MMT Batch 2
10 Feb 1966

BAYESIAN STATISTICS, INFORMATION, AND SIGNAL DETECTION

ROC curves and parametric representation

In general, $A(\mathrm{II})$ contains all the available information about the world required for a decision on the state of II, while $G_D(\mathrm{II})$ gives all the relevant information gained from the observation D.

If the universe (II) consists of two hypotheses $\{H, J\}$, the assessment function is

$$A(\mathrm{II}) = \{c P_D(H), c P_D(J)\}$$

Because $A(\mathrm{II})$ is invariant with changes in the multiplier, c, all the information is contained in the assessment ratio

$$\frac{P_D(H)}{P_D(J)} = \frac{P(H)}{P(J)} \frac{P_H(D)}{P_J(D)} \underset{\text{def}}{=} \alpha_D$$

$\lambda_D = \dfrac{P_H(D)}{P_J(D)}$ is usually called the "likelihood ratio" and is often taken to be the important quantity in the detection situation, instead of the assessment ratio. If $\dfrac{P(H)}{P(J)}$ is taken to be unity, the likelihood ratio is equal to the assessment ratio.

We want to find for some hypothesis pair $\{H, J\}$ how well we are likely to be able to correctly identify the correct hypothesis by taking some observation D which may have the possible results $\{D_1, D_2 \cdots D \cdots D_n\}$. For any particular D, all the relevant information is contained in α_D. We thus need to find how α_D is distributed if H is true and how it is distributed if J is true. Either these distributions or any function of them which

permits their recovery, completely specifies how well H and J may be distinguished by an observation from D.

It is convenient to assume a uniform prior assessment function $\alpha = 1$, so that we can concentrate on λ_D instead of the numerically equal α_D.

A rational decision ~~whether~~ to act as if H were true must be based on whether α_D (or λ_D) exceeds some criterion α_c (or λ_c).

If H is true λ_D exceeds λ_c with some probability $P_H(\lambda_D > \lambda_c)$

If J is true λ_D exceeds λ_c with some other probability $P_J(\lambda_D > \lambda_c)$

The information relevant to discriminating between H and J is contained in these Probabilities, just as much as in the distributions of λ_D under the conditions H or J being true, since

$$\int_{\lambda_0=\lambda_c}^{\infty} P_H(\lambda_D = \lambda_0)\, d\lambda_0 = P_H(\lambda_D \geq \lambda_c).$$

Furthermore, all the information is contained in just one of the two distributions, since

$$\frac{P_H(\lambda_D = \lambda_c)}{P_J(\lambda_D = \lambda_c)} = \lambda_c$$

Proof:

$$P_H(\lambda_D = \lambda_c) = \sum_D P_H(D \text{ where } \lambda_D = \lambda_c) = \sum_D \lambda_c P_H(D \text{ where } \lambda_D = \lambda_c)$$

$$= \lambda_c \sum_D P_H(D \text{ where } \lambda_D = \lambda_c)$$

$$= \lambda_c P_J(\lambda_D = \lambda_c)$$

All the information about the distribution of λ_D under the two

hypotheses is therefore contained in a plot of either $P(\lambda_D=\lambda_c)$ or $P(\lambda_D>\lambda_c)$ against λ_c under either hypothesis.

Often, in a practical situation, it is possible in principle to find the probability that H is chosen if H actually is true (a signal is detected when it is there), without knowing what value of λ_c is implied. We then can, by varying the unknown λ_c obtain a function of $P_H(\lambda_D>\lambda_c)$ as a function of some z_c with an unknown monotonic relationship with λ_c. Knowing only z_c, not λ_c, we cannot infer for any point a value for $P_J(\lambda_D>\lambda_c)$ and hence we cannot infer any values of λ_D and have no estimate of how well the two hypotheses may be distinguished.

