5 Decision Theory

5.1 Notation

Here we introduce the basic notation.

- $\theta \in \Theta$: an unknown quantity affecting decision process.
- $a \in \mathcal{A}$: action to be taken.
- $x \in \mathcal{X}$: data
- $L(\theta_0, a) : \Theta \times \mathcal{A} \to R$: loss for performing action a if "true" value of θ is θ_0.
- $p(x|\theta)$: density of X.
- $\pi(\theta)$: ("prior") distribution for θ.
- $\delta : \mathcal{X} \to \mathcal{A}$ a decision rule that maps data to actions.
- $R(\theta, \delta)$ the Risk function for decision rule δ, being the expected loss (as a function of θ, with expecation over x).
- $r(\pi, \delta)$ the integrated risk (also known as the "Bayes Risk"), with respect to a distribution π on θ, $r(\pi, \delta) := \int R(\theta, \delta)\pi(d\theta)$.

In statistics we usually want to explicitly consider data that we are using in deciding to make a decision. However, decision theory also can be applied to problems where there are no "data", and actions are taken directly on the basis of $\pi(\theta)$.

Note that statistical inference is one special case of a decision problem. Suppose we want to estimate a parameter θ. The "action" we want to take is to report our estimate. Thus the action space consists of the possible estimates we might report, and so $\mathcal{A} = \Theta$. In this case the "action" is usually to report an estimate, which is often denoted $\hat{\theta}$. So the loss function becomes a function $L(\theta, \hat{\theta})$.

Decision theory plays an important role in economic theory as well as statistical theory. In economics, the problem is usually framed in terms of *utility* function rather than a *loss* function. You can think of utility as the negation of a loss, so $U(\theta_0, a) = -L(\theta_0, a)$. You might also think of the loss function as a "cost function".

Example: taking umbrella to work

You are trying to decide whether or not to bring an umbrella to work with you. It is not raining now, but you are concerned that it might rain during your walk home. You don't really like to carry your umbrella if you don't need it, but on the other hand if it rains in the evening you would be glad you had taken it for your walk home.

The action space is $\mathcal{A} = \{\text{take umbrella}, \text{don't take umbrella}\}$. The unknown quantity affecting your decision is whether it will rain in the evening. So $\Theta = \{\text{rain}, \text{no rain}\}$. You might also have data x to be the weather forecast.

Your loss function L tells you how sad you will be if certain action-θ combinations. For example, we might assume that your loss looks like this:

$\theta \backslash a$	umbrella	no umbrella
no rain	1	0
rain	2	10

Table 1: Example loss function for taking umbrella to work problem

Here the units of loss are arbitrary, but this captures the idea that the worst case is if it rains and you have no umbrella. Having your umbrella with you when it rains saves you much of that sadness, although you might still be a bit sad that it rains (your feet might get wet!). And you have a mild inconvenience if you take your umbrella and it does not rain.

5.2 Choosing an action: the conditional Bayesian Principle

Here we just state the principle, which should seem natural. Then we will see some examples before taking another look at motivation for the principle.

This principle says you should "Choose action a to minimize the posterior expected loss, which is given by

$$\text{posterior expected loss} = \int L(\theta, a) p(\theta|x) d\theta \tag{1}$$

Note that the action that minimizes this will depend on x, so this defines a decision rule $\delta(\cdot)$ that maps data to actions. This decision rule also depends on the prior π so you might write $\delta^\pi(\cdot)$.

Note that this principle tells us how we (rational decision makers) "should" behave, and is not always a good description of how people actually behave in practice!

Example: umbrella problem

Suppose that after looking at the weather forecast your posterior probability of it raining during your walk home is p. Then (under the loss function in Table 5.1) your expected loss if you do or do not take an umbrella is

$$EL(\text{umbrella}) = 2p + 1(1-p) = 1 + p \tag{2}$$

$$EL(\text{no umbrella}) = 10p + 0(1-p) = 10p. \tag{3}$$

So by the conditional Bayesian principle you should take the umbrella if $1 + p < 10p$, or in other words if $p > 1/9$.

5.3 More Examples

1. *You are a competitor on a quiz show. So far, you have won five thousand dollars, and are given the choice between i) leaving with the five thousand dollars; ii) attempting a bonus question, which, if you answer correctly, will give you a Toyota Prius car. Which do you choose?*

   Answer 1: You have seen the quiz show before, and the bonus questions are quite hard: you think you are able to answer them correctly about half the time. That is, your personal probability that you will answer the bonus question correctly is 0.5. On the other hand, you have always wanted a Prius, and your current car is on its last legs so you are going to buy a new one soon anyway. The Prius costs more than 20,000 dollars. So, treating dollars as utility, the expected utility for i) is 5,000 dollars, and for ii) is 10,000 dollars, so you choose ii).

