To: *C. M. Caves*
From: *C. M. Caves*
Subject: **Note on decision theory**

1997 May 21; extensively modified on 2003 January 10; modified again on 2006 November 15

Consider a decision among hypotheses H_i, which have prior probabilities $p(H_i)$. The decision among the hypotheses is based on data D that have conditional probabilities $p(D|H_i)$. The probability that hypothesis H_i is true, given the data D, is given by Bayes's theorem:

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D)} \; . \tag{1}$$

Although this probability sums up entirely one's knowledge about the hypotheses given the data, a decision cannot be based solely on these probabilities. One requires in addition a knowledge of the costs of the various decisions. These costs are quantified by a cost matrix $C(H_i|H_j)$, which is the cost of adopting hypothesis H_i when H_j is true. The average cost of adopting hypothesis H_i in the presence of data D is given by

$$\overline{C}(H_i|D) = \sum_j C(H_i|H_j)p(H_j|D) = \sum_j C(H_i|H_j)p(H_j)\frac{p(D|H_j)}{p(D)} \; . \tag{2}$$

Here the average is taken over hypotheses. Notice that the cost matrix and the prior probabilities enter as a product $C(H_i|H_j)p(H_j)$.

Now let E_D be the decision function: for data D, E_D is the adopted hypothesis. The expected cost, averaged over hypotheses and data, is given by

$$\overline{C} = \sum_{D,j} C(E_D|H_j)p(H_j|D)p(D) = \sum_D p(D)\overline{C}(E_D|D) \; . \tag{3}$$

Minimizing the expected cost is thus equivalent to choosing for each data set D the hypothesis $E_D = H_i$ that has the smallest average cost $\overline{C}(H_i|D)$ for that data. Since the cost function and the prior probabilities appear as a product in $\overline{C}(H_i|D)$, one cannot separate the effects of the cost function and the priors on the decision function.

We can also write the expected cost as

$$\overline{C} = \sum_j p(H_j)\overline{C}(H_j) \; , \tag{4}$$

where

$$\overline{C}(H_j) = \sum_D C(E_D|H_j)p(D|H_j) \tag{5}$$

is the average cost given that hypothesis H_j is true. In Eq. (5) the average is taken over the data.

One can give a neater formulation by noting that any decision function partitions the data sets into classes $D_i = \{D \mid E_D = H_i\}$, one class for each hypothesis. Only the class is relevant to the decision, so we can regard the classes as the outcomes of the data collection. The outcome D_i leads to a decision for the corresponding hypothesis H_i. With this simplification, the average cost for adopting hypothesis H_i is given by

$$\overline{C}_i = \sum_j C(H_i|H_j)p(H_j|D_i) = \sum_j C(H_i|H_j)p(H_j)\frac{p(D_i|H_j)}{p(D_i)} \;, \tag{6}$$

and the expected cost is

$$\overline{C} = \sum_i \overline{C}_i p(D_i) = \sum_{i,j} C(H_i|H_j)p(H_j|D_i)p(D_i) = \sum_{i,j} C(H_i|H_j)p(D_i|H_j)p(H_j) \;. \tag{7}$$

Neater though this formulation is, it obscures the process of optimizing over decision functions, so we revert to the previous formulation for the remainder of the document. The neater formulation comes into its own in quantum decision theory.

One very useful cost function assigns no cost to correct decisions and and a uniform, positive cost to every error:

$$C(H_i|H_j) = 1 - \delta_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases} \tag{8}$$

In this case, the average cost of adopting hypothesis H_i in the presence of data D,

$$\overline{C}(H_i|D) = \sum_{j \neq i} p(H_j|D) = 1 - p(H_i|D) \;, \tag{9}$$

is simply the probability of having made an error. The expected cost

$$\overline{C} = \sum_D p(D)\overline{C}(E_D|D) = P_e \tag{10}$$

is the total error probability. The total error probability is minimized by using a decision function that given the data D, chooses the hypothesis with the highest posterior probability.

A similar, but more complicated cost function allows for the possibility of making no decision when the probability of error is too high. To formulate this cost function, we need to introduce (formally) an additional "no-decision" hypothesis H_0. We know H_0 isn't true, so we assign it zero prior probability, i.e., $p(H_0) = 0$; this choice makes the associated costs irrelevant, so we might just as well choose them to be zero, i.e., $C(H_0|H_0) = C(H_j|H_0) = 0$. Now, letting Roman indices designate the actual hypotheses, we choose the rest of the cost function to be $C(H_0|H_j) = 1$, and $C(H_i|H_j) = (1 - \delta_{ij})/A$, where A is a nonnegative constant. The cost of making no decision sets the unit cost, and the cost of making any error is $1/A$. Now the average cost for adopting the no-decision hypothesis in the presence of data D is

$$\overline{C}_0(D) = \sum_j p(H_j|D) = 1 \;, \tag{11}$$

and the average cost of adopting hypothesis H_i in the presence of data D is

$$\overline{C}(H_i|D) = \frac{1}{A}\sum_{j\neq i} p(H_j|D) = \frac{1}{A}\big(1 - p(H_i|D)\big) \ . \tag{12}$$

Minimizing the expected cost thus means choosing the hypothesis H_i with the biggest posterior probability $p(H_i|D)$, provided that the cost of that decision is no bigger than the cost of no decision, i.e.,

$$\frac{1}{A}\big(1 - p(H_i|D) \leq 1 \qquad \Longleftrightarrow \qquad p(H_i|D) \geq 1 - A \ ; \tag{13}$$

Otherwise one makes no decision by choosing H_0. The constant A is a threshold for making a decision: if the error probability is greater than A, then you make no decision. For $A \geq 1$, this decision function reduces to the previous one. In the limit $A \to 0$, the decision function becomes completely averse to errors; it chooses an actual hypothesis if and only if that hypothesis is confirmed unambiguously by the data and otherwise makes no decision.

We can formulate this differently by using the neater formulation where the data is divided into classes corresponding to the different decisions. In this case the expected cost (3) becomes

$$\begin{aligned}
\overline{C} &= \sum_{\alpha,\beta} C(H_\alpha|H_\beta)p(D_\alpha|H_\beta)p(H_\beta) \\
&= \sum_{\alpha,j} C(H_\alpha|H_j)p(D_\alpha|H_j)p(H_j) \\
&= \sum_{j} p(D_0|H_j)p(H_j) + \frac{1}{A}\sum_{i,j}(1-\delta_{ij})p(D_i|H_j)p(H_j) \\
&= p(D_0) + \frac{1}{A}\sum_{j} p(H_j)\sum_{i\neq j} p(D_i|H_j) \ .
\end{aligned} \tag{14}$$

When $A \to 0$, minimizing the expected cost requires that we divide the data up so that $p(D_i|H_j) = 0$ for $i \neq j$, i.e., no errors, and then the expected cost becomes the probability for all the data for which there are errors and for which we therefore make no decision.