Minimum Probability of Error Decision Rule*

Clayton Scott
Robert Nowak

Consider the binary hypothesis test

$$\mathcal{H}_0 : x \sim f_0(x)$$

$$\mathcal{H}_1 : x \sim f_1(x)$$

Let π_i, denote the *a priori* probability of hypothesis $\mathcal{H}_i$. Suppose our decision rule declares "$\mathcal{H}_0$ is the true model" when $x \in R_0$, and it selects $\mathcal{H}_1$ when $x \in R_1$, where $R_1 = {R_0}'$. The probability of making an error, denoted P_e, is

$$\begin{aligned} P_e &= Pr\,[\text{declare}\,\mathcal{H}_0\,\text{and}\,\mathcal{H}_1\,\text{true}] + Pr\,[\text{declare}\,\mathcal{H}_1\,\text{and}\,\mathcal{H}_0\,\text{true}] \\ &= Pr\,[\mathcal{H}_1]\,Pr\,[\mathcal{H}_0 \mid \mathcal{H}_1] + Pr\,[\mathcal{H}_0]\,Pr\,[\mathcal{H}_1 \mid \mathcal{H}_0] \\ &= \int \pi_1 f_1(x)\,dx + \int \pi_0 f_0(x)\,dx \end{aligned} \quad (1)$$

In this module, we study the minimum probability of error decision rule, which selects R_0 and R_1 so as to minimize the above expression.

Since an observation x falls into one and only one of the decision regions R_i, in order to minimize P_e, we assign x to the region for which the corresponding integrand in (1) is smaller. Thus, we select $x \in R_0$ if $\pi_1 f_1(x) < \pi_0 f_0(x)$, and $x \in R_1$ if the inequality is reversed. This decision rule may be summarized concisely as

$$\Lambda(x) \equiv \frac{f_1(x)}{f_0(x)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} \equiv \eta$$

Here, $\Lambda(x)$ is called the **likelihood ratio**, η is called a **threshold**, and the overall decision rule is called the Likelihood Ratio Test[1] .

Example 1

1 Normal with Common Variance, Uncommon Means

Consider the binary hypothesis test of a scalar x

$$\mathcal{H}_0 : x \sim \mathcal{N}\left(0, \sigma^2\right)$$

*Version 1.11: May 25, 2004 4:08 pm -0500

†http://creativecommons.org/licenses/by/1.0

[1] http://workshop.molecularevolution.org/resources/lrt.php

$$\mathcal{H}_1 : x \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

where μ and σ^2 are known, positive quantities. Suppose we observe a single measurement x. The likelihood ratio is

$$\begin{aligned} \Lambda\left(x\right) &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}} \\ &= e^{\frac{1}{\sigma^2}\left(\mu x-\frac{\mu^2}{2}\right)} \end{aligned} \quad (2)$$

and so the minimum probability of error decision rule is

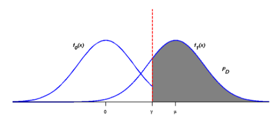
$$e^{\frac{1}{\sigma^2}\left(\mu x-\frac{\mu^2}{2}\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} = \eta$$

The expression for $\Lambda\left(x\right)$ is somewhat complicated. By applying a sequence of monotonically increasing functions to both sides, we can obtain a simplified expression for the optimal decision rule without changing the rule. In this example, we apply the natural logarithm and rearrange terms to arrive at

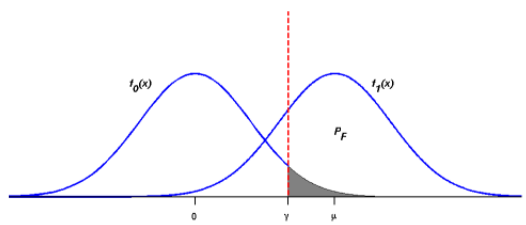
$$x \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{\mu}\ln\left(\eta\right) + \frac{\mu}{2} \equiv \gamma$$

Here we have used the assumption $\mu > 0$. If $\mu < 0$, then dividing by μ would reverse the inequalities.

This form of the decision rule is much simpler: we just compare the observed value x to a threshold γ. Figure 1 depicts the two candidate densities and a possible value of γ. If each hypothesis is *a priori* equally likely ($\pi_0 = \pi_1 = \frac{1}{2}$), then $\gamma = \frac{\mu}{2}$. Figure 1 illustrates the case where $\pi_0 > \pi_1$ ($\gamma > \frac{\mu}{2}$).