We can, again in principle, determine $P_J(\lambda_D>\lambda_c)$ directly for each z_c. Knowing, for each z_c, both $P_H(\lambda_D>\lambda_c)$ and $P_J(\lambda_D>\lambda_c)$ we can infer λ_c from the integral relationships

$$\int_{\lambda_0=\lambda_c}^{\infty} P_H(\lambda_D=\lambda_0)\,d\lambda_0 = P_H(\lambda_D>\lambda_c) = \int_{z=z_c}^{\infty} P_H(\lambda_D=\lambda_0)\frac{d\lambda_0}{dz}\,dz$$

whence

$$\frac{\frac{d}{dz_c}P_H(\lambda_D>\lambda_c)}{\frac{d}{dz_c}P_J(\lambda_D>\lambda_c)} = \frac{P_H(\lambda_D=\lambda_c)\frac{d\lambda_0}{dz}_{\lambda_0=\lambda_c}}{P_J(\lambda_D=\lambda_c)\frac{d\lambda_0}{dz}_{\lambda_0=\lambda_c}}$$

$$= \frac{P_H(\lambda_D=\lambda_c)}{P_J(\lambda_D=\lambda_c)}$$

$$= \lambda_c$$

But

$$\frac{\frac{d}{dz_c} P_H(\lambda > \lambda_c)}{\frac{d}{dz_c} P_J(\lambda > \lambda_c)} = \frac{d(P_H(\lambda_D > \lambda_c))}{d(P_J(\lambda_D > \lambda_c))}$$

So that all the relevant information about the distribution of λ_D under the two hypotheses is contained in a plot of $P_H(\lambda_D > \lambda_c) = y$ against $P_J(\lambda_D > \lambda_c) = x$, without any reference to z_c or explicit reference to λ_c. Such a plot, because of its history in radar technology, is known commonly as an ROC (Reciever operating characteristic).

With fixed hypotheses $\{H, J\}$ and a fixed type of observation D, there is in some sense a fixed discrimination a priori possible. Hence, the ROC is sometimes called an "isosensitivity curve".

The slope of the ROC, $\frac{dy}{dx} = \frac{d P_H(\lambda_D > \lambda_c)}{d P_J(\lambda_D > \lambda_c)}$ is called the criterion, and is often symbolized by β rather than λ_c.

NOTE IN SUMMARY

1) All the information relevant to a decision about H is contained in $A(H)$ or in $\alpha_D = \frac{P_D(H)}{P_D(J)}$

2) To determine all possible discrimination behaviour exhibited by a rational observer it is necessary to have the information conveyed by the distributions of α_D under H and under J. The part of this information given by the observation is contained in λ_D.

3) The ROC contains all the information in the distributions of λ_D, which may be recovered from it.

4) The points on the ROC may, in principle, be observed directly by experiment.

5) Observation 4 makes the ROC, which is only one of many ways of representing the relevant information, very important in practical situations.

6) If we are concerned with the intrinsic discriminability of H and J by means of observations from D, λ_c does not enter into consideration. The shape and placement of the ROC contain the information. λ_c pertains to actions by an observer with a particular value system.

PARAMETRIC REPRESENTATION OF DISCRIMINATION

The form of the ROC is constrained. It must pass through (0,0) and (1,1) and as it is traced out, its slope must never increase. Both constraints are apparent from the integral representations of $P_H(\lambda_D > \lambda_c)$ and $P_J(\lambda_D > \lambda_c)$. Within these constraints an infinity of possible ROC's may pass through any specified point in the ROC space (the space in which the coordinates are $P_H(\lambda_D > \lambda_c)$ and $P_J(\lambda_D > \lambda_c)$).

Some families of hypothesis pairs lend themselves to single parameter representations. We shall deal often with the case where each hypothesis is that the "signal" may be specified by a finite number of precisely determined real numbers, while the observation set consists of these numbers with a random Gaussian (normal) variable added to each. In such conditions, a single parameter, called d' by the Michigan school, completely specifies the λ distributions.