   Answer 2: You are happy with your car, and don't really like the Prius. Of course, you could always sell the Prius. Perhaps you would get 18,500 for it. But it would be a bit of work, so maybe the Prius is worth 18,000 to you when you've accounted for that. Also, although you think you could answer previous bonus questions right about half the time, you are feeling unlucky today, so your probability that you will get today's question right is more like 0.25. So your utility for i) is still 5,000 dollars, but for ii) it is 0.25*18,000 = 4,500. You choose i).

2. *(Purchasing Insurance) You are buying a new Ipod at Best Buy. It costs 300 dollars, and comes with a 1 year warranty. They ask you if you would like to purchase the extended 3-year warranty for 30 dollars. Do you?*

   Possible Answer: You know that they must be expecting to make money on the warranty, so you figure that it is used by fewer than one person in ten. If you buy the warranty, that will

cost you 30 dollars. If you do not buy the warranty, then if your ipod breaks during years 2 or 3 you will have to replace it. But costs are dropping, and things always get better, so you will presumably spend less than 300 to replace it for something better. So you figure your utility for buying it is -30, and your utility for not buying it is more than (-300)*0.1 = -30. So you don't buy it.

3. *(Purchasing Insurance 2). You have just bought a new house, for 500,000 dollars. To insure the house against fire for one year costs 1,000 dollars. Do you do it?*

   Possible Answer: Fire is pretty unlikely. Indeed, you know that they expect to make money on insurance, so presumably it affects less than 1 in 500 houses per year. On the other hand, if your house burned down you would lose everything, and, to you, losing your house is unthinkable: much more than 500 times worse than losing 1,000 dollars. So you buy the insurance. (Note: Of course, in most cases the people lending you money to buy a house insist you insure it!)

4. *Classification with 2 classes (e.g. hypothesis testing, H_0 vs H_1).* Let $c_{i|j}$ denote the loss for choosing H_i if H_j is true. So $c_{1|0}$ is the cost of rejecting H_0 if H_0 is true, or in other words the cost of a type I error. And $c_{0|1}$ is the cost of a type II error. Assume that if you make the right choice then you lose zero: $c_{0|0} = c_{1|1} = 0$.

   The action space and the space of unknowns is the same: we don't know which hypothesis is true, and we wish to select one. Thus $\Theta = \mathcal{A} = \{H_0, H_1\}$.

   Posterior expected loss$(H_0) = p(H_1|x)c_{0|1}$

   Posterior expected loss$(H_1) = p(H_0|x)c_{1|0}$

   Choose H_1 if $p(H_1|x)c_{0|1} > p(H_0|x)c_{1|0}$. That is if $p(H_1|x) > p(H_0|x)c_{1|0}/c_{0|1}$.

   If costs are equal, then threshold is 0.5. If cost of type I error is higher than cost of type II error then threshold is higher.

5. *(Point Estimation). In current genetic studies, the most commonly-used type of marker is called a SNP (Single Nucleotide Polymorphism). Each SNP is a single position in the genome, where different copies of the genome carry one of two different bases. For example, at a C/G SNP, some genome copies carry the C, and others have a G. Since we each have two copies of our genome (one from mother, or one from father), at a C/G SNP each of us will have either 0, 1 or 2 Cs. That is our "genotype" can be thought of as a 0, 1 or 2 variable.*

   Suppose now that we have a method for measuring genotypes. This method acknowledges that the measurement may have errors, so instead of just giving a genotype for each individual, it

instead gives probabilities p_0, p_1 and p_2 for the genotype to be 0,1 or 2. Given these probabilities, what would you report to be the genotype?

Answer 1 (0-1 loss; mode): You decide that if the genotype is incorrect, you will lose 1 unit, whereas if it is correct, you lose 0. This is referred to as 0-1 loss, and can be written as

$$L(g; \hat{g}) = I(\hat{g} \neq g),$$

where $L(g; \hat{g})$ denotes the loss ("consequence") you suffer if you report $\hat{g}$ ("action") when the truth is g ("Event"). Note that 0-1 loss assumes that, when you are wrong, it does not matter *how* wrong you are.

If you report anything but 0, 1 or 2 as your genotype you are certain to be wrong, so your expected loss would be 1. If you report $g \in \{0, 1, 2\}$ then your expected loss is $1 - p_g$. To minimize your expected loss you report $\hat{g} = \arg\max p_g$. That is, you report the modal genotype.

Answer 2 (quadratic loss; mean): You decide that, when a mistake is made, some mistakes are worse than others. For example, if the truth is 0, and you guess 2, then this is worse than guessing 1, even though both guesses are wrong. A common way to quantify this is via quadratic loss, which penalizes big mistakes quite harshly:

$$L(g; \hat{g}) = (g - \hat{g})^2.$$

Now you want to choose $\hat{g}$ to minimize your expected loss

$$E[L(g; \hat{g})] = \sum_g p_g (g - \hat{g})^2.$$

Differentiating with respect to $\hat{g}$, and setting the derivative to zero, yields $\hat{g} = \sum_g p_g g$. That is, you estimate the genotype by the mean, averaging over the uncertainty.