(a)



(b)

Figure 1: The two candidate densities, and a threshold corresponding to $\pi_0 > \pi_1$

If we plot the two densities so that each is weighted by its *a priori* probability of occuring, the two curves will intersect at the threshold γ (see Figure 2). (Can you explain why this is? Think back to our derivation of the LRT). This plot also offers a way to visualize the probability of error. Recall

$$\begin{aligned} P_e &= \int \pi_1 f_1(x)\,dx + \int \pi_0 f_0(x)\,dx \\ &= \int \pi_1 f_1(x)\,dx + \int \pi_0 f_0(x)\,dx \\ &= \pi_1 P_M + \pi_0 P_F \end{aligned} \tag{3}$$

where P_M and P_F denote the miss and false alarm probabilities, respectively. These quantities are depicted in Figure 2.

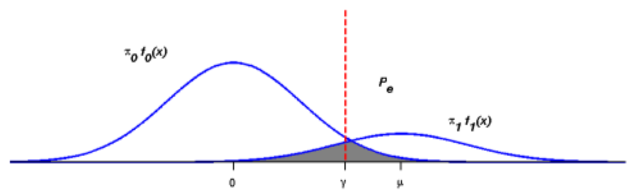


Figure 2: The candidate densities weighted by their *a priori* probabilities. The shaded region is the probability of error for the optimal decision rule.

We can express P_M and P_F in terms of the Q-function as

$$P_e = \pi_1 Q\left(\frac{\mu - \gamma}{\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sigma}\right)$$

When $\pi_0 = \pi_1 = \frac{1}{2}$, we have $\gamma = \frac{\mu}{2}$, and the error probability is

$$P_e = Q\left(\frac{\mu}{2\sigma}\right)$$

Since $Q(x)$ is monotonically decreasing, this says that the "difficulty" of the detection problem decreases with decreasing σ and increasing μ.

In the preceding example, computation of the probability of error involved a one-dimensional integral. If we had multiple observations, or vector-valued data, generalizing this procedure would involve multi-dimensional integrals over potentially complicated decision regions. Fortunately, in many cases, we can avoid this problem through the use of sufficient statistics.

Example 2

Suppose we have the same test as in the previous example (Example 1), but now we have N independent observations:

$$\mathcal{H}_0 : x_n \sim \mathcal{N}\left(0, \sigma^2\right), n = 1, \dots, N$$

$$\mathcal{H}_1 : x_n \sim \mathcal{N}\left(\mu, \sigma^2\right), n = 1, \dots, N$$

where $\mu > 0$ and $\sigma^2 > 0$ and both are known. The likelihood ratio is

$$\begin{aligned} \Lambda(x) &= \frac{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}}{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_n^2}{2\sigma^2}}} \\ &= \frac{e^{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} (x_n-\mu)^2}}{e^{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} x_n^2}} \\ &= e^{\frac{1}{2\sigma^2} \sum_{n=1}^{N} 2x_n\mu - \mu^2} \\ &= e^{\frac{1}{\sigma^2} \left(\mu \sum_{n=1}^{N} x_n - \frac{N\mu^2}{2}\right)} \end{aligned} \tag{4}$$

As in the previous example (Example 1), we may apply the natural logarithm and rearrange terms to obtain an equivalent form of the LRT:

$$t \equiv \sum_{n=1}^{N} x_n \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{\mu} \ln(\eta) + \frac{N\mu}{2} \equiv \gamma$$

The scalar quantity t is a sufficient statistic for the mean. In order to evaluate the probability of error without resorting to a multi-dimensional integral, we can express P_e in terms of t as

$$P_e = \pi_1 Pr\left[t < \gamma \mid \mathcal{H}_1 \text{ true}\right] + \pi_0 Pr\left[t > \gamma \mid \mathcal{H}_0 \text{ true}\right]$$

Now t is a linear combination of normal variates, so it is itself normal. In particular, we have $t = Ax$, where $\begin{pmatrix} 1 & \dots & 1 \end{pmatrix}$ is an N-dimensional row vector of 1's, and x is multivariate normal with mean 0 or $\mu = (\mu, \dots, \mu)^T$, and covariance $\sigma^2 I$. Thus we have

$$t|\mathcal{H}_0 \sim \mathcal{N}\left(A0, A\sigma^2 I A^T\right) = \mathcal{N}\left(0, N\sigma^2\right)$$

$$t|\mathcal{H}_1 \sim \mathcal{N}\left(A\mu, A\sigma^2 I A^T\right) = \mathcal{N}\left(N\mu, N\sigma^2\right)$$

Therefore, we may write P_e in terms of the Q-function as

$$P_e = \pi_1 Q\left(\frac{N\mu - \gamma}{\sqrt{N}\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sqrt{N}\sigma}\right)$$

In the special case $\pi_0 = \pi_1 = \frac{1}{2}$,

$$P_e = Q\left(\frac{\sqrt{N}\mu}{\sigma}\right)$$

Since Q is monotonically decreasing, this result provides mathematical support for something that is intuitively obvious: The performance of our decision rule improves with increasing N and μ, and decreasing σ.