Often, also, we are not very interested in the entire ROC, but want a parameter which indicates more or less precisely how easy the discrimination is. One such parameter is the probability that the correct choice can be made between the two hypothesis given an observation for each one separately. This is the common 2-alternative forced choice (2AFC) experiment. Overall, the probability of a correct choice in a 2AFC experiment gives a quick and useful measure of the likely ease of discrimination between the hypotheses, without giving indications of how often observations will occur that give strong discrimination or weak discrimination. We shall consider further the use of this parameter.

Sometimes we are not interested in the overall discriminability of the hypotheses, but rather we want to know how often easy discriminations occur and how often the task will be difficult. Single parameter representations of families of ROC's sometimes can be found to answer such questions.

DAVE GREEN'S THEOREM for 2-AFC EXPERIMENTS

An observation is made from each hypothesis (ie one presentation of signal H and one of signal J) but it is not known to the observer which observation was of which hypothesis (signal).

The optimum procedure to get most correct judgements is to say H for that observation giving greater α_D. Say for observation 1, $\alpha_D = \alpha_1$, and for observation 2, $\alpha_D = \alpha_2$.

$P(\text{correct}) = P_H(\alpha_1 > \alpha_2)$ if H was true in observation 1

Consider only the case in which H is as likely to be in observation 1 as in observation 2. We need then only consider those instances in

which H is actually in observation 1, since the probabilities are symmetric.

If J gives $\alpha = \alpha_J$

$$P(\text{correct}) = P_H(\alpha > \alpha_J)$$

$$= \int_{\alpha_J}^{\infty} P_H(\alpha = \alpha_c)\, d\alpha_c$$

This occurs with probability $P_J(\alpha = \alpha_J)$

Then over all observations

$$P(\text{correct}) = \int_{\alpha_J=0}^{\infty} P_J(\alpha = \alpha_J)\left[\int_{\alpha_c=\alpha_J}^{\infty} P_H(\alpha = \alpha_c)\, d\alpha_c\right] d\alpha_J$$

$$= \int_{\alpha_J=0}^{\infty} P_H(\alpha > \alpha_J)\; P_J(\alpha = \alpha_J)\, d\alpha_J$$

$$= \int_0^1 P_H(\alpha > \alpha_J)\; d\, P_J(\alpha > \alpha_J)$$

$$= \int_0^1 y\, dx \quad \text{in ROC space}$$

which is the area under the ROC curve

(Dave Green presented this remarkable result at the Psychonomic Society meeting at Niagara Falls in 1964)

Note: The derivation of Green's theorem was carried out in terms of α rather than λ. This was to demonstrate the interchangeability of the two measures when the prior assessment ratio is unity. Either derivation can, without this restriction, be done either with α or with λ, though decisions must be based on α.

MMT Batch 2
10 Feb 1966

Green's theorem indicates that the area under an ROC curve is a fundamental parameter to represent the intrinsic discriminability of two hypotheses by a particular set of observations. Apart from the fact that it gives a good overall representation of the detection capabilities of the observer, it has little to recommend it. It has to be computed for each ROC separately, it is not additive, either directly or through any obvious transform, and it has a range from .5 to 1 rather than from 0 to ∞ as would be more desireable for a measure of discriminability. In its favour is the fact that it is based on the ROC, which is in principle measureable by experiment.

What is really needed is some parametric representation of an ROC with a monotonic functional relationship to the area, but which is additive, and has a range from 0 to ∞. At the moment I know of no such rational parameter, but perhaps by the end of this seminar we can develope one.

Parametric representations are known for particular families of ROCs. The most important, because it is most used in the literature and is well tabulated, is d'.

d' is well defined in particular for the general case where the hypotheses concern which of two exactly defined waveforms are transmitted through a noisy channel where the noise is additive and Gaussian. It is defined more generally here, and the statements in energy terms included where appropriate.

MMT Batch 2
10 Feb 1966

Definition of d' and Derivation in terms of Energy

Definition: If z is some monotonic transform of the likelihood ratio L such that $P_H(z \geq z_c)$ and $P_J(z = z_c)$ are both Gaussian distributions with unit variance as a function of z_c, then d' is the separation of the means of the two distributions.

