Notes on Decision Making for ECE 413 Fall 2004 [1]

1 Introduction

Most people have found decision-making to be one of the most fruitful or enjoyable or empowering or aggravating or enervating of human activities. Some people make decisions rapidly and effortlessly (sometimes thoughtlessly), others ponder deeply before deciding, while many more let their "I dare not" wait upon their "I would" like the poor cat i' the adage and dither endlessly. The concern that gives one pause is that the data on which the decision is to be based usually have uncertainties associated with them. A decision could possibly be in error because the data are "faulty" and such errors may have tremendous economic or social consequences as well. The results and effects of one's decisions must be taken into account in the decision-making, this often resulting in decisions that "come from the heart rather than the head." While such decisions may be appropriate in more romantic settings, a more prosaic and rational approach to decision-making is all that one can hope for in a scientific study. Thus, in this set of notes, we consider rational methodologies for making decisions based on uncertain data. The methods that we study were originated by statisticians and probabilists but are now pervasive in engineering, science, and business applications.

Decision-making entered the realms of electrical and computer engineering during World War II and the years immediately prior thereto in the form of a problem in radar system design. A simple model for a radar system consists of a transmitter that emits electromagnetic pulses in the direction of a "target" that may (or may not!) be out there. The radar receiver listens for echoes that are reflected off the hypothetical target. The length of time between the transmission of a pulse and its reception is used to estimate the range of the target. Other signal and system parameters such as Doppler shift in frequency and antenna orientation can be used to estimate the target velocity, compass orientation, etc., but such bells and whistles will not be part of our study, and we will consider only the fundamental decision problem: "Is a target present?" Note that the word target comes from the original military application, and we shall continue to use it with the caveat that the target is not necessarily something that we desperately want to shoot down: it is merely a convenient name for the object whose presence or absence we are attempting to determine.

Although the question "Is a target present?" has the seemingly blindingly obvious answer "A target is present if and only if an echo is heard", it really is not as simple as that. The echoes, if present at all, are generally very weak signals (remember inverse square laws in electromagnetic propagation?) that are difficult to distinguish from the ambient noise. On the other hand, when a target is not present at all, the receiver may occasionally mistake the noise for echoes reflected off a target. As a specific example, suppose that n radar pulses are transmitted. When a target is present, the n reflected pulses may be masked by the noise and it is possible that not all n are detected. Conversely, when the target is not present, the receiver may at times mistake the noise for echoes. Thus, suppose that the radar receiver has concluded that it has detected k echoes, where k ranges from 0 to n. Based on this data, the radar system must make a decision. Either it decides that a target is present, or it decides that a target is absent. Now, it is eminently reasonable that if the receiver has heard n echoes, then it should decide that a target is present, while if it has heard no echoes at all, then it should decide that there is no target. But what if the receiver has heard k echoes, $0 < k < n$? Well, if k is small, the receiver should decide that the target is absent, while if k is large, it should decide that the target is present. But where does one draw the dividing line between large and small? Should the dividing line be $n/2$? And can one be sure that $n/2$ is the "best possible" choice? And what do we mean by best choice anyway? In figuring out where to draw this dividing line, we will consider the case of a paranoid "hawk" who wants the reception of even one echo to be interpreted as the target being present, and also the case of an ostrich-like "dove" who refuses to regard anything less than the complete unanimity of n echoes as indicative of the presence of the target. Should such prejudices be entertained in the system design? And if so, how can they be included?

[1] These notes are an abridged version of course notes by Dilip V. Sarwate, *An Introduction to Decision-Making under Uncertainty*, 2002.

We shall also consider the awkward fact that, regardless of the choice of the dividing line and how it was arrived at, the receiver's decision will sometimes be incorrect. The system designer's fondest hope is that the receiver decision will always be correct, but it is more realistic to accept the fact that occasionally, a radar receiver does make mistakes. The system may occasionally decide that a target is present when in fact there is no target present at all: an event that is picturesquely named a False Alarm by the radar community but, more mundanely, is deemed a Type I Error by statisticians. On the other hand, a radar system may occasionally decide that a target is not present when in fact a target is present, a pooh-poohing that is called a Missed Detection in the radar community (False Dismissal by those who see beauty in symmetry of nomenclature), and is called a Type II Error by statisticians. We will study the probabilities of these events and consider how to minimize the probabilities by careful choice of decision rules. We will consider how to take into account our beliefs about whether we expect to see a target or not by quantizing them into a priori probabilities of the two hypotheses. Finally, the two different types of errors may have different consequences that might be measurable in terms of costs. For a given set of costs, we shall study how to make decisions so as to minimize the average cost of the erroneous decisions.

2 Binary Hypothesis Testing

The simplest paradigm in hypothesis testing involves two possible hypotheses about the state of nature. Exactly one of these hypotheses is true, and we wish to decide which hypothesis is true. To help us in making the decision, we observe the value taken on by a random variable X whose distribution depends on the state of nature, that is, which hypothesis is true. For example, in the radar problem described above, the two hypotheses are H_0: the target is absent and H_1: the target is present.

