SOME INTERESTING DECISION FUNCTIONS

Michael F. Capobianco
Polytechnic Institute of Brooklyn

Basic Considerations of Statistical Decision Theory

The problem is to decide which of q possible states of nature $\theta_1, \theta_2, \ldots, \theta_q$ is the true one by observing the outcome of some experiment which has n possible outcomes $x_1, x_2, \ldots, x_n$. For each θ_j there is a probability distribution vector

$$P_j = \begin{bmatrix} p_{1j} \\ p_{2j} \\ \vdots \\ p_{nj} \end{bmatrix}$$

where $p_{ij} = P(x_i | \theta_j)$. We also form a loss vector for each θ_j

$$W_j = \begin{bmatrix} w_{1j} \\ w_{2j} \\ \vdots \\ w_{qj} \end{bmatrix}$$

where w_{ij} = the loss incurred in making decision d_i when θ_j is the true state of nature; $w_{ij} \geq 0$ and $= 0$ if and only if $i = j$. In order to make decisions we need a mechanism for choosing a θ_j upon observing an x_i. Such a mechanism is called a decision function and can be represented by a matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \cdots & a_{1n} \\ \vdots & & \\ a_{q1} & \cdots\cdots\cdots & a_{qn} \end{bmatrix}$$

where $a_{ij} = P(d_i | x_j)$.

The idea is to find a matrix A that in some sense minimizes the loss. We can compute the expected loss, called the risk, for any A under a given state of nature. This is denoted by $R(A, \theta_j)$ and

$$R(A, \theta_j) = W_j' \; A \; P_j$$

where W_j' is the transpose of W_j. In the absence of any further information one way of choosing a single decision function is by using the minimax criterion, i.e., choose the function with risk equal to $\min_A \max_\theta R(A, \theta)$.

Another possibility is to use the maximum likelihood decision function, i.e., choose the function A such that

$$a_{ij} = \begin{cases} 1 & \text{if } p_{ji} \geq p_{j\ell} \text{ for all } \ell \\ 0 & \text{otherwise} \end{cases}$$

This is an example of an non-randomized decision function; each column of A has a single entry equal to 1 and all other entries equal to 0. In such a case we say that A is non-randomized. Clearly each column of any A must add up to 1.

If an a priori distribution is available, i.e., if one has a vector

$$P = \begin{bmatrix} P(\theta_1) \\ P(\theta_2) \\ \vdots \\ P(\theta_q) \end{bmatrix}$$

where $P(\theta_i)$ = the probability that θ_i is the true state of nature, one can then find the expected risk (or Bayes risk)

$$\sum W_j' \; AP_j \; P$$

where $\sum = \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix}$, a row vector of q 1's. One now chooses the decision function with the minimum expected risk. This is called the Bayes decision function.

One property that a decision function should have is that of admissibility. To explain this term we introduce first the notion of dominance. If $R(A, \theta_j) \leq R(B, \theta_j)$ for all θ_j and strict inequality holds for at least one θ_j, then A is said to dominate B. A decision function is admissible if it is not dominated by any other one.

Proportional Likelihood Decision Function

We argue as follows: There seems to be a weakness in the maximum likelihood criteria in that it chooses that state of nature θ_j which yields the observed x_i with the highest probability, even though some other state of nature may yield x_i with a probability almost as high. It seems reasonable that it would be better to give all states of nature a chance of being chosen which is proportional to their respective probabilities of yielding x_i. We, therefore, propose to form the matrix A with

$$a_{ij} = \frac{P(x_j | \theta_i)}{\sum_{i=1}^{q} P(x_j | \theta_i)} = \frac{p_{ji}}{\sum_{i=1}^{q} p_{ji}}$$

The following example shows that such a decision function may be admissible.

