Decision Theory

1 The loss function

Given a parameter θ living in some space Θ, we would like to construct estimators of θ based on some data.
There are various ways of constructing estimators and thus our main task is to compare them.
To do this we introduce the notion of the loss function which measures to some extent how far the estimator is from the parameter θ.
Thus $L(\theta,\hat\theta)$ is the loss function.

Ex 1) $L(\theta,\hat\theta)=(\theta-\hat\theta)^2$ (probably the most important one)

2) $L(\theta,\hat\theta)=|\theta-\hat\theta|$

3) $L(\theta,\hat\theta)=|\theta-\hat\theta|^p$, $p>0$ the L^p-loss

4) $L(\theta,\hat\theta)=\begin{cases}0 & \text{if } \theta=\hat\theta\\ 1 & \text{if } \theta\neq\hat\theta\end{cases}$ zero-one loss

5) $L(\theta,\hat\theta)=\int \log\dfrac{f(x;\theta)}{f(x;\hat\theta)}\, f(x;\theta)\,dx$

which is called the Kullback-Leibler loss

Now assume we harvested some data $X_1,\dots,X_n$.
The estimator $\hat\theta$ is a function of $X_1,\dots,X_n$.

Definition The risk of an estimator is

$$R(\theta,\hat\theta)=E_\theta\left[L(\theta,\hat\theta)\right]=\int L(\theta,\hat\theta(x))f(x;\theta)\,dx$$

where $f(x;\theta)$ is the density of the sample $X_1,\dots,X_n$.

2 Comparison of estimators and risks

In principle, we would like to have an estimator with the smallest possible risk. However the risk actually depends on the parameter θ range.

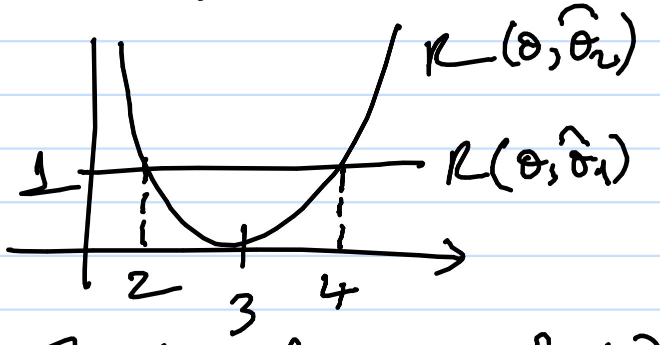
Ex $X \sim N(\theta, 1)$ and $\hat{\theta}_1 = X$, $\hat{\theta}_2 = 3$. Which of these is better in terms of quadratic loss?

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

and $R(\theta, \hat{\theta}_1) = E[(X-\theta)^2] = Var(X) = 1$

$$R(\theta, \hat{\theta}_2) = E[(3-\theta)^2] = (3-\theta)^2.$$

Thus depending on the range of θ we have one or the other being smaller



Therefore for $\theta \in [2,4)$, $\hat{\theta}_2$ is a better estimator, but for $\theta \notin [2,4)$, $\hat{\theta}_1$ is better.

Thus the risk functions are not comparable everywhere.

Ex2 This is one important example as it introduces the key concept of Bayes estimator

Take $X_1, \dots, X_n \sim$ Bernoulli(p). Consider the square error loss and let $\hat{p} = \overline{X}$. This has

$$R(p, \hat{p}) = E[(\hat{p}-p)^2] = E[(\overline{X}-p)^2] = $$

$$= Var(\overline{X}) = \frac{p(1-p)}{n}.$$

This $\hat{p}_1$ is the MLE and also the MVUE for p.
Another estimator is

$$\hat{p}_2 = \frac{\sum x_i + \alpha}{\alpha + \beta + n} \quad \text{for } \alpha, \beta > 0$$

This is obtained via the Bayes' theory of estimators taking the prior parameter p to be distributed uniformly in $[0,1]$. See the appendix for more on this.

The idea here is that the risk of $\hat{p}_2$ is

$$E[(\hat{p}_2 - p)^2] = E\left[\left(\frac{Y+\alpha}{\alpha+\beta+n} - p\right)^2\right] =$$

$Y = \sum_{i=1}^{n} X_i$

$$= Var\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(E\left[\frac{Y+\alpha}{\alpha+\beta+n}\right] - p\right)^2$$

$$= \frac{1}{(\alpha+\beta+n)^2} Var(Y) + \frac{(np+\alpha-p(\alpha+\beta+n))^2}{(\alpha+\beta+n)^2}$$

$$= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \frac{((1-p)\alpha - p\beta)^2}{(\alpha+\beta+n)^2}$$

$$= \frac{np(1-p) + ((1-p)\alpha - p\beta)^2}{(\alpha+\beta+n)^2}$$

$$= \frac{np - np^2 + (\alpha - p(\alpha+\beta))^2}{(\alpha+\beta+n)^2}$$

If we choose this to be independent of p, then we need to make sure that

$$np - np^2 + (\alpha - p(\alpha+\beta))^2 =$$
$$= np - np^2 + \alpha^2 - 2\alpha(\alpha+\beta)p + p^2(\alpha+\beta)^2$$
$$= \alpha^2 + p(n - 2\alpha(\alpha+\beta)) + p^2((\alpha+\beta)^2 - n)$$