This definition means that if a transformation $z = z(L)$ can be found that yields the two normal distributions, then d' is defined for those hypotheses and the available set of observations. Otherwise d' is not defined.

Evaluation of d' for a particular family of hypotheses

Let an observation X consist of $2WT+1$ independent elements $\{X_0, X_1 \cdots X_K \cdots X_{2WT}\}$. ($2WT+1$ is chosen because the example is drawn from communications, though the formulation does not depend on this fact. If $\{X_0, X_1 \cdots X_{2WT}\}$ are samples of the voltage equally spaced in time (where if the voltage is measured across a 1 ohm resistor, ΣX_K^2 represents the energy of the observation) then these $2WT+1$ independent measurements just suffice to specify a continuous waveform from a source of bandwidth W operating for time T).

Each element $\{X_K\}$ is a noisy representation either of a "signal" element S_K, or of a "no signal" element n_K, which we shall assume to be zero. The generalization to n_K finite will be obvious. For each S_K, substitute $e_K = S_K - n_K$, and for each n_K, substitute zero.

The noise is Gaussian, with mean zero and variance σ^2, so that if the observation X is of signal plus noise (condition SN),

$$P_{SN}(X_K) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_K - S_K)^2}{2\sigma^2}}$$ which is Gaussian with

mean S_k and variance σ^2, while if the observation X is of noise and no signal (condition N)

$$P_N(X_k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_k)^2}{2\sigma^2}}$$

These two probability distributions not only describe X_k, but are the elements of $G_D(H) = \{P_H(D), P_J(D)\}$ where D is a particular value of X_k observed. If the prior assessment ratio is unity, we have

$$A_{X_k}(H) = \left\{ e^{-\frac{(X_k - S_k)^2}{2\sigma^2}}, e^{-\frac{X_k^2}{2\sigma^2}} \right\} \quad \text{where } H = \{SN, N\}$$
$$\left(= G_{X_k}(H)\right)$$

As discussed earlier in this batch of notes, only L_{X_k}, the likelihood ratio, or α_{X_k} the assessment ratio, is needed when the universe contains only 2 hypotheses. Here, the log-likelihood-ratio $= \mathcal{L}_{X_k} = \ln(L_{X_k})$ is most easy to work with, because it is additive over observations and because the elements of $G_{X_k}(H)$ are exponentials.

$$\mathcal{L}_{X_k} = \frac{-(X_k - S_k)^2 + X_k^2}{2\sigma^2} = \frac{X_k S_k}{\sigma^2} - \frac{S_k^2}{2\sigma^2} = \frac{(X_k - S_k) S_k}{\sigma^2} + \frac{S_k^2}{2\sigma^2}$$

Adding the evidence from $X_0, X_1 \cdots X_k \cdots X_{2WT}$,

$$\mathcal{L}_X = \sum_0^{2WT} \mathcal{L}_{X_k} = \Sigma \frac{X_k S_k}{\sigma^2} - \Sigma \frac{S_k^2}{2\sigma^2} \qquad (1)$$

$$= \Sigma \frac{(X_k - S_k) S_k}{\sigma^2} + \Sigma \frac{S_k^2}{2\sigma^2} \qquad (2)$$

If N is true (no signal) then X_k is normally distributed with mean zero and variance σ^2, so that $\frac{X_k S_k}{\sigma^2}$ is normally distributed with mean zero and variance $\frac{S_k^2}{\sigma^2}$, which means that $\mathcal{L}_{X_k}$ is normally distributed with mean $\frac{S_k^2}{2\sigma^2}$ and variance $\frac{S_k^2}{\sigma^2}$.

