Lecture 14

Bayesian inference

So far we have tried to minimize $R(\theta,\delta)$ over all possible δ, and hope that will be uniform in θ.

Restricted to

<u>unbiased</u> estimators

$$\hat{\delta} = \underset{\delta \text{ unbiased}}{\operatorname{argmin}} \; R(\theta,\delta)$$

~~If~~ If

$$R(\theta,\hat{\delta}) \le R(\theta,\delta) \quad \forall \delta \text{ unbiased} \quad \forall \theta$$

we call $\hat{\delta}$ UMVU. <u>Exponential families</u> important.

<u>equivariant</u> estimators

$$\hat{\delta} = \underset{\delta \text{ equivariant}}{\operatorname{argmin}} \; R(\theta,\delta)$$

Risk does not depend on θ, so if

$$R(\theta,\hat{\delta}) \le R(\theta,\delta) \quad \forall \delta \text{ equivariant}$$

we call $\hat{\delta}$ MRE. Group families important

Now change focus, change risk to average risk

Treat θ as a random variable with a distribution Λ, and define the average risk of estimator δ

$$r(\Lambda, \delta) = \int R(\theta, \delta)\, d\Lambda(\theta).$$

Define, given a distribution Λ, the Bayes estimator

$$\hat{\delta} = \arg\min_{\delta} r(\Lambda, \delta) = \arg\min_{\delta} \int R(\theta, \delta)\, d\Lambda(\theta)$$

What does it mean? How do we use it?

(i) A mathematical tool

Way to solve problems, connection to minimax theory and minimax estimator. Solve problems otherwise intractable.

(ii) As a way to use past experience

Previous research/exploration/analysis suggests a distribution Λ for a parameter. Use this, gather data to update the distribution.

Ex: Coin toss, $\theta = p(\text{head})$.

View

θ also of (n) r.v., (u) $\sim \Lambda$.

(iii) As a description of a state of mind

"Hard-core Bayesianism".

A subjective "feeling about" θ. Use data X obs of $\mathbb{X}$ ($\mathbb{X}|\theta$ has a distribution we model), to update belief or feeling about θ

(4) $\theta \sim \Lambda$ assume a priori dist. of θ

$\mathbb{X}|\theta \sim$ model

$\Rightarrow$

$\Theta|\mathbb{X} \sim$ aposteriori dist. of θ.

(iv) As a method to generate estimators.

(i) simple prior Bayes $\theta \sim \Lambda(\lambda)$ ← hyperparameter fixed value

(ii) empirical Bayes

$\theta \sim \Lambda(\lambda)$ ← estimated from data

(iii) hierarchical Bayes

$\theta \sim \Lambda(\lambda)$ — hyperprior distribution.

$\lambda \sim \Gamma(\gamma)$ ← fixed parameter

(iv) Robust Bayes

make sure the estimator works well for each member of the prior class.

Some notation:

$\underline{X} \sim f(x|\theta)$ density of $\underline{X}$ given θ (p.d.f or p.m.f.)

(4) $\theta \sim \pi, \lambda$ density $\pi(\theta|\lambda), \gamma(\lambda)$

hyperparameter

(4) $\theta|x,\lambda$ $\lambda|x$ $\sim$ posterior distribution

$\pi(\theta|x,\lambda)$, $\gamma(\lambda|x)$ densities.

Marginal distributions used:

$\pi(\theta|x,\lambda) =$ (a posteriori distribution of (4) given data (sample) and hyperparameter λ)

treat λ as a fixed parameter and Bayes theorem to get this

$$= \frac{f(x|\theta)\pi(\theta|\lambda)}{m(x|\lambda)}$$

$\int f(x|\theta)\pi(\theta|\lambda)\,d\theta$ marginal distribution of x given λ, integrating out θ.

This means, if we forget about the hyperparameter λ

$$\text{“}\ \pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta}\ \text{”}$$

(4)

but everything depends on hyperparameter λ, made explicit above

How do we calculate the Bayes estimator?
Given a loss function $L(\theta, d)$.

<u>Before</u> data is sampled:

$$\Theta \sim \lambda$$

Risk is

$$E\, L(\Theta, d)$$

for estimating

$$g(\Theta)$$

Then

$$\hat{d} = \arg\min_{d} E(L(\Theta, d))$$

is Bayes estimator.