NOTE: In the context of signal processing, the foregoing problem may be viewed as the problem of detecting a constant (DC) signal in additive white Gaussian noise:

$$\mathcal{H}_0 : x_n = w_n, n = 1, \dots, N$$

$$\mathcal{H}_1 : x_n = A + w_n, n = 1, \dots, N$$

where A is a known, fixed amplitude, and $w_n \sim \mathcal{N}\left(0, \sigma^2\right)$. Here A corresponds to the mean μ in the example.

The next example explores the minimum probability of error decision rule in a **discrete** setting.

Example 3

1 Repetition Code

Suppose we have a friend who is trying to transmit a bit (0 or 1) to us over a noisy channel. The channel causes an error in the transmission (that is, the bit is flipped) with probability p, where $0 \leq p < \frac{1}{2}$, and p is known. In order to increase the chance of a successful transmission, our friend sends the same bit N times. Assume the N transmissions are statistically independent. Under these assumptions, the bits you receive are Bernoulli random variables: $x_n \sim \text{Bernoulli}(\theta)$. We are faced with the following hypothesis test:

$\mathcal{H}_0$	$\theta = p$	0 sent
$\mathcal{H}_1$	$\theta = 1 - p$	1 sent

Table 1

We decide to decode the received sequence $x = (x_1, \dots, x_N)^T$ by minimizing the probability of error. The likelihood ratio is

$$\begin{aligned} \Lambda(x) &= \frac{\prod_{n=1}^{N} (1-p)^{x_n} p^{1-x_n}}{\prod_{n=1}^{N} p^{x_n} (1-p)^{1-x_n}} \\ &= \frac{(1-p)^k p^{N-k}}{p^k (1-p)^{N-k}} \\ &= \left(\frac{1-p}{p}\right)^{2k-N} \end{aligned} \quad (5)$$

where $k = \sum_{n=1}^{N} x_n$ is the number of 1s received.

NOTE: k is a sufficient statistic for θ.

The LRT is

$$\left(\frac{1-p}{p}\right)^{2k-N} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} = \eta$$

Taking the natural logarithm of both sides and rearranging, we have

$$k \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{N}{2} + \frac{1}{2} \frac{\ln(\eta)}{\ln\left(\frac{1-p}{p}\right)} = \gamma$$

In the case that both hypotheses are equally likely, the minimum probability of error decision is the "majority-vote" rule: Declare $\mathcal{H}_1$ if there are more 1s than 0s, declare $\mathcal{H}_0$ otherwise. In the event $k = \gamma$, we may decide arbitrarily; the probability of error is the same either way. Let's adopt the convention that $\mathcal{H}_0$ is declared in this case.

To compute the probability of error of the optimal rule, write

$$\begin{aligned} P_e &= \pi_0 Pr\left[\text{declare}\, \mathcal{H}_1 \mid \mathcal{H}_0 \,\text{true}\right] + \pi_1 Pr\left[\text{declare}\, \mathcal{H}_0 \mid \mathcal{H}_1 \,\text{true}\right] \\ &= \pi_0 Pr\left[k > \gamma \mid \mathcal{H}_0 \,\text{true}\right] + \pi_1 Pr\left[k \leq \gamma \mid \mathcal{H}_1 \,\text{true}\right] \end{aligned} \quad (6)$$

Now k is a binomial random variable, $k \sim \text{Binomial}(N, \theta)$, where θ depends on which hypothesis is true. We have

$$\begin{aligned} Pr\left[k > \gamma \mid \mathcal{H}_0\right] &= \sum_{k=\lfloor\gamma\rfloor+1}^{N} f_0(k) \\ &= \sum_{k=\lfloor\gamma\rfloor+1}^{N} \begin{pmatrix} N \\ k \end{pmatrix} p^k (1-p)^{N-k} \end{aligned} \quad (7)$$

and

$$Pr\left[k \leq \gamma \ \mid \mathcal{H}_1\right] = \sum_{k=0}^{\lfloor\gamma\rfloor} \left(\begin{array}{c} N \\ k \end{array}\right) (1-p)^k p^{N-k}$$

Using these formulae, we may compute P_e explicitly for given values of N, p, π_0 and π_1.

1 MAP Interpretation

The likelihood ratio test is one way of expressing the minimum probability of error decision rule. Another way is

Rule 1:
Declare hypothesis i such that $\pi_i f_i(x)$ is maximal.

This rule is referred to as the **maximum *a posteriori***, or **MAP** rule, because the quantity $\pi_i f_i(x)$ is proportional to the posterior probability of hypothesis i. This becomes clear when we write $\pi_i = Pr\left[\mathcal{H}_i\right]$ and $f_i(x) = f(x|\mathcal{H}_i)$. Then, by Bayes rule, the posterior probability of $\mathcal{H}_i$ given the data is

$$Pr\left[\mathcal{H}_i \ \mid x\right] = \frac{Pr\left[\mathcal{H}_i\right] f(x|\mathcal{H}_i)}{f(x)}$$

Here $f(x)$ is the unconditional density or mass function for x, which is effectively a constant when trying to maximiaze with respect to i.