The sum of a number of independent Gaussian variables is itself Gaussian, the mean being the sum of the means and the variance being the sum of the variances

Hence

$\mathcal{L}_{N,X}$ is Gaussian with mean $-\frac{E}{2\sigma^2}$ and variance $\frac{E}{\sigma^2}$

where $E = \sum s_k^2$

By the same argument, if SN is true $(x_k - s_k)$ is Gaussian with mean zero and variance $\frac{s_k^2}{\sigma^2}$. Using (2)

$\mathcal{L}_{SN,X}$ is Gaussian with mean $\frac{E}{2\sigma^2}$ and variance $\frac{E}{\sigma^2}$

Changing scale by a factor of $\sqrt{\frac{E}{\sigma^2}}$ we find $z = \sqrt{\frac{\sigma^2}{E}} \ln L$ to be a variable monotonically related to L, on which the criterial distributions

$P_N(z = z_c)$ and $P_{SN}(z = z_c)$

are Gaussian with unit variance and means $-\frac{1}{2}\sqrt{\frac{E}{\sigma^2}}$ and $+\frac{1}{2}\sqrt{\frac{E}{\sigma^2}}$ respectively.

According to the definition

$$d' = \sqrt{\frac{E}{\sigma^2}}$$

If we identify the numbers s_k, x_k, σ^2, etc. with voltages measured across a 1 ohm resistor and with energies developed in the resistor, we can write:

Average noise energy per sample (mean square noise value) $= \sigma^2$

Total noise energy $= (2WT+1)\sigma^2 \approx 2WT\sigma^2$

Noise power $= 2W\sigma^2$

Noise power per unit bandwidth $= 2\sigma^2 \underset{\text{def}}{=} N_0$

Energy in a sample of the undistorted signal $= s_k^2$

Total energy in the undistorted signal $= \Sigma s_k^2 = E$

In energy terms

$$d' = \sqrt{\frac{E}{N_0/2}} = \sqrt{\frac{2E}{N_0}}$$ which is the standard form

If SN and N refer not to "signal" and "no signal" then $(s_k - n_k)$ must be substituted for s_k, and E is now the energy in the signal (waveform) obtained by point for point subtraction of the N signal from the SN signal. This energy E is called the "difference energy" of the two signals, and must not be confused with the difference of the energies of the two signals.

This comment shows the formal identity between detection and discrimination problems.

MMT Batch 2
10 Feb 1966

INFORMATION

If we continue to talk in terms of waveforms and energy, we can identify the observation as that of the output of a noisy transmission channel whose input is one of the two signals. The lack of knowledge of the observer about the signals is the distortion in the channel.

In the case considered in the derivation of $d' = \sqrt{\frac{2E}{N_o}}$, the distortion consisted of additive Gaussian noise, and apart from this noise the observer knew exactly what would be at the channel output for each of the (2) possible inputs.

Consider the signal $S_{\frac{1}{2}} = \left\{ \frac{S_0}{2}, \frac{S_1}{2} \cdots \frac{S_{2WT}}{2} \right\}$

This signal has a difference energy $\frac{E}{4}$ from either N or SN, where N is $\{0, 0, \cdots 0\}$ and SN is $\{S_0, S_1, \cdots S_{2WT}\}$ and is therefore equally ~~detectable~~ discriminable from either, and from all signals with the same difference energy.

We can define an ~~el~~ ensemble of signals having difference energy $\frac{E}{4}$ from $S_{1/2}$, or an informationally equivalent set of signals formed from this ensemble by subtracting $\frac{S_k}{2}$ from the k^{th} element of each. Each signal of the derived ensemble has a total energy $E/4$.

Shannon (Math. Theory of Comm.) gives the information transmission channel capacity for ~~each second~~ a transmitter of power P through noise of power N, where both signal and noise have bandwidth W and last for time T,

as

$$U_c = WT \log_2\left(\frac{P+N}{N}\right)$$

The ensemble of signals of power P transmitted for time T is the ensemble of signals having energy PT. Our ensemble has energy $E/4$ or power $\frac{E}{4T}$. The noise power $N = N_0 W$ where N_0 is the noise power per unit bandwidth, as before.