<u>After</u> data is sampled: x observed of $X \mid \Theta$

Prior is replaced by posterior distribution of

$$\Theta \mid x$$

The Bayes estimator is

$$\hat{\delta}(x) = \arg\min_{\delta} E(L(\Theta, \delta(x)) \mid x)$$

Theorem

Assume $\Theta \sim \Lambda$, $X \mid \Theta = \theta \sim P_\theta$. Loss function $L(\theta, d)$ non-negative for estimating $g(\Theta)$. Assume

a) There is an estimator with finite risk.

b) For almost all x (P_θ-a.s.),

$$\delta_\Lambda(x) = \arg\min_\delta E\left(L(\Theta, \delta(x)) \mid X = x\right)$$

exists.

Then $\delta_\Lambda(X)$ is a Bayes estimator.

Proof.

Assume δ is an arbitrary estimator with

$$\int R(\theta, \delta)\, d\Lambda(\theta) = \int E_\theta(L(\theta, \delta(X)))\, d\Lambda(\theta)$$

$$= E_\Lambda \int L(\Theta, \delta(x))\, dP_\theta(x)$$

$$= \int E_\Lambda(L(\Theta), \delta(x)) \mid X = x)\, dP_\theta(x) < \infty$$

(there is at least one such by a)). Since $L \geq 0$ and P_θ is a prob. measure, that means that

$$E_\Lambda(L(\Theta, \delta(x)) \mid X = x) < \infty$$

Use Bayes theorem and the posterior distribution

and thus by b)

$$E(L(\Theta, \delta(x)) \mid X = x) \geq E(L(\Theta, \delta_\Lambda(x)) \mid X = x)$$

P_θ-a.s. Do $\int \ldots dP_\theta$ of both sides to obtain

$$\int R(\theta, \delta) d\Lambda(\theta) \geq \int R(\theta, \delta_\Lambda) d\Lambda(\theta)$$

for all δ, i.e. δ_Λ is a Bayes estimator #

By taking a few particular loss functions we get the following special cases

<u>Corollary</u>

Assume the assumptions of the theorem hold.

(i) If $L(\theta, d) = (d - g(\theta))^2$ is quadratic loss then

$$\underline{\delta_\Lambda(x) = E(g(\Theta) \mid x)}$$

(If $L(\theta, d) = (d - g(\theta))^2 w(\theta)$ quadratic loss with weights then

$$\delta_\Lambda(x) = \frac{E(g(\Theta) w(\Theta) \mid x)}{E(w(\Theta) \mid x)} \Big)$$

(ii) If $L(\theta, d) = |d - g(\theta)|$ is absolute value loss then $\delta_\lambda(x)$ is (any) median of the conditional distribution of $\Theta \mid \underline{X} = x$.

(iii) If

$$L(\theta, d) = \begin{cases} 0 & |d-\theta| \le c \\ 1 & |d-\theta| > c \end{cases} = 1\{|d - \overset{g(\theta)}{\theta}| > c\}$$

then $\delta_\lambda(x)$ is the midpoint of interval I of length $2c$ which maximizes $P(\Theta \in I \mid x)$

#

<u>Proof.</u>

(i) We want to find $\delta_\lambda(x)$ that minimizes

$$E((g(\Theta) - \delta(x))^2 \mid x).$$

That is possible by the theorem and the minimizing value is (conditional L^2-projection...)

$$E(g(\Theta) \mid x)$$

(ii) Yes, and conditional median solves it

(iii) Yes, just use $E(1\{|d - \Theta| > c\} \mid x)$
$= P(|d - \Theta| > c \mid x)$ and change minimize this to maximize $P(|d - \Theta| \le c \mid x)$.

#

We can next state a result on uniqueness of the Bayes estimator.

<u>Corollary.</u>

Assume the loss function $L(\theta, d)$ is strictly convex in d. Assume

i) $\int R(\theta, \delta_\Lambda)\, d\Lambda(\theta)$ is finite

(ii) Let Q be the marginal distribution of X

$$Q(A) = \int P_\theta(X \in A)\, d\Lambda(\theta)$$

Then ~~with~~ ~~Q~~ ~~Q~~ Q-a.s. implies P_θ-a.s. for all θ.

~~Proof (read on your own, carefully)~~ #

Then the Bayes estimator δ_Λ is unique P_θ-a.s. for all θ.

