36

Decision Theory

Decision theory is trivial, apart from computational details (just like playing chess!).

You have a choice of various actions, a. The world may be in one of many states $\mathbf{x}$; which one occurs may be influenced by your action. The world's state has a probability distribution $P(\mathbf{x} \mid a)$. Finally, there is a utility function $U(\mathbf{x}, a)$ which specifies the payoff you receive when the world is in state $\mathbf{x}$ and you chose action a.

The task of decision theory is to select the action that maximizes the expected utility,

$$\mathcal{E}[U \mid a] = \int \mathrm{d}^K \mathbf{x} \; U(\mathbf{x}, a) P(\mathbf{x} \mid a). \tag{36.1}$$

That's all. The computational problem is to maximize $\mathcal{E}[U \mid a]$ over a. [Pessimists may prefer to define a loss function L instead of a utility function U and minimize the expected loss.]

Is there anything more to be said about decision theory?

Well, in a real problem, the choice of an appropriate utility function may be quite difficult. Furthermore, when a sequence of actions is to be taken, with each action providing information about $\mathbf{x}$, we have to take into account the effect that this anticipated information may have on our subsequent actions. The resulting mixture of forward probability and inverse probability computations in a decision problem is distinctive. In a realistic problem such as playing a board game, the tree of possible cogitations and actions that must be considered becomes enormous, and 'doing the right thing' is not simple, because the expected utility of an action cannot be computed exactly (Russell and Wefald, 1991; Baum and Smith, 1993; Baum and Smith, 1997).

Let's explore an example.

▶ 36.1 Rational prospecting

Suppose you have the task of choosing the site for a Tanzanite mine. Your final action will be to select the site from a list of N sites. The nth site has a net value called the return x_n which is initially unknown, and will be found out exactly only after site n has been chosen. [x_n equals the revenue earned from selling the Tanzanite from that site, minus the costs of buying the site, paying the staff, and so forth.] At the outset, the return x_n has a probability distribution $P(x_n)$, based on the information already available.

Before you take your final action you have the opportunity to do some prospecting. Prospecting at the nth site has a cost c_n and yields data d_n which reduce the uncertainty about x_n. [We'll assume that the returns of

the N sites are unrelated to each other, and that prospecting at one site only yields information about that site and doesn't affect the return from that site.]

Your decision problem is:

> given the initial probability distributions $P(x_1)$, $P(x_2)$, ..., $P(x_N)$, first, decide whether to prospect, and at which sites; then, in the light of your prospecting results, choose which site to mine.

For simplicity, let's make everything in the problem Gaussian and focus on the question of whether to prospect once or not. We'll assume our utility function is linear in x_n; we wish to maximize our expected return. The utility function is

The notation $P(y) = \text{Normal}(y; \mu, \sigma^2)$ indicates that y has Gaussian distribution with mean μ and variance σ^2.

$$U = x_{n_a}, \tag{36.2}$$

if no prospecting is done, where n_a is the chosen 'action' site; and, if prospecting is done, the utility is

$$U = -c_{n_p} + x_{n_a}, \tag{36.3}$$

where n_p is the site at which prospecting took place.

The prior distribution of the return of site n is

$$P(x_n) = \text{Normal}(x_n; \mu_n, \sigma_n^2). \tag{36.4}$$

If you prospect at site n, the datum d_n is a noisy version of x_n:

$$P(d_n \,|\, x_n) = \text{Normal}(d_n; x_n, \sigma^2). \tag{36.5}$$

▷ Exercise 36.1.[2] Given these assumptions, show that the prior probability distribution of d_n is

$$P(d_n) = \text{Normal}(d_n; \mu_n, \sigma^2 + \sigma_n^2) \tag{36.6}$$

(mnemonic: when independent variables add, variances add), and that the posterior distribution of x_n given d_n is

$$P(x_n \,|\, d_n) = \text{Normal}\left(x_n; \mu_n', \sigma_n^{2\prime}\right) \tag{36.7}$$

where

$$\mu_n' = \frac{d_n/\sigma^2 + \mu_n/\sigma_n^2}{1/\sigma^2 + 1/\sigma_n^2} \quad \text{and} \quad \frac{1}{\sigma_n^{2\prime}} = \frac{1}{\sigma^2} + \frac{1}{\sigma_n^2} \tag{36.8}$$

(mnemonic: when Gaussians multiply, precisions add).