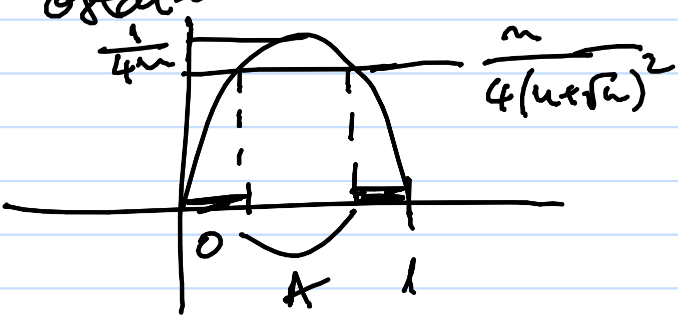
Thus we need to choose $\alpha+\beta = \sqrt{n}$ &
$\alpha(\alpha+\beta) = \frac{n}{2}$

which solves for $\alpha = \beta = \frac{\sqrt{n}}{2}$. This leads to $\hat{p}_2 = \frac{Y + \sqrt{n}/2}{n+\sqrt{n}}$ and the risk is

$$R(\hat{p}_2, p) = \frac{\alpha^2}{(\alpha+\beta+n)^2} = \frac{n}{4(n+\sqrt{n})^2} < \frac{1}{4n}$$

which is independent of p.

Thus if we draw the dependence on p we obtain



As we can see the first estimator is smaller on the set A

Notice that as $n \to \infty$, $\frac{n}{4(n+\sqrt{n})^2} \sim \frac{1}{4n}$.

One natural way of defining the risk is

$$\overline{R}(\hat\theta) = \sup_\theta R(\theta, \hat\theta)$$

and the Bayes risk as

$$r(f, \hat\theta) = \int R(\theta, \hat\theta) f(\theta) d\theta$$

where $f(\theta)$ is a prior for θ.

In the previous case we have

$$\overline{R}(\hat p_1) = \sup_p \frac{p(1-p)}{n} = \frac{1}{4n}$$

$$\overline{R}(\hat p_2) = \frac{n}{4(n+\sqrt{n})^2} < \frac{1}{4n}.$$

As we can see this means that $\hat p_2$ is better than $\hat p_1$ in general, however in absolute terms $\hat p_2$ is better only for p outside a region which shrinks to 0 around 1/2.

Therefore $\hat p_1$ could be the preferred estimator though.

The Bayes' risk with the uniform prior f for $\hat p_1$ and $\hat p_2$ are given by

$$r(f, \hat p_1) = \int_0^1 R(p, \hat p_1) dp = \int_0^1 \frac{p(1-p)}{n} dp = \frac{1}{6n}$$

$$r(f, \hat p_2) = \int_0^1 \frac{n}{4(n+\sqrt{n})^2} dp = \frac{n}{4(n+\sqrt{n})^2} > \frac{1}{6n}$$

for n large enough, $(n \geq 20)$. Thus this suggests $\hat p_1$ is better than $\hat p_2$.

Definition A decision rule that minimizes the Bayes risk is called a Bayes rule/estimator.

Formally $\hat{\theta}$ is a Bayes rule with respect to f if

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$$

where the infimum is taken over all possible estimators $\tilde{\theta}$.

An estimator that minimizes the maximum risk is called the minimax rule. Formally, $\hat{\theta}$ is minimax if

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

where the infimum is taken over all estimators $\tilde{\theta}$.

Bayes Estimators

Assume f to be a given fixed prior. Then

$$f(\theta|x) = \frac{f(x|\theta) f(\theta)}{m(x)} = \frac{f(x|\theta) f(\theta)}{\int f(x|\theta) f(\theta) d\theta}$$

where $m(x)$ is the pmf/pdf marginal of X.

The posterior risk of an estimator $\hat{\theta}(x)$ is

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta$$

Theorem 1) $r(f, \hat{\theta}) = \int r(\hat{\theta}|x) m(x) dx$.

2) If $\hat{\theta}(x)$ is the value of θ which minimizes $r(\hat{\theta}|x)$, then $\hat{\theta}(x)$ is a Bayes estimator.