<u>Proof</u> (read on your own, carefully) #

An example on when condition (ii) is <u>not</u> satisfied is ~~Note~~ Assume

<u>Ex</u>: ~~X~~ $X \sim Bin(n, p)$ distribution and Λ is two-point distribution: Λ puts mass $1/2$ on $p=0$ and mass $1/2$ on $p=1$. Let $\delta(X)$ be an arbitrary estimator such that $\delta(0)=0$, $\delta(n)=1$ (so $\delta(k)$ for $k=1,\dots,n-1$ does not matter). Assume we have squared

error loss. Then the Bayes risk is

$$\int R(\theta,\delta)\, d\Lambda(\theta)$$
$$= \frac{1}{2} R(0,\delta) + \frac{1}{2} R(1,\delta)$$
$$= \frac{1}{2} E_0\left((0-\delta(X))^2\right) + \frac{1}{2} E_p\left((1-\delta(X))^2\right)$$

But if $p=0$, $X=0$ with prob 1 and if $p=1$, $X=n$ with prob 1 so the above is

$$\cancel{= \frac{1}{2} E_0\left((0-\delta(0))^2\right) + \frac{1}{2} E_1(}$$
$$= \frac{1}{2}(0-\delta(0))^2 + \frac{1}{2}(1-\delta(n))^2$$

which is $=0$ if $\delta(X)$ is as above, and this is the smallest possible value. So then the Bayes estimator is <u>not unique.</u>

Furthermore the marginal distribution of X is

$$Q(A) = \cancel{P(X\in A)} = \int P_\theta(X\in A)\, d\Lambda(\theta)$$
$$= P_0(X\in A)\frac{1}{2} + P_1(X\in A)\frac{1}{2}$$

and the set $A=\{0,n\}$ has Q mass 1, but for instance

$\cancel{Q(A)=}$

if $p=1/2$ ~~that~~ ~~$\neq$~~

$$P_\theta(X\in A) = P_\theta(0) + P_\theta(n)$$
$$= \left(\tfrac{1}{2}\right)^n + \left(\tfrac{1}{2}\right)^n < 1$$

so the condition (ii) is <u>not</u> satisfied

#

How does one evaluate a Bayes estimator?
Different (several) risk functions:

(i) The Bayes risk

$$\int R(\theta, \delta)\, d\Lambda(\theta)$$

(ii) The conditional risk given $\Theta = \theta$, the previously studied risk

$$R(\theta, \delta)$$

(iii) The a posteriori risk

$$E(L(\Theta, \delta(X)) \mid X = x)$$

The Bayes risk is <u>the</u> criterion for the Bayes estimator.
(ii) can be used to make sure that one does not get "very good" results simply by choosing a "proper" prior Λ. (iii) is of use to the hard core Bayesian.

<u>Ex</u>: Assume $X \in Bin(n, p)$. Choose as prior distribution for p a Beta distribution $B(a, b)$ with density

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}; \qquad 0 < x < 1$$

and

$$E(p) = \frac{a}{a+b} \qquad Var(p) = \frac{ab}{(a+b)^2(a+b+1)}$$

We want to find an estimator (the Bayes estimator) for $g(p)$. We need to get conditional distribution of p given x. Joint density of X and p is (w.r.t. product of Lebesgue measure and counting measure)

$$\binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{x+a-1}(1-p)^{n-x+b-1}$$

$x = 0, 1, 2, \ldots, n$, $0 < p < 1$. Integrate out p to get the marginal density (pmf) of x; this will be a function of only x. Divide by this to get conditional density of p given x

$$C(a,b,x)\, p^{x+a-1}(1-p)^{n-x+b-1}$$

Recognize as Beta distribution with parameters

$$a' = a + x$$

$$b' = b + n - x$$

Assume we have a quadratic loss and $g(p) = p$. Then we have

$$\delta_\Lambda(x) = E(p|x) \overset{\text{Beta distri.}}{=} \frac{a'}{a'+b'} = \frac{a+x}{a+b+n}$$

$$= \underbrace{\left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b}}_{\text{Bayes estimator before observations taken}} + \underbrace{\left(\frac{n}{a+b+n}\right)\frac{x}{n}}_{\text{Standard estimator (UMVU etc.)}}$$

and the Bayes estimator is a weighted average of the two!

(i) $a \to \infty, b \to \infty$, $\frac{a}{b}$ fixed $\Rightarrow Var(p) \to 0$

all mass of p concentrates on $\frac{a}{a+b}$

so $\Lambda_{a,b} \to \delta_{\frac{a}{a+b}}$ ← Dirac measure

Then

$$\delta_\Lambda \to \frac{a}{a+b}$$ Bayes before data gathered so data has <u>no effect</u>.

(ii) a, b fixed, $n \to \infty$

$\Rightarrow \delta_\Lambda$ "$\to$" $\frac{x}{n}$ (rather $\frac{\delta_\Lambda(x)}{x/n} \to 1$)

and the a priori distribution has no effect on the Bayes estimator

#

Finally note that $a = b = 0$ gives

$$\delta_\Lambda(x) = \frac{x}{n}$$ (the ordinary estimator).

But $B(0,0)$ is not a proper distribution since then

$$\int_0^1 f_p(p)\,dp = \infty$$