Note: The results regarding 0-1 loss and quadratic loss are quite general, and are not limited only to this example, or to discrete outcomes. That is, reporting a mode of a distribution as the point estimate is to implicitly adopt 0-1 loss; reporting the mean as a point estimate is to implicitly adopt quadratic loss.

6. *(Forecasting: Proper scoring rules).* Proper scoring rules arise when you are asked to forecast the probability of an event.

   Suppose your final exam is a multiple choice exam, and you have to assign, for each answer A, B, C and D a probability that it is correct. That is the "action" is a probability vector (q_A, q_B, q_C, q_D) of non-negative numbers summing to 1. The unknown quantity θ that affects

your response is which answer is actually correct. That is $\Theta = \{A, B, C, D\}$. What loss function is approriate for this task?

Here are two loss functions that make some sense:

i) (Brier score) $L(\theta; q) = \sum_{i \in \Theta} (q_i - I(\theta = i))^2$

ii) ("log loss") $L(\theta; q) = -\log(q_\theta)$.

Note that the Brier score measures the distance between the vector q and the unit vector that corresponds to θ (e.g. (1,0,0,0) if $\theta = A$). The log loss is intuitive in that you lose more if you assign small probability to the correct answer. (Note that you lose ∞ if you assign 0 probability to the correct answer!)

Here is another loss function that might seem to make sense, but actually is not recommended, for reasons we will come to in a moment:

iii) $L(\theta; q) = 1 - q_\theta$.

The reason that i) and ii) are recommended over iii) is that they are *proper scoring rules*. The meaning of this is as follows. Suppose that after reading the question and answers you have decided that answers A to D have, for you, (posterior) probabilities (p_A, p_B, p_C, p_D). Then under loss functions i) and ii) your posterior expected loss would be minimized by reporting $q = p$. That is, these loss functions would tell you to "tell the truth" about your probabilities on A, B, C and D.

However, under loss function iii) your posterior expected loss would be minimized by a q other than p.

Exercise: show that i), and ii) are proper scoring rules, but iii) is not. Find the optimal q (as a function of p) for iii).

The bottom line: if you are ever in the position of organizing a "forecasting competition" where you have to assess how good a probabilistic forecast is by comparing it with what actually happened, then you should use a proper scoring rule so that competitors are not incentivized to "lie" and report something other than their actual probabilities to win the competition.

5.4 Frequentist decision theory, Risk and Admissibility

Definitions (Following Berger, Statistical Decision Theory and Bayesian Analysis, p9):

Risk Function

The *risk function* of a decision rule $\delta(x)$ is defined to be the expected loss, with expectation taken over repetitions of the experiment/data.

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(X))] = \int L(\theta, \delta(x))p(x|\theta)\, dx \tag{4}$$

Note that $R(\theta, \delta)$ measures how badly the decision rule δ would do on average, if the true value of θ were θ, and the rule was applied repeatedly to a large number of datasets. Ideally we'd like to choose a δ with a small risk. The problem is that risk depends on θ which is unknown, and risk functions of different δ will often cross: that is, for two decision rules δ_1 and δ_2, δ_1 may be better for one value of θ and worse for another.

Example: type I and type II error probabilities as risk functions

Suppose we consider the problem of comparing hypotheses H_0, H_1. So we can assume $\mathcal{A} = \Theta = \{0, 1\}$ with $\theta = j$ denoting that H_j holds, and $a = j$ denoting that we select H_j.

Define loss $L(\theta, a) = 1(a \neq \theta)$. Then

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(X))] = \Pr(\delta(X) \neq \theta|\theta) \tag{5}$$

so $R(0, \delta)$ is $\Pr(\delta(X) = 1|\theta = 0)$ or the probability of type I error. Similarly $R(1, \delta)$ is the probability of type II error.

If one test has a smaller type I error *and* a smaller type II error than another, then one might argue that the first test is better. This leads us to the concept of admissibility.

Admissibility

What if one rule is always riskier than another? Then intuitively we would never want to use it. That is the idea behind *admissibility*.

Definition: A decision rule δ_1 is *R-better* ("Risk-better") than a decision rule δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all θ, with strict inequality for some θ.

Definition: A decision rule δ is *admissible* if there is no R-better decision rule. Otherwise it is *inadmissible*.