To start with, let's evaluate the expected utility if we do no prospecting (i.e., choose the site immediately); then we'll evaluate the expected utility if we first prospect at one site and then make our choice. From these two results we will be able to decide whether to prospect once or zero times, and, if we prospect once, at which site.

So, first we consider the expected utility without any prospecting.

Exercise 36.2.[2] Show that the optimal action, assuming no prospecting, is to select the site with biggest mean

$$n_a = \underset{n}{\text{argmax}}\ \mu_n, \tag{36.9}$$

and the expected utility of this action is

$$\mathcal{E}[U \,|\, \text{optimal } n] = \max_n \mu_n. \tag{36.10}$$

[If your intuition says 'surely the optimal decision should take into account the different uncertainties σ_n too?', the answer to this question is 'reasonable – if so, then the utility function should be *nonlinear* in x'.]

Now the exciting bit. Should we prospect? Once we have prospected at site n_p, we will choose the site using the decision rule (36.9) with the value of mean μ_{n_p} replaced by the updated value μ'_n given by (36.8). What makes the problem exciting is that we don't yet know the value of d_n, so we don't know what our action n_a will be; indeed the whole value of doing the prospecting comes from the fact that the outcome d_n may alter the action from the one that we would have taken in the absence of the experimental information.

From the expression for the new mean in terms of d_n (36.8), and the known variance of d_n (36.6), we can compute the probability distribution of the key quantity, μ'_n, and can work out the expected utility by integrating over all possible outcomes and their associated actions.

Exercise 36.3.[2] Show that the probability distribution of the new mean μ'_n (36.8) is Gaussian with mean μ_n and variance

$$s^2 \equiv \sigma_n^2 \frac{\sigma_n^2}{\sigma^2 + \sigma_n^2}. \tag{36.11}$$

Consider prospecting at site n. Let the biggest mean of the other sites be μ_1. When we obtain the new value of the mean, μ'_n, we will choose site n and get an expected return of μ'_n if $\mu'_n > \mu_1$, and we will choose site 1 and get an expected return of μ_1 if $\mu'_n < \mu_1$.

So the expected utility of prospecting at site n, then picking the best site, is

$$\mathcal{E}[U \,|\, \text{prospect at } n] = -c_n + P(\mu'_n < \mu_1)\,\mu_1 + \int_{\mu_1}^{\infty} \mathrm{d}\mu'_n \, \mu'_n \, \text{Normal}(\mu'_n; \mu_n, s^2). \tag{36.12}$$

The difference in utility between prospecting and not prospecting is the quantity of interest, and it depends on what we would have done without prospecting; and that depends on whether μ_1 is bigger than μ_n.

$$\mathcal{E}[U \,|\, \text{no prospecting}] \;=\; \begin{cases} -\mu_1 & \text{if } \mu_1 \geq \mu_n \\ -\mu_n & \text{if } \mu_1 \leq \mu_n. \end{cases} \tag{36.13}$$

So

$$\mathcal{E}[U \,|\, \text{prospect at } n] - \mathcal{E}[U \,|\, \text{no prospecting}] = \begin{cases} -c_n + \displaystyle\int_{\mu_1}^{\infty} \mathrm{d}\mu'_n \, (\mu'_n - \mu_1) \, \text{Normal}(\mu'_n; \mu_n, s^2) & \text{if } \mu_1 \geq \mu_n \\ -c_n + \displaystyle\int_{-\infty}^{\mu_1} \mathrm{d}\mu'_n \, (\mu_1 - \mu'_n) \, \text{Normal}(\mu'_n; \mu_n, s^2) & \text{if } \mu_1 \leq \mu_n. \end{cases} \tag{36.14}$$

We can plot the change in expected utility due to prospecting (omitting c_n) as a function of the difference $(\mu_n - \mu_1)$ (horizontal axis) and the initial standard deviation σ_n (vertical axis). In the figure the noise variance is $\sigma^2 = 1$.

