Decision Theory

Sofia Olhede

EPFL

October 13, 2020

Loss functions IV

- The risk is

?

$$
\begin{aligned}
R(\lambda, \widetilde{\lambda}) &= \mathbb{E}_\lambda\left[\frac{n\lambda\bar{Y}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{Y}}{n-1}\right)\right] \\
&= \mathbb{E}_\lambda\left[\lambda\bar{Y} - 1 - \log(\lambda\bar{Y})\right] + \frac{\mathbb{E}_\lambda\left[\lambda\bar{Y}\right]}{n-1} - \log\left(\frac{n}{n-1}\right) \\
&= \underbrace{\mathbb{E}_\lambda\left[\lambda\bar{Y} - 1 - \log(\lambda\bar{Y})\right]}_{\text{risk for } \lambda} + g(n). \qquad (1)
\end{aligned}
$$

+ a positive addition

- To derive the simplification we write $\bar{Y} = \frac{n-1}{n}\bar{Y} + \frac{1}{n}\bar{Y}$.
- Note that $\mathbb{E}_\lambda(\bar{Y}) = \lambda^{-1}$. Thus

$$
g(n) = \frac{1}{n-1} - \log\left(\frac{n}{n-1}\right).
$$

- We claim that $g(n) > 0$ once $n \geq 2$.

Loss functions V

Random variable A has [first-order stochastic] dominance over random variable B if for any outcome x, A gives at least as high a probability of receiving at least x as does B, and for some x, A gives a higher probability of receiving at least x.

- Using that $\log(x) = \int_1^x t^{-1}\, dt$ this follows if

$$\frac{1}{x} > \log(x+1) - \log(x), \quad x > 1$$
$$\Leftrightarrow \frac{1}{x} > \int_x^{x+1} t^{-1} dt, \quad x > 1. \tag{2}$$

?

This inequality holds by a rectangle area bound on the integral, as follows:

$$\frac{1}{x} = [(1+x) - x]\frac{1}{x} = \int_x^{x+1} \frac{1}{x} dt > \int_x^{x+1} \frac{1}{t} dt,$$

when $x > 1$.

- It therefore follows $R(\widetilde{\lambda}, \lambda) > R(\widehat{\lambda}, \lambda)$ and so $\widetilde{\lambda}$ dominates $\widehat{\lambda}$.

Decision Theory

- We can push generality even further, and obtain an all encompassing framework.
- Called decision theory, it views inference as a game between nature and the statistician.
- Recall our general framework for statistical inference:

Reminder on how we do things

1. Model phenomenon by distribution $F(y_1, \ldots, y_n; \theta)$ for some $n \geq 1$.
   of how data was generated
2. Distributional form is known but $\theta \in \Theta$ is not known.
3. Observe realisation of $(Y_1, \ldots, Y_n)$ from this distribution.
4. Use $(Y_1, \ldots, Y_n)$ in order to make assertions concerning the true value of , and quantify the uncertainty associated with these assertions.

- The decision theory framework formalises step (4) to include estimation, testing, and confidence intervals.

Decision Theory II

- In the decision theory framework we usually have these formal constructs:
  - * A family of distributions $\mathcal{F}$ usually assumed to admit densities (frequencies). → eg 0 or 1 in Bernoulli
  - * A parameter space Θ that is used to parameterize $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This models the possible realizations that may happen.
  - * The space on which observations are taken, the data space $\mathcal{Y}^n$.
  - * The action space which represents the space of possible actions or decisions or plays/moves available to the statistician.
  - * A loss function $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$.
  - * A set $\mathcal{D}$ of decision rules. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{Y}^n \mapsto \mathcal{A}$. This corresponds to possible strategies.