$$P_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} \qquad P_2 = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{1}{6} \end{bmatrix}$$

$$W_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad W_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad A_6 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad A_7 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \qquad A_8 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Sample Calculation:

$$R(A_4, \theta_1) = W_1' A_4 P_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} = \frac{2}{3}$$

Tabulation

	$R(A, \theta_1)$	$R(A, \theta_2)$
A_1	0	1
A_2	$\frac{1}{3}$	$\frac{5}{6}$
A_3	$\frac{1}{3}$	$\frac{1}{2}$
A_4	$\frac{2}{3}$	$\frac{1}{3}$
A_5	$\frac{1}{3}$	$\frac{2}{3}$
A_6	$\frac{2}{3}$	$\frac{1}{2}$
A_7	$\frac{2}{3}$	$\frac{1}{6}$
A_8	1	0

Note that A_2, A_4, A_5 and A_6 are inadmissible. Also A_3 and A_7 are both maximum likelihood functions. Now for the decision function proposed above.

$$A = \begin{bmatrix} \frac{1}{2} & \frac{2}{5} & \frac{2}{3} \\ \frac{1}{2} & \frac{3}{5} & \frac{1}{3} \end{bmatrix}$$

Sample Calculation

$$a_{22} = \frac{\frac{1}{2}}{\frac{1}{3} + \frac{1}{2}} = \frac{\frac{1}{2}}{\frac{5}{6}} = \frac{3}{5}$$

Now

$$R:(A, \theta_1) = W_1' AP_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{2}{5} & \frac{2}{3} \\ \frac{1}{2} & \frac{3}{5} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$

$$= \frac{1}{6} + \frac{1}{5} + \frac{1}{9} = \frac{43}{90},$$

and

$$R(A, \theta_2) = W_2' AP_2 = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{2}{5} & \frac{2}{3} \\ \frac{1}{2} & \frac{3}{5} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{1}{6} \end{bmatrix}$$

$$= \frac{1}{6} + \frac{1}{5} + \frac{1}{9} = \frac{43}{90}$$

so that A is admissible. Note that all losses were taken as equal in order to make things most favorable for a maximum likelihood decision function. If we make this assumption in general i.e., suppose W_i is a q x 1 vector with 0 in the i^{th} position and the quantity W in all other positions, then we can see that proportional likelihood decision functions become dominated by maximum likelihood decision functions, and in fact have a risk of W for each θ_i as $q \to \infty$. We say, therefore, that they are asymptotically inadmissible. The proof is as follows: Let A be the proportional likelihood function

$$R(A, \theta_i) = W_i' AP_i = W_i' \left[\sum_{k=1}^{n} a_{jk} p_{ki} \right]$$

$$= W \sum_{j \neq i}^{n} \sum_{k=1}^{n} a_{jk} p_{ki}$$

$$= W \sum_{j \neq i}^{n} \sum_{k=1}^{n} \frac{p_{kj} p_{ki}}{\sum_{\ell=1}^{q} p_{k\ell}} = W \sum_{k=1}^{n} p_{ki} \sum_{j \neq i} \frac{p_{kj}}{\sum_{\ell=1}^{q} p_{k\ell}}$$

$$= W \sum_{k=1}^{n} p_{ki} \frac{\sum_{j=1}^{q} p_{kj} - p_{ki}}{\sum_{\ell=1}^{q} p_{k\ell}} = W \sum_{k=1}^{n} p_{ki} (1 - a_{ik}),$$

which approaches

$$W \sum_{k=1}^{n} p_{ki} = W$$

as $q \to \infty$ because

$$a_{ik} = \frac{p_{ki}}{\sum_{\ell=1}^{q} p_{k\ell}} \leq \frac{1}{\sum_{\ell=1}^{q} p_{k\ell}} \to 0$$

as $q \to \infty$.