Pf 1) $r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) d\theta$

$f(x|\theta)f(\theta) =$
$= f(\theta|x) m(x)$

$$= \int \left(\int L(\theta, \hat{\theta}(x)) f(x|\theta) dx \right) f(\theta) d\theta$$
$$= \int \int L(\theta, \hat{\theta}(x)) f(\theta|x) m(x) dx d\theta$$
$$= \int \left(\int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta \right) m(x) dx$$
$$= \int r(\hat{\theta}|x) m(x) dx$$

2) If we take $\hat{\theta}(x)$ the minimizer of $r(\hat{\theta}|x)$ then $r(f, \tilde{\theta}) = \int r(\tilde{\theta}|x) m(x) dx \geq$
$$\geq \int r(\hat{\theta}(x)|x) m(x) dx = r(f, \hat{\theta})$$
and thus $\hat{\theta}$ is the Bayes estimator.

Theorem If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then the Bayes estimator is given by
$$\hat{\theta}(x) = \int \theta f(\theta|x) d\theta = E[\theta|X=x].$$
The Bayes rule minimizes
$$r(\hat{\theta}|x) = E_\theta[(\theta - \hat{\theta}(x))^2 | X=x]$$
$$= \int (\theta - \hat{\theta}(x))^2 f(\theta|x) d\theta$$
and thus $\hat{\theta}(x) = E[\theta|X=x]$.

Pf Indeed $r(\hat{\theta}|x) = \int (\theta - \hat{\theta})^2 f(\theta|x) d\theta$ and thus the value $\hat{\theta}(x)$ which minimizes this is $E[\theta|X=x]$.

Appendix How to compute the Bayes estimators.

Assume that we have $f(x|\theta)$ and we make the assumption that θ follows a certain distribution called prior.

Then the Bayes estimator of θ given the data X is the posterior distribution

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

where $f(x) = \int f(x|\theta)f(\theta)d\theta$.

Typically, if we observe $x_1, \dots, x_n$, then

$$\mathcal{L}_n(\theta) = f(x_1|\theta)\cdots f(x_n|\theta).$$

Thus $f(\theta|x^n) = C_n \mathcal{L}_n(\theta) f(\theta)$

for some constant C_n.

The mean of the posterior distribution is a good source of estimators for the parameter θ.

Ex $X_1, \dots, X_n \sim$ Bernoulli(p). If we take the uniform dist for p, then

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^S(1-p)^{n-S}$$

where $S = \sum_{i=1}^n x_i$. Thus the posterior dist of p is Beta$(S+1, n-S+1)$. In general Beta(α, β) has density $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

The mean of Beta(α,β) is

$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x\, x^{\alpha-1}(1-x)^{\beta-1}\,dx =$$

$$= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha+1-1}(1-x)^{\beta-1}\,dx =$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} =$$

$$= \frac{\alpha}{\alpha+\beta}.$$

Thus we have $p \mid x^n \sim \text{Beta}(s+1, n-s+1)$

The mean of this is $\bar{p} = \dfrac{s+1}{n-s+1+s+1} = \dfrac{s+1}{n+2}$.

Thus the estimator can be written as

$$\bar{p} = \frac{s+1}{n+2}.$$

In general, if we take as prior Beta(α,β) then $\bar{p} = \dfrac{s+\alpha}{\alpha+\beta+n}$

We can write this as

$$\bar{p} = \frac{n}{\alpha+\beta+n}\cdot\frac{s}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\, p_0$$

where $p_0 = \dfrac{\alpha}{\alpha+\beta}$ is the prior mean.

<u>Ex</u> If $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ and we take the prior to be $N(a, b^2)$, then $\theta \mid X^n \sim N(\bar{\theta}, \tau^2)$ where $\bar{\theta} = w\bar{X} + (1-w)a$

$$w = \frac{1/s^2}{1/s^2 + 1/b^2}, \quad 1/\tau^2 = 1/s^2 + 1/b^2$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Indeed $f(x|\theta) \propto e^{-\sum \frac{(x_i-\theta)^2}{2\sigma^2}}$

and then $f(x|\theta)f(\theta) \propto e^{-\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-a)^2}{2b^2}}$

and this can be written as

$$e^{-\frac{\theta^2}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)+\theta\left(\frac{n\bar{x}}{\sigma^2}+\frac{a}{b^2}\right)}$$

$$= e^{-\frac{\theta^2}{2\alpha^2}+\theta\beta} = e^{-\frac{\theta^2}{2\alpha^2}+2\frac{\theta}{2\alpha^2}\cdot\alpha^2\beta} \propto e^{-\frac{(\theta-\alpha^2\beta)^2}{2\alpha^2}}$$

with $\frac{1}{\alpha^2} = \frac{n}{\sigma^2}+\frac{1}{b^2}$ & $\beta = \frac{n\bar{x}}{\sigma^2}+\frac{a}{b^2}$

which means that $\theta|x \sim N(\alpha^2\beta, \alpha^2)$.

We can write the mean as

$$\bar{\theta} = w\bar{x}+(1-w)a$$

where $w = \frac{\alpha^2}{\sigma^2/n} = \frac{1/(\sigma^2/n)}{1/(\sigma^2/n)+1/b^2}$

which is a linear combination between $\bar{x}$, the MLE & a.