Decision Theory III

data space action

decision rule / function $\delta : \mathcal{Y}^n \to \mathcal{A}$

- The statistician would pick δ to limit losses. As the losses are random, actions are associated with risk.
- Given a decision rule $\delta : \mathcal{Y}^n \mapsto \mathcal{A}$ the risk is:
- $R(\delta, \theta) = \mathbb{E}\{\mathcal{L}(\delta(\mathrm{Y}, \theta)\}$. We compare decision rules in terms of their risk functions. the smaller the risk, the better!
- Risk is not uniform, but depends on θ the true state of nature.
- So comparisons can be made picking a decision rule which is always better
  1. Uniform comparisons (hard). Seek dominance everywhere in Θ.
  2. Minimax (relaxed). Compare worst-case risks over Θ.
  3. Bayes (relaxed). Compare average risk over Θ.

Risk

ie look at all worst performances
and pick the "best worse"

- Rather than look at risk at every value of θ minimax risk concentrates (focuses) on the maximum risk:
- Definition (Minimax decision rule): Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{f_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If

$$\sup_{\theta\in\Theta} R(\delta,\theta) \leq \sup_{\theta\in\Theta} R(\delta',\theta),$$

  then δ is a minimax rule.
- Rather than the behaviour at all θ, Bayes risk is a weighted average.
- Definition (Bayes Risk): Let $\pi(\theta)$ be a probability density (or frequency) and let δ be a decision rule for the experiment $(\{f_\theta\}_{\theta\in\Theta}, \mathcal{L})$. The π-Bayes risk is defined as

prior on Θ ie the very-likely values of θ

"posterior risk"

$$r(\delta,\pi) = \int_\Theta R(\delta,\theta)\pi(\theta)\,d\theta = \int_\Theta\int_\mathcal{Y} \mathcal{L}(\delta(\boldsymbol{y}),\theta) f_\theta(\boldsymbol{y})\pi(\theta)\,d\boldsymbol{y}\,d\theta.$$

  If $\delta\in\mathcal{D}$ is such that $r(\delta,\pi)\leq r(\delta',\pi)$ for all $\delta'\in\mathcal{D}$ then δ is a decision rule wrt π.

Risk II

- The choice of prior $\pi(\theta)$ places different emphasis for different values of θ considering what prior knowledge we have.
- Minimax rules are useful to establish the fundamental inferential complexity or a statistical experiment.
- Using them for more practical purposes requires caution.
- Motivated as follows: we do not know anything about θ so let us insure ourselves against the worst thing that can happen.
- Minimax is quite a conservative point-of-view.

Minimax it's not so good

- Bayes rules are quite attractive as they can nearly never be uniformly dominated.
- Intuitively, if you can show your rule to be Bayes for a nice prior, you know you are doing reasonably well.

Bayes is better

Hypothesis Testing

- Model a phenomenon by a distribution $F(y_1, \ldots, y_n; \theta)$ on $\mathcal{Y}^n$ for some $n \geq 1$.
- Distributional form is known but $\theta \in \Theta$ is unknown.
- Observe $\begin{pmatrix} Y_1 & \ldots & Y_n \end{pmatrix}^T \in \mathcal{Y}^n$ from this distribution.
- Use the observed data $\begin{pmatrix} Y_1 & \ldots & Y_n \end{pmatrix}^T$ to make statements about θ.
- The first assertion we wish to make is hypothesis testing. Given two disjoint regions Θ_0 and Θ_1, i.e. we have $\Theta_0 \cap \Theta_1 = \emptyset$, which interval is more likely to contain the true value of θ?
- We assume we know $\theta \in \Theta_0 \cup \Theta_1$.
- We need to use $\begin{pmatrix} Y_1 & \ldots & Y_n \end{pmatrix}^T \in \mathcal{Y}^n$ to decide between the two possibilities.

eg. in a trial: guilty or not guilty regions

eg in hospital: sick vs healthy

Hypothesis Testing II

- Often in science two concurrent theories need to be confronted with the empirical evidence.

