ST308 - Lent term
Bayesian Inference

Kostas Kalogeropoulos

Statistical Decision Theory - Bayes Estimators

Summary of previous lecture

- Frequentist and subjective probability
- Setup of Frequentist and Bayesian Inference
- Examples: Conjugate models

Outline

Topics covered: Loss function, Frequentist, Posterior and Bayes Risk, Bayes (decision) Rules, Minimax criterion, Point Estimation, Bayes Estimators

Outline

1. Statistical Decision Theory
2. Point Estimation

Statistical Inference

- Collect or consider the data $x = (x_1, \ldots, x_n)$ from an experiment.
- Assign model-likelihood to real world problem with uncertainty.
- (In Bayesian Inference) Assign a prior to the parameters.
- Based on the above **decide** on
    - A best guess for θ - Point Estimation
    - A range of values for θ - Interval Estimation
    - Choosing between H_0 and H_1 - Hypothesis Testing
    - A best guess/range for a future value of x - Prediction

Statistical Decision Theory

Given observations x and model $f(x|\theta)$, a statistical decision problem consists of

1. The parameter space Θ.
2. A set $\mathcal{A}$ of all possible actions a.
3. A loss function $L(a, \theta) : \mathcal{A} \times \Theta \to \mathcal{R}$, reflecting the loss for action a and true parameter value θ.

Notes:

- The sets $\mathcal{A}$, Θ could be finite or infinite.
- The negative of $L(a, \theta)$ is called utility function.

Decision Rule and Frequentist Risk

Decision Rule

The function $\delta(x) : \mathcal{R} \to \mathcal{A}$ that indicates the action a after observing $X = x$.

Every decision rule is associated with a random risk.

Frequentist Risk

$$R(\delta(x), \theta) = E_{X|\theta}\left(L(\delta(x), \theta)\right) = \int L(\delta(x), \theta) f(x|\theta) dx$$

Posterior Risk

$$\rho(\delta(x), \theta) = E_{\theta|x}\left(L(\delta(x), \theta)\right) = \int L(\delta(x), \theta) \pi(\theta|x) d\theta$$

Using Frequentist Risk

How to choose between two decision rules, $\delta_1(x)$ and $\delta_2(x)$?

1. Choose a risk function, e.g. frequentist risk.
2. If $R(\delta_1(x), \theta) < R(\delta_2(x), \theta)$ for all $\theta \in \Theta$, then $\delta_1(x)$ is uniformly better than $\delta_2(x)$.

If there exists $\delta^*(x)$ such that $R(\delta^*(x), \theta) \leq R(\delta(x), \theta)$ for all $\delta(x) \in \mathcal{D}$ and all $\theta \in \Theta$, then $\delta^*(x)$ is an admissible decision rule.

Issue: It is usually very difficult to minimise frequentist risk for all $\theta \in \Theta$.

The Minimax criterion

Minimax criterion

A minimax estimator $\delta^M(x)$ satisfies

$$\max_{\theta \in \Theta} R(\delta^M(x), \theta) = \min_{\delta \in \mathcal{D}} \left[\max_{\theta \in \Theta} R(\delta(x), \theta) \right]$$

Notes:

- Optimisation under worst case scenario - too conservative.
- Very difficult to find.
- Does not use prior information.
- inf and sup may also be used.

Bayes Risk criterion

Bayes Risk

Given the prior $\pi(\theta)$ and $\delta(x)$, Bayes risk is the function

$$r(\delta(x), \pi(\theta)) = E_\theta\left[R(\delta(x), \theta)\right] = \int R(\delta(x), \theta)\pi(\theta)d\theta$$

The decision rule $\delta^B(x)$ that minimises the Bayes risk is called Bayes rule.

Note on Bayes Risk criterion

- Bayes risk is a number. Hence, given a loss function $L(a, \theta)$, $f(x|\theta)$ and $\pi(\theta)$ there exists an optimal solution to the statistical decision problem.

- Bayes risk can also be written as an expectation of the posterior risk wrt X

$$r(\delta(x), \pi(\theta)) = E_X\left[\rho(\delta(x), \pi(\theta))\right] = \int \rho(\delta(x), \pi(\theta)) m(x) dx$$

  where $m(x)$ is the marginal likelihood. It unifies the two approaches based on the frequentist and posterior risks.