Hence

$$U_c = WT \log_2\left(\frac{E/4T + WN_0}{WN_0}\right)$$

$$= WT \log_2\left(\frac{1}{8WT} \cdot \frac{2E}{N_0} + 1\right)$$

$$= \log_2\left(\frac{1}{WT}\frac{(d')^2}{8} + 1\right)^{WT} \qquad \lim_{m\to\infty}\left(1+\frac{x}{m}\right)^m = e^x$$

$$\underset{WT\to\infty}{=} \log_2\left(e^{(d')^2/8}\right)$$

$$= \frac{(d')^2}{8}\log_2 e$$

$$\text{or } d' = \sqrt{8 \ln 2}\,\sqrt{U_c}$$

where d' is measured between the most discriminable members of the ensemble which contains both signals and has constant difference energy from some signal not contained in the ensemble.

If d' is a useful parameter, so is the channel capacity of a channel having the same distortion but having as input all the equal energy signals. Furthermore, U_c may in principle be determined for a channel with any statistically specified distortion, while d' is only truly defined if unit normal criterial distributions can be defined.

MMT Batch 2
10 Feb 1966

How good a representation of detectability is d'?

If it is known that the criterial distributions are normal with unit variance, the separation of their means is sufficient to describe the distributions and hence the ROC completely. In such a case, d' is sufficient to present all the available information about how the observations can effect discrimination between the two hypotheses.

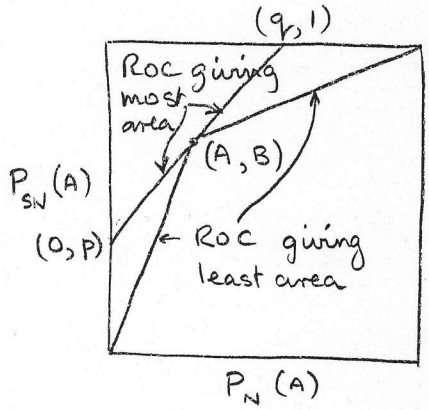
If a second observation of the same kind as the first is made in the unit-normal case, then the d' of the combined observation is given by $\sqrt{2d'^2}$. In other words, the square of d' is additive over observations. d' has a range from zero to infinity, as a good discrimination measure should.

All in all, d' is a good measure of discriminability when the unit-normal condition is fulfilled. If the assumptions are not met, d' is not defined, but in an experiment we may not know whether or not the assumptions hold, and may not want to determine a complete ROC to assess the discriminative value of the observations. Very often only one point in ROC space is determined, and d' calculated as if the unit-normal assumption were satisfied.

Limits of discriminability from a single point in ROC space (Suggested by Don Norman)

If the probability of correctly identifying H is found for a single criterion, and for the same criterion the probability of saying H when J is true is found, the two probabilities determine a point in ROC space. The intrinsic discriminability is given by a monotonic transform of the area under the ROC, and the ROC

is constrained to pass through the measured point and not to increase slope. For each measured point, the possible ROCs determine both a minimum and a maximum area, while the true ROC would determine some area in between.



If the measured point gives $P_{SN}(A)$ (probability of Accepting SN if it is true)

$$P_{SN}(A) = A$$

and $P_N(A)$ (Probability of Accepting SN if N is true)

$$P_N(A) = B$$

then the minimum area is given by an ROC which is a straight line from (0,0) to (A,B) and thence to (1,1). The ROC giving most area is a straight line from (0,p) to (q,1) passing through (A,B) where p is appropriately chosen.

The maximum area is given (if my calculations are correct) by

$$1 - \frac{1}{2}\,\frac{A(2B-1)^2}{1-B}$$

while the minimum is

$$\frac{1+B-A}{2}.$$

If we assume that the unit-normal assumptions hold, (A,B) defines a value of d', which gives an ROC with some determinable area. As a numerical example, consider $P_{SN}(A) = .75$ $P_N(A) = .25$, which define d' = 1.349 and a Gaussian ROC with an area of .83. The maximum area is .875, and the minimum .75. Hence, if we know nothing of the observation conditions, the Gaussian estimate could be no more than .08 high or .045 low in predicting 2AFC probability of a correct choice.