Example (Berger, p10, example 4)

Assume $X \sim N(\theta, 1)$ and we want to estimate θ under loss $L(\theta, a) = (\theta - a)^2$. Consider the decision rules $\delta_\theta(x) = cx$. Show that the rules δ_c are inadmissible for $c > 1$.

in fact the rules are admissible for $0 \leq c \leq 1$ although this is harder to show. The fact that δ_0 is admissible shows that "admissible does not imply sensible or useful"!

Stein's paradox

This phenomenon, discovered by Charles Stein in 1956, is one of the most suprising results in statistics.

Let θ denote a vector of J unknown parameters. Suppose that for each j we have a single observation X_j that is a noisy estimate of θ_j:

$$X_j \sim N(\theta_j, \sigma^2). \tag{6}$$

Assume also that these measurements are independent. In this setting it is intuitive (and common) to use X_j as an estimate of θ_j. That is, to use

$$\hat{\theta} = \mathbf{X}. \tag{7}$$

The amazing result is that, for $J \geq 3$, this estimator is inadmissible, under squared loss

$$L(\theta, \hat{\theta}) = \sum_{j=1}^{J} (\theta_j - \hat{\theta}_j)^2. \tag{8}$$

Interestingly the proof is constructive: there is a simple analytic form for an estimator that is R-better. For example, James and Stein showed that

$$\hat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{X}\|^2}\right) \mathbf{X}. \tag{9}$$

is R-better than $\hat{\theta}$. In fact it turns out that this estimator is also inadmissible! But it can be made admissible by some minor modifications. Despite this, I think it would be reasonable to say that nobody actually uses this estimator in practice!

Note that if $(m-2)\sigma^2 < \|\mathbf{X}\|^2$ then the estimator (9) simply takes the natural estimator $\mathbf{X}$ and shrinks it towards the origin 0. For this reason this kind of estimator is sometimes called a "shrinkage

estimator". Actually you don't have to shrink towards the origin 0. There are analogues of (9) that shrink to any value ν, and these are R-better than $\hat{\theta}$ for any ν!

Again, JS-type estimators are seldom used in practice. But Bayesian versions of these estimators are perhaps more commonly used. The following is a simple exercise. Suppose you assume a prior on θ in which $\theta_j \sim N(\mu_0, \sigma_0)$ independently for each θ_j. Then the posterior mean of θ is a shrinkage estimator for θ (with shrinkage towards the prior mean μ_0). Personally I find the argument for shrinkage based on a prior belief much easier to understand, intuitively, than the argument for shrinkage based on admissibility.

Bayes Risk and Bayes Rule

Admissibility is all very well for comparing situations where one rule is always better than another. But what if the risk functions cross (as is more common)? That is, what if one rule is better for some θ, but worse for others? One way to compare rules is to average over values for θ, and compare the expected risk. Of course, the result will depend on what distribution π for θ you average over. So we define the integrated risk to be a function of π:

$$r(\pi, \delta) := \int R(\theta, \delta)\pi(d\theta). \tag{10}$$

Note that this is a frequentist concept in the sense that it asks about average performance of δ over different data sets. (but the different data sets have different values for θ, drawn from π). However, somewhat confusingly in my opinion, r is often referred to as the "Bayes Risk", and any decision rule δ minimizing r is referred to as a "Bayes rule".

As you might guess, if we are Bayesian we might select π to be our prior distribution, although we don't have to. The fundamental result is the following: Let δ_B^π denote the decision rule arising from performing a Bayesian analysis with prior π, and then selecting the action that minimizes the posterior expected loss. Then δ_B minimizes r. (That is, δ_B is a Bayes rule, which gives a post-hoc justification for this terminology.)

See Results 1 and 2 on p159-160 of Berger for more formal statement and proof.

Bayesian methods are Admissible

It is easy to show that, with mild regularity conditions, Bayesian methods - specifically, choosing the action that minimizes the posterior expected loss - are admissible. Let's do the finite discrete case ($\theta \in \{\theta_1, \ldots, \theta_J\}$) by contradiction [Theorem 7 from Berger (p253)]

Assume that Θ is discrete, say $\Theta = \{\theta_1, \theta_2, \dots\}$, and that the prior π gives positive probability to each $\theta_i \in \Theta$. Assume also that the Bayes Risk is finite. Then a Bayes rule δ_B^π with respect to π is admissible.

Proof: Suppose not. Then there exists a δ such that $R(\theta_i, \delta) \leq R(\theta_j, \delta_B^\pi)$ for all i, with strict inequality for some $i = k$ say. Hence

$$r(\pi, \delta) = \sum_i R(\theta_i, \delta)\pi(\theta_i) < \sum_i R(\theta_i, \delta_B^\pi)\pi(\theta_i) = r(\pi, \delta_B^\pi), \tag{11}$$

the inequality being strict due to the conditions that the Bayes Risk is finite and the prior places positive probability on all θ_i. But this contradicts the fact that δ_B^π is a Bayes rule.