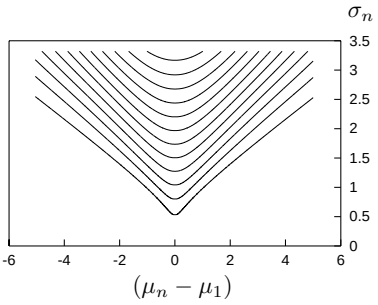

Figure 36.1. Contour plot of the gain in expected utility due to prospecting. The contours are equally spaced from 0.1 to 1.2 in steps of 0.1. To decide whether it is worth prospecting at site n, find the contour equal to c_n (the cost of prospecting); all points $[(\mu_n - \mu_1), \sigma_n]$ above that contour are worthwhile.

▶ 36.2 Further reading

If the world in which we act is a little more complicated than the prospecting problem – for example, if multiple iterations of prospecting are possible, and the cost of prospecting is uncertain – then finding the optimal balance between exploration and exploitation becomes a much harder computational problem. *Reinforcement learning* addresses approximate methods for this problem (Sutton and Barto, 1998).

▶ 36.3 Further exercises

▷ Exercise 36.4.[2] The four doors problem.

A new game show uses rules similar to those of the three doors (exercise 3.8 (p.57)), but there are four doors, and the host explains: 'First you will point to one of the doors, and then I will open one of the other doors, guaranteeing to choose a non-winner. Then you decide whether to stick with your original pick or switch to one of the remaining doors. Then I will open another non-winner (but never the current pick). You will then make your final decision by sticking with the door picked on the previous decision or by switching to the only other remaining door.'

What is the optimal strategy? Should you switch on the first opportunity? Should you switch on the second opportunity?

▷ Exercise 36.5.[3] One of the challenges of decision theory is figuring out exactly what the utility function is. The utility of money, for example, is notoriously nonlinear for most people.

In fact, the behaviour of many people cannot be captured by a coherent utility function, as illustrated by the *Allais paradox*, which runs as follows.

Which of these choices do you find most attractive?

A. £1 million guaranteed.
B. 89% chance of £1 million;
10% chance of £2.5 million;
1% chance of nothing.

Now consider these choices:

C. 89% chance of nothing;
11% chance of £1 million.
D. 90% chance of nothing;
10% chance of £2.5 million.

Many people prefer A to B, and, at the same time, D to C. Prove that these preferences are inconsistent with any utility function $U(x)$ for money.

Exercise 36.6.[4] Optimal stopping.

A large queue of N potential partners is waiting at your door, all asking to marry you. They have arrived in random order. As you meet each partner, you have to decide on the spot, based on the information so far, whether to marry them or say no. Each potential partner has a desirability d_n, which you find out if and when you meet them. You must marry one of them, but you are not allowed to go back to anyone you have said no to.

There are several ways to define the precise problem.

(a) Assuming your aim is to maximize the desirability d_n, i.e., your utility function is $d_{\hat{n}}$, where $\hat{n}$ is the partner selected, what strategy should you use?

(b) Assuming you wish very much to marry *the most desirable* person (i.e., your utility function is 1 if you achieve that, and zero otherwise); what strategy should you use?

(c) Assuming you wish very much to marry the most desirable person, and that your strategy will be 'strategy M':

> Strategy M – Meet the first M partners and say no to all of them. Memorize the maximum desirability $d_{\max}$ among them. Then meet the others in sequence, waiting until a partner with $d_n > d_{\max}$ comes along, and marry them. If none more desirable comes along, marry the final Nth partner (and feel miserable).

– what is the optimal value of M?

Exercise 36.7.[3] Regret as an objective function?

The preceding exercise (parts b and c) involved a utility function based on regret. If one married the tenth most desirable candidate, the utility function asserts that one would feel regret for having not chosen the most desirable.