According to the MAP interpretation, the optimal decision boundary is the locus of points where the weighted densities (in the continuous case) $\pi_i f_i(x)$ intersect one another. This idea is illustrated in Example 2.

2 Multiple Hypotheses

One advantage the MAP formulation of the minimum probability of error decision rule has over the LRT is that it generalizes easily to M-ary hypothesis testing. If we are to choose between hypotheses $\mathcal{H}_i$, $i = \{1, \dots, M\}$, the optimal rule is still the MAP rule (Rule 1, p. 7)

3 Special Case of Bayes Risk

The Bayes risk criterion for constructing decision rules assigns a cost C_{ij} to the outcome of declaring $\mathcal{H}_i$ when $\mathcal{H}_j$ is in effect. The probability of error is simply a special case of the Bayes risk corresponding to $C_{00} = C_{11} = 0$ and $C_{01} = C_{10} = 1$. Therefore, the form of the minimum probability of error decision rule is a specialization of the minimum Bayes risk decision rule: both are likelihood ratio tests. The different costs in the Bayes risk formulation simply shift the threshold to favor one hypothesis over the other.

4 Problems

Exercise 1
Generally speaking, when is the probability of error **zero** for the optimal rule? Phrase your answer in terms of the distributions underlying each hypothesis. Does the LRT agree with your answer in this case?

Exercise 2
Suppose we measure N independent values $x_1, \dots, x_N$. We know the variance of our measurements ($\sigma^2 = 1$), but are unsure whether the data obeys a Laplacian or Gaussian probability law:

$$\mathcal{H}_0 : f_0(x) = \frac{1}{\sqrt{2}} e^{-\left(\sqrt{2}|r|\right)}$$

$$\mathcal{H}_1 : f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{r^2}{2}}$$

4.1

Show that the two densities have the same mean and variance, and plot the densities on the same graph.

4.2

Find the likelihood ratio.

4.3

Determine the decision regions for different values of the threshold η. Consider all possible values of $\eta > 0$

NOTE: There are three distinct cases.

4.4

Draw the decision regions and decision boundaries for $\eta = \left\{\frac{1}{2}, 1, 2\right\}$.

4.5

Assuming the two hypotheses are equally likely, compute the probability of error. Your answer should be a number.

Exercise 3

4.1 Arbitrary Means and Covariances

Consider the hypothesis testing problem

$$\mathcal{H}_0 : x \sim \mathcal{N}(\mu_0, \Sigma_0)$$

$$\mathcal{H}_1 : x \sim \mathcal{N}(\mu_1, \Sigma_1)$$

where $\mu_0 \in \mathbb{R}^d$ and $\mu_0 \in \mathbb{R}^d$, and Σ_0, Σ_1 are positive definite, symmetric $d \times d$ matrices. Write down the likelihood ratio test, and simplify, for the following cases. In each case, provide a geometric description of the decision boundary.

4.1.1

$\Sigma_0 = \Sigma_1$, but $\mu_0 \neq \mu_1$.

4.1.2

$\mu_0 = \mu_1$, but $\Sigma_0 \neq \Sigma_1$.

4.1.3

$\mu_0 \neq \mu_1$ and $\Sigma_0 \neq \Sigma_1$.

Exercise 4

Suppose we observe N independent realizations of a Poisson random variable k with intensity parameter λ:

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We must decide which of two intensities is in effect:

$$\mathcal{H}_0 : \lambda = \lambda_0$$

$$\mathcal{H}_1 : \lambda = \lambda_1$$

where $\lambda_0 < \lambda_1$.

4.1

Give the minimum probability of error decision rule.

4.2

Simplify the LRT to a test statistic involving only a sufficient statistic. Apply a monotonically increasing transformation to simplify further.

4.3

Determine the distribution of the sufficient statistic under both hypotheses.

NOTE: Use the characteristic function to show that a sum of IID Poisson variates is again Poisson distributed.

4.4

Derive an expression for the probability of error.

4.5

Assuming the two hypotheses are equally likely, and $\lambda_0 = 5$ and $\lambda_1 = 6$, what is the minimum number N of observations needed to attain a probability of error no greater than 0.01?

NOTE: If you have numerical trouble, try rewriting the log-factorial so as to avoid evaluating the factorial of large integers.

Exercise 5

In Example 3, suppose $\pi_0 = \pi_1 = \frac{1}{2}$, and $p = 0.1$. What is the smallest value of N needed to ensure $P_e \leq 0.01$?