The null hypothesis H_0 which states that $\theta \in \Theta_0$

$$H_0 \; : \; \theta \in \Theta_0.$$

The alternative hypothesis that postulates $\theta \in \Theta_1$

$$H_1 \; : \; \theta \in \Theta_1.$$

Definition (Test Function): A test function is a map $\delta : \mathcal{Y}^n \mapsto \{0, 1\}$.

- Obtaining 0 or 1 must be decided on whether or not the sample satisfies a certain condition:

critical region

$$\delta(\mathrm{Y}) = \begin{cases} 1, & \text{if} \quad T(Y_1 \ldots Y_n) \in C \\ 0, & \text{if} \quad T(Y_1 \ldots Y_n) \notin C \end{cases} .$$

Hypothesis Testing III

- T is a statistic called a test statistic and;
- C is a subset of the range of T and is called the critical region.
- We can write

$$\delta(\mathrm{Y}) = \mathrm{I}(T(Y_1 \ldots Y_n) \in C).$$

- To choose good test functions we need to quantify the performance of a test function.
- Remark that, obviously, δ has a Bernoulli distribution:

$$\delta = \begin{cases} 1, & \text{if} \quad T(Y_1 \ldots Y_n) \in C \\ 0, & \text{if} \quad T(Y_1 \ldots Y_n) \notin C \end{cases}.$$

- So a good test function must have a sampling distribution concentrated around the right decision.
- The difference from point estimation is that our action space is discrete.
- But how do we compare δ?

Hypothesis Testing IV

- What possible errors are there?
- Take action 0 when H_1 is true–this is a type II error. Take action 1 when H_0 is true–this is a type I error.

?

- If we abused terminology we could define

$$\mathrm{MSE}\{\delta, H_i\} = \mathbb{E}_\theta(\delta - i)^2.$$

test result — the index of the right hypothesis

- We can then deduce that

$$\begin{aligned}\mathrm{MSE}\{\delta, H_i\} &= \begin{cases} \mathbb{E}_\theta(\delta) & \text{if} \quad \theta \in \Theta_0 \\ \mathbb{E}_\theta(1-\delta) & \text{if} \quad \theta \in \Theta_1 \end{cases} \\ &= \begin{cases} \Pr_\theta(\delta = 1) & \text{if} \quad \theta \in \Theta_0 \\ \Pr_\theta(\delta = 0) & \text{if} \quad \theta \in \Theta_1 \end{cases} \\ &= \begin{cases} \Pr_\theta(\delta = 1) & \text{if} \quad \theta \in \Theta_0 \\ 1 - \Pr_\theta(\delta = 1) & \text{if} \quad \theta \in \Theta_1 \end{cases}.\end{aligned}$$

?

Asymmetry of false positive versus false negative.

Hypothesis Testing V

- In decision theory terms, the action space is $\mathcal{A} = \{0, 1\}$ and the loss function is the so-called "0–1" loss,
- The loss is then

$$\mathcal{L}(a, \theta) = \begin{cases} 1 & \text{if} \quad \theta \in \Theta_0 \ \& \ a = 1 \text{ (Type } I \text{ error)} \\ 1 & \text{if} \quad \theta \in \Theta_1 \ \& \ a = 0 \text{ (Type } II \text{ error)} \\ 0 & \text{o/w} \quad \text{(no error)} \end{cases} . \left.\vphantom{\begin{cases}1\\1\end{cases}}\right\} \text{ie wrong guess}$$

- Thus, we lose 1 unit whenever committing a type I or type II error.
- The risk function is

$$\mathcal{R}(\delta, \theta) = \begin{cases} \Pr\{\delta = 1\} & \text{if} \quad \theta \in \Theta_0 \text{ (Prob type } I \text{ error)} \\ \Pr\{\delta = 0\} & \text{if} \quad \theta \in \Theta_1 \text{ (Prob type } II \text{ error)} \end{cases} .$$

probability of δ being 0.