Many people working in learning theory and decision theory use 'minimizing the maximal possible regret' as an objective function, but does this make sense?

Imagine that Fred has bought a lottery ticket, and offers to sell it to you before it's known whether the ticket is a winner. For simplicity say the probability that the ticket is a winner is 1/100, and if it is a winner, it is worth £10. Fred offers to sell you the ticket for £1. Do you buy it?

The possible actions are 'buy' and 'don't buy'. The utilities of the four possible action–outcome pairs are shown in table 36.2. I have assumed that the utility of small amounts of money for you is linear. If you don't buy the ticket then the utility is zero regardless of whether the ticket proves to be a winner. If you do buy the ticket you end up either losing one pound (with probability 99/100) or gaining nine (with probability 1/100). In the minimax regret community, actions are chosen to minimize the maximum possible regret. The four possible regret outcomes are shown in table 36.3. If you buy the ticket and it doesn't win, you have a regret of £1, because if you had not bought it you would have been £1 better off. If you do not buy the ticket and it wins, you have a regret of £9, because if you had bought it you would have been £9 better off. The action that minimizes the maximum possible regret is thus to buy the ticket.

	Action	
Outcome	Buy	Don't buy
No win	−1	0
Wins	+9	0

Table 36.2. Utility in the lottery ticket problem.

	Action	
Outcome	Buy	Don't buy
No win	1	0
Wins	0	9

Table 36.3. Regret in the lottery ticket problem.

Discuss whether this use of regret to choose actions can be philosophically justified.

The above problem can be turned into an investment portfolio decision problem by imagining that you have been given one pound to invest in two possible funds for one day: Fred's lottery fund, and the cash fund. If you put £f_1 into Fred's lottery fund, Fred promises to return £$9f_1$ to you if the lottery ticket is a winner, and otherwise nothing. The remaining £f_0 (with $f_0 = 1 - f_1$) is kept as cash. What is the best investment? Show that the minimax regret community will invest $f_1 = 9/10$ of their money in the high risk, high return lottery fund, and only $f_0 = 1/10$ in cash. Can this investment method be justified?

Exercise 36.8.[3] Gambling oddities (from Cover and Thomas (1991)). A horse race involving I horses occurs repeatedly, and you are obliged to bet all your money each time. Your bet at time t can be represented by

a normalized probability vector $\mathbf{b}$ multiplied by your money $m(t)$. The odds offered by the bookies are such that if horse i wins then your return is $m(t+1) = b_i o_i m(t)$. Assuming the bookies' odds are 'fair', that is,

$$\sum_i \frac{1}{o_i} = 1, \tag{36.15}$$

and assuming that the probability that horse i wins is p_i, work out the optimal betting strategy if your aim is *Cover's aim*, namely, to maximize the *expected value of* $\log m(T)$. Show that the optimal strategy sets $\mathbf{b}$ equal to $\mathbf{p}$, independent of the bookies' odds $\mathbf{o}$. Show that when this strategy is used, the money is expected to grow exponentially as:

$$2^{nW(\mathbf{b},\mathbf{p})} \tag{36.16}$$

where $W = \sum_i p_i \log b_i o_i$.

If you only bet once, is the optimal strategy any different?

Do you think this optimal strategy makes sense? Do you think that it's 'optimal', in common language, to ignore the bookies' odds? What can you conclude about 'Cover's aim'?

Exercise 36.9.[3] Two ordinary dice are thrown repeatedly; the outcome of each throw is the sum of the two numbers. Joe Shark, who says that 6 and 8 are his lucky numbers, bets even money that a 6 will be thrown before the first 7 is thrown. If you were a gambler, would you take the bet? What is your probability of winning? Joe then bets even money that an 8 will be thrown before the first 7 is thrown. Would you take the bet?

Having gained your confidence, Joe suggests combining the two bets into a single bet: he bets a larger sum, still at even odds, that an 8 and a 6 will be thrown before two 7s have been thrown. Would you take the bet? What is your probability of winning?