Now for a maximum likelihood decision function A,

$$R(A, \theta_i) = W_1' AP_i = W \sum_{j \neq i} \sum_{k=1}^{n} a_{jk} p_{ki}$$

where

$$a_{jk} = \begin{cases} 1 & \text{if } p_{kj} \geq p_{k\ell} \text{ for all } \ell \\ 0 & \text{otherwise .} \end{cases}$$

Therefore,

$$R(A, \theta_i) = W \sum_{k=1}^{n} p_{ki} \sum_{j \neq i} a_{jk} = W \sum_{k=1}^{n} p_{ki}(1 - a_{ik})$$

$$= W \sum p_{ki} ,$$

where the sum is taken over all k such that $p_{ki} < p_{k\ell}$ for some ℓ. This is less than W for at least one θ_i.

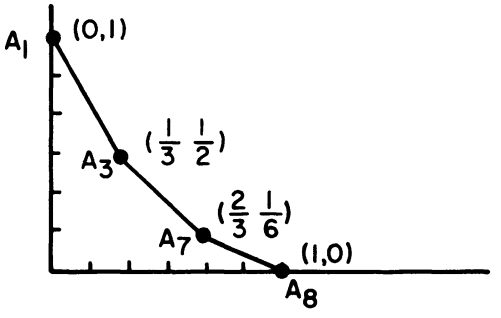
Proportional Bayes Decision Functions

We will assume that in our above example we have on a priori distribution given by

$$P = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}$$

and we will show how to find the Bayes decision function.

If we plot $R(A, \theta_1)$ against $R(A, \theta_2)$ for all the admissible A's and join these points by a broken line we have



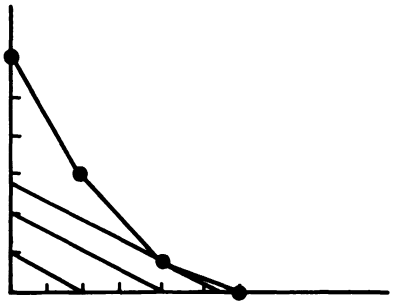
It can easily be shown that any admissible randomized decision function can be obtained from the non-randomized admissible ones in the following way: Select two non-randomized admissible functions, say A_3 and A_7, which are jointed by a straight line segment. Choose A_3 with probability α and A_7 with probability $1 - \alpha$. This yields a randomized function A such that

$$R(A, \theta_i) = \alpha R(A_3, \theta_i) + (1 - \alpha) R(A_7, \theta_i), \; i = 1, 2$$

The point $(R(A, \theta_1), R(A, \theta_2))$ lies on the line segment jointing A_3 and A_7. Hence, the points of the entire broken line are the risk pairs for all admissible functions. To find which of these is the Bayes function we form the equation

$$\frac{1}{3} x + \frac{2}{3} y = k$$

where, x is the risk under θ_1, and y is the risk under θ_2, and let k vary from zero up until this line first touches our broken line of admissible functions. As we can see from the diagram below, A_7 is the Bayes decision function.



A problem arises in this procedure if the line

$$P(\theta_1) x + P(\theta_2) y = k \qquad (1)$$

has a slope equal to that of one of the line segments of admissible functions. In such a case there will be an infinite number of Bayes functions. In our example this would happen if $P(\theta_1) = P(\theta_2) = \frac{1}{2}$. Then (1) has a slope of -1, and so does the line segment joining A_3 and A_7. We now must choose one of the functions along this segment. To do this we argue as follows:

$$R(A_3, \theta_1) < R(A_7, \theta_1)$$

while

$$R(A_7, \theta_2) < R(A_3, \theta_2) .$$

Therefore if θ_1 were the true state of nature, A_3 would be better, while if θ_2 were true, A_7 would be better. Hence, we propose choosing A_3 with probability $P(\theta_1)$ and A_7 with probability $P(\theta_2)$. It seems that the resulting function is in some sense better than either A_3 or A_7, but it is not clear how this can be expressed mathematically.

The procedures discussed can all be generalized to more than two states of nature. We used the above example in the interests of clarity of exposition.