} L(P_{\hat{\theta}|X}, t, \{\lambda(\theta)\})$

Note that unless $\sum_\theta \lambda(\theta) = 1$, we get $g(\{\lambda(\theta)\}) = -\infty$

Thus $\sup g(\{\lambda(\theta)\})$ is attained for $\lambda(\theta) =$ a pmf on θ, and

$$\sup_{\{\lambda(\theta)\}} g(\{\lambda(\theta)\}) = \sup_{\{\lambda(\theta)\}} \inf_{P_{\hat{\theta}|X}} \sum_\theta \lambda(\theta) E_\theta[\ell(\theta, \hat{\theta})] = \sup_\pi R^*_\pi$$

with $\pi(\theta) = \lambda(\theta)$ is the worst-case Bayes risk

Because of weak duality, we always have

$$\sup_{\pi} R^*_\pi \leq R^*$$

and strong duality, i.e.

$$R^* = \sup_{\pi} R^*_\pi$$

holds if

- θ is finite and $\mathcal{X}$ is finite, or
- θ is finite and $\inf_{\theta,\hat{\theta}} \ell(\theta, \hat{\theta}) > -\infty$

and also under very general conditions (see [PW:28.3.4]...)

We have thus established the minimax theorem

When strong duality holds the minimax risk is obtained by a deterministic $\hat{\theta}(x)$