Example: Vaccination problem

A public health organisation considers vaccination to prevent a disease. A test to determine immunity exists. Let X denote the test outcome (x_1: positive, x_2: negative) and θ whether the person is immune to the disease (θ_1: immune, θ_2: susceptible). The likelihood is given below

$f(x\|\theta)$	x_1	x_2
θ_1	0.65	0.35
θ_2	0.25	0.75

Example: Vaccination problem (cont'd)

Consider the actions regarding vaccination (a_1: yes, a_2: no). The loss function is

$L(a, \theta)$	a_1	a_2
θ_1	8	0
θ_2	0	20

Which of the 4 strategies (decision rules) is better?

1. Vaccinate everyone $\delta_1(x_1) = \delta_1(x_2) = a_1$
2. Vaccinate positives $\delta_2(x_1) = a_1$, $\delta_2(x_2) = a_2$
3. Vaccinate negatives $\delta_3(x_1) = a_2$, $\delta_3(x_2) = a_1$
4. Don't vaccinate anyone $\delta_4(x_1) = \delta_4(x_2) = a_2$

Example: Vaccination problem (cont'd)

$$R(\delta_1(x), \theta_1) = E_{X|\theta}(L(\delta_1(x), \theta_1)) = \sum_{i=1}^{2} L(\delta_1(x_i), \theta_1) f(x_i|\theta_1)$$
$$= L(a_1, \theta_1) f(x_1|\theta_1) + L(a_1, \theta_1) f(x_2|\theta_1) = ... = 8$$

Similarly we can get

$R(a, \theta)$	$\delta_1(x)$	$\delta_2(x)$	$\delta_3(x)$	$\delta_4(x)$
θ_1	8	5.2	2.8	0
θ_2	0	15	5	20

No admissible (optimal) strategy (decision rule).

Example: Vaccination problem (cont'd)

Suppose that the prior is $\pi(\theta_1) = 0.6$ and $\pi(\theta_2) = 0.4$.

$$r(\delta_1(x), \pi(\theta)) = E_\theta(R(\delta_1(x), \theta)) = \sum_{i=1}^{2} R(\delta_1(x), \theta_i)\pi(\theta_i)$$
$$= R(\delta_1(x), \theta_1)\pi(\theta_1) + R(\delta_1(x), \theta_2)\pi(\theta_2) = ... = 4.8$$

Similarly we can get $r(\delta_2(x), \pi(\theta)) = 9.12$, $r(\delta_3(x), \pi(\theta)) = 3.68$ and $r(\delta_4(x), \pi(\theta)) = 8$.

Hence, the optimal strategy (decision rule) is $\delta_3(x)$, to vaccinate those who are tested negative for immunity.

Outline

1. Statistical Decision Theory
2. Point Estimation

Point Estimation problem

- Collect or consider the data $x = (x_1, \ldots, x_n)$ from an experiment.
- Assign model-likelihood to real world problem with uncertainty.
- (In Bayesian Inference) Assign a prior to the parameters.
- Based on the above decide on a best guess for θ - Point Estimation.

Decision theory elements

- **Action:** report a value for the θ, action set $\mathcal{A} = \Theta$.
- **Decision rule:** $\delta(x)$ is an estimator also denoted with $\hat{\theta}$, e.g. $\bar{x}$, $\frac{1}{n}\sum_i x_i^2$ etc.
- **Loss function:** absolute error, quadratic error, 0-1 loss etc.

Note: In the case of quadratic error loss function $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, frequentist risk is the mean squared error (MSE)

$$R(\hat{\theta}, \theta) = E_{X|\theta}(\hat{\theta} - \theta)^2 = MSE(\hat{\theta})$$

Example: Partial ordering with frequentist risk

Let $x = (x_1, \ldots, x_n)$ be a random sample from $\mathrm{N}(\theta, \sigma^2)$. Consider the quadratic loss function, i.e. the MSE as frequentist risk.

Consider $\delta_1(x) = \bar{x}$, the minimum variance unbiased estimator for this problem, and the naive estimator $\delta_2(x) = 100$.

We get $R(\delta_1(x), \theta) = \frac{1}{n}\sigma^2$, $R(\delta_2(x), \theta) = (100 - \theta)^2$.

Note that $\delta_1(x)$ is not better than $\delta_2(x)$ for all $\theta \in \mathcal{R}$ in terms of MSE!

Bayes Estimators

Theorem (Construction of Bayes estimators)

Bayes estimators minimise the Bayes risk $r(\delta(x), \pi(\theta))$

This can be achieved if for every $x \in \mathcal{X}$ we select the value $\delta(x)$ that minimises the posterior risk $\rho(\delta(x), \pi(\theta))|x)$, since

$$r(\delta(x), \pi(\theta)) = \int \rho(\delta(x), \pi(\theta)|x)m(x)dx$$

Minimising the posterior risk is the same as minimising the Bayes risk.