- Thus

$$\mathcal{R}(\delta, \theta) = \Pr\{\delta = 1\} \mathrm{I}(\theta \in \Theta_0) + \Pr\{\delta = 0\} \mathrm{I}(\theta \in \Theta_1).$$

Can we simultaneously control both errors??? NO :(.

Hypothesis Testing VI

- Let us understand why.
- Let $\delta(Y_1 \ldots Y_n) = \mathrm{I}(T(Y_1 \ldots Y_n) \in C)$.
- Suppose we wish to reduce the type I error probability $\Pr_\theta(\delta = 1)$ if $\theta \in \Theta_0$.

  eg we take a lower threshold for being guilty in trial
- To do this, we would replace C by a subset $C^* \subset C$. This gives $\delta_*(Y_1 \ldots Y_n) = \mathrm{I}(T(Y_1 \ldots Y_n) \in C^*)$.

  eg instead of C = 8 crimes or more = $[8, \infty)$ we will have C^* = 9 crimes or more = $[9, \infty)$
- Observe that for $\theta \in \Theta_0$ we have

$$\Pr_\theta\{\delta_* = 1\} = \Pr\{T \in C^*\} \leq \Pr\{T \in C\} = \Pr_\theta\{\delta = 1\}.$$

checks if result is in the new region

- Whilst for $\theta \in \Theta_1$ we have

$$\Pr_\theta\{\delta_* = 0\} = \Pr\{T \notin C^*\} \geq \Pr\{T \notin C\} = \Pr_\theta\{\delta = 0\}.$$

- By reducing the type I error probability we increased the type II error probability! → obvious but a good reminder

Hypothesis Testing VII

- We have to make a philosophical choice regarding the importance of different errors.
- In applications, one type of error (false positive or negative) is typically more severe.
- Say this is the type I error, and exploit the asymmetry: fix a tolerance ceiling for the probability of this error.
- Given this ceiling, consider only test functions that respect it, and focus on minimising type II error (i.e. maximising power).

Hypothesis Testing VIII

- The Neyman-Pearson Framework
- We declare that we only consider test functions $\delta : \mathcal{X} \mapsto \{0, 1\}$ such that

observation space

$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \{\delta : \sup_{\theta \in \Theta_0} \Pr_\theta\{\delta = 1\} \leq \alpha\}.$$

i.e. the probability of the success hypothesis is kept limited (below α)

- i.e. rules for which prob of type I error is bounded above by α.
- Jargon: we fix a significance level for our test.

bc i.e. we need good data but also anomalies

- Within this restricted class of rules, choose δ to minimize prob of type II error:

$$\Pr\{\delta(\boldsymbol{X}) = 0\} = 1 - \Pr\{\delta(\boldsymbol{X}) = 1\}.$$

- Equivalently, maximize the power

$$\beta(\theta, \delta) = \Pr\{\delta(\boldsymbol{X}) = 1\} = \mathbb{E}\,\mathrm{I}\{\delta(\boldsymbol{X}) = 1\} = \mathbb{E}\{\delta(\boldsymbol{X}))\}, \quad \theta \in \Theta_1.$$

Hypothesis Testing IX

- Neyman-Pearson setup naturally exploits any asymmetric structure.
- But, if natural asymmetry absent, need judicious choice of H_0.
- Consider simplest situation: $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.
- The Neyman Pearson Lemma: Let $\boldsymbol{Y}$ have joint density/frequency f where $f \in \{f_0, f_1\}$. We wish to test

$$H_0 : \quad f = f_0 \quad \text{and} \quad f = f_1.$$

or?

- If $\Lambda(\mathrm{Y}) = f_1(\mathrm{Y})/f_0(\mathrm{Y})$ is a continuous random variable, then there exists a $k > 0$ such that $\Pr\{\Lambda(\mathrm{Y}) \geq k | H_0\} = \alpha$ and the test whose test function is given by $\delta(\mathrm{Y}) = \mathrm{I}(\Lambda(\mathrm{Y}) \geq k)$ is a most powerful (MP) test of H_0 versus H_1 at significance level α.