Quadratic error loss function

Suppose that we have $L(a, \theta) = (a - \theta)^2$. The Bayes estimator minimises the posterior risk

$$\rho(a, \theta) = \int (a - \theta)^2 \pi(\theta|x) d\theta.$$

We can write

$$\frac{\partial \rho(a, \theta)}{\partial a} = \int \frac{\partial}{\partial a}(a - \theta)^2 \pi(\theta|x) d\theta = \int 2(a - \theta)\pi(\theta|x) d\theta$$
$$= 2\left\{ a \int \pi(\theta|x) d\theta - \int \theta \pi(\theta|x) \right\} d\theta = 2(a - E(\theta|x))$$

Setting $\frac{\partial \rho(a,\theta)}{\partial a} = 0$ gives $a = E(\theta|x)$. Also $\frac{\partial^2}{\partial \alpha^2} \rho(\alpha, \theta) = 2 > 0$.

Quadratic error loss $\longrightarrow$ Bayes estimator is the posterior mean.

Linear error loss function

Theorem: Assume that for positive k_0, k_1 the loss function is

$$L(a,\theta) = \begin{cases} k_0(a-\theta) & \text{if } a > \theta \\ k_1(\theta - a) & \text{if } a \leq \theta \end{cases}$$

The Bayes estimator is the $\frac{k_1}{k_0+k_1}$-th percentile of $\pi(\theta|x)$ denoted by q.

$$\frac{k_1}{k_0+k_1} = P(\theta \leq q|x) = \int_{-\infty}^{q} \pi(\theta|x)d\theta$$

Proof: We will show that $E_{\theta|x}(L(q,\theta)) \leq E_{\theta|x}(L(a,\theta))$. Assume $q < a$.

Linear error loss function (cont'd)

If $\theta \leq q < a$:

$$L(q,\theta) - L(a,\theta) = k_0(q-\theta) - k_0(a-\theta) = k_0(q-a)$$

If $q < a < \theta$:

$$L(q,\theta) - L(a,\theta) = k_1(\theta-q) - k_0(\theta-a) = k_1(a-q)$$

If $q < \theta < a$:

$$\begin{aligned} L(q,\theta) - L(a,\theta) &= k_1(\theta-q) - k_0(a-\theta) = k_1(\theta-q) + k_0(\theta-a) \\ &< k_1(\theta-q) < k_1(a-q) \end{aligned}$$

So putting everything together we get

$$L(q,\theta) - L(a,\theta) \leq \begin{cases} k_0(q-a) & \text{if } \theta \leq q \\ k_1(a-q) & \text{if } q < \theta \end{cases}$$

Linear error loss function (cont'd)

Taking expectation wrt to the posterior yields

$$
\begin{aligned}
E_{\theta|x}(L(q,\theta) - L(a,\theta)) &\leq k_0(q-a)P(\theta \leq q) + k_1(a-q)P(\theta > q) \\
&= k_0(q-a)\frac{k_1}{k_0+k_1} + k_1(a-q)\left(1 - \frac{k_1}{k_0+k_1}\right) = \ldots = 0
\end{aligned}
$$

So $E_{\theta|x}(L(q,\theta)) \leq E_{\theta|x}(L(a,\theta))$, i.e. q minimises the posterior risk. $\square$

Special case: For $k_0 = k_1 = 1$ we get the absolute error loss function

$$L(a,\theta) = |a - \theta|$$

Hence the $1/(1+1) = 0.5-$percentile, or else the posterior median minimises the posterior risk and therefore is the Bayes estimator.

$0-1$ loss function

Finally consider the $0-1$ loss function

$$L(a,\theta)=\begin{cases} 0 & \text{if } |a-\theta|\le\epsilon \\ 1 & \text{if } |a-\theta|>\epsilon \end{cases}$$

The posterior risk is the probability

$$P(|a-\theta|>\epsilon|x)$$

and is minimised when the following probability is maximised

$$P(|a-\theta|\le\epsilon|x)$$

This occurs at the posterior mode of $\pi(\theta|x)$ (draw a graph to check it).

Facts about Bayes Estimators

- Bayes estimators are also minimax estimators. But their risk (Bayes risk) is smaller.
- Bayes estimators are typically admissible estimators.
- For improper priors, Bayes estimators may not exist. If they do, they are called generalised Bayes estimators.
- Bayes estimators are biased.
- Like maximum likelihood estimators, Bayes estimators are asymptotically unbiased and efficient and normally distributed.
- **Famous examples:** Lasso and Ridge Regression estimators. More in week 5

Reading

J.O. Berger:
Sections 1.3 1.5 2.4.1 2.4.2 4.3.1 4.4.1 and 4.4.2