The Likelihood Matrix The likelihood matrix is an array with two rows, one for each hypothesis, and as many columns as the possible values of the observation X. The entries in each row are the probabilities of the various possible observations conditioned on which hypothesis is true. As an example, consider the radar problem with $n = 3$. The likelihood matrix might be the following:

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.0	0.1	0.3	0.6
H_0	0.4	0.3	0.2	0.1

In practice, the numbers in the table might be provided by systems engineers that test the system with and without targets present. Note that since all four possible observation values are taken into account, the row sums of the matrix are both one.

Now, recall that we observe the value of X and then decide whether H_0 or H_1 is true. The decision rule can be conveniently displayed on the likelihood matrix by underlining the appropriate entry in the matrix. Note that one entry in each column is underlined, specifying that when X has the value shown at the top of the column, the decision is that the hypothesis corresponding to the row containing the underlined entry is true. For example, the hawk rule is shown below, where the decision is that H_1 is true (the target is present) whenever at least one echo has been detected. Thus, when it is observed that $X = 2$, the decision is that H_1 is true, because the underlined entry is in the H_1 row of the likelihood matrix, and similarly for the other entries.

	$X=0$	$X=1$	$X=2$	$X=3$		
H_1	0.0	$\underline{0.1}$	$\underline{0.3}$	$\underline{0.6}$	$\leftarrow$	underlines indicate the "hawk" decision rule
H_0	$\underline{0.4}$	0.3	0.2	0.1		(used just for this example)

Error probabilities We are observing the value of a random variable X and then deciding whether H_0 or H_1 is the true hypothesis. If H_0 is, in fact, the true state of nature and we decide that H_1 is the true hypothesis, then we have committed the grievous error called a false alarm. For the hawk decision rule described above, we almost always decide that H_1 is the true hypothesis, and thus, have a false alarm whenever H_0 is, in fact, the true state of nature but we have observed the value of X to be 1, or 2 , or 3. By convention, we denote by p_{false_alarm} the conditional probability:

$$p_{false\_alarm} = P[decide\ H_1\ true\ |H_0]$$

Note that p_{false_alarm} is the sum of the entries on the H_0 row of the likelihood matrix that are not underlined. For the decision rule given above $p_{false\_alarm} = P[decide\ H_1\ true\ |H_0] = 0.3 + 0.2 + 0.1 = 0.6$. Note that p_{false_alarm} is rather large. This is a direct result of our hawk rule: we wish to decide that a target is present on the least provocation, and thus, on more than half the occasions when a target is in fact not present, our eagerness causes us to mistakenly declare that a target is present.

A similar analysis can be carried out for the case when H_1 is, in fact, the true state of nature and we decide that H_0 is the true hypothesis. We have now committed the error called a missed detection. For the hawk decision rule described above, we almost always decide that H_1 is the true hypothesis, and thus, have a missed detection whenever H_1 is, in fact, the true state of nature but we have observed the value of X to be 0. By convention, we denote by p_{miss} the conditional probability:

$$p_{miss} = P[decide\ H_0\ true\ |H_1]$$

Note that p_{miss} is the sum of the entries of the H_1 row of the likelihood matrix that are not underlined. The "hawk" decision rule decides that H_1 is true unless $X = 0$, and $P[X = 0|H_1] = 0.0$. Therefore $p_{miss} = 0.0$, which is unusually good. Of course this small value p_{miss} is earned at the expense of the large value of p_{false_alarm} noted above.

The above analysis applies to all decision rules, not just the "hawk" decision rule. When H_i is, in fact, the true state of nature and we decide that H_{1-i} is the true hypothesis, the conditional error probability (p_{false_alarm} if $i = 0$, p_{miss} if $i = 1$) is the sum of the entries on the H_i row of the likelihood matrix that are not underlined. The entries not underlined indicate precisely those values of X for which the decision rule is not choosing H_i, and the conditional probabilities that X takes on these values given H_i can be read off right from the likelihood matrix. Our analysis also allows us to illustrate trade-offs between the two types of error probabilities. If the underlining in some column is moved from one row to the other, then one error probability increases (since there is one more entry not underlined to include in the sum) and correspondingly, the other error probability decreases (because there is one fewer entry not underlined to include in the other sum.) For example, in our radar problem if we choose to decide that there is no target present when $X = 1$, then p_{false_alarm} is reduced from 0.6 to just 0.3 while p_{miss} increases from 0.0 to 0.1, which makes the two error probabilities more equal. Whether such equality is good or bad is quite another matter.

3 Three popular decision rules

So far, we have discussed decision rules in general. For each value of X, it is possible to choose to whether to decide in favor of H_0 or in favor of H_1 quite arbitrarily. No matter what the decision rule is, the trick of adding up all the entries not underlined in the rows of the likelihood matrix gives us a handy method for calculating the conditional probabilities, p_{false_alarm} and p_{miss}. But are there reasons for favoring one decision rule over another? How do we go about choosing a decision rule that gives good performance in some sense? and can we find some reasons to explain to our boss why we chose the rule the way we did?

3.1 Maximum likelihood (ML) decision rule

The ML decision rule favors the hypothesis which gives the maximum probability (which the statisticians insist on calling likelihoods) for the event that we have observed. Operationally, the ML decision rule can be stated as follows: In each column of the likelihood matrix, choose the largest entry as the one to be underlined. If both entries in a column of the likelihood matrix are identical, then either can be chosen to be underlined. The choice may depend on other considerations such as whether we wish to minimize p_{false_alarm} or p_{miss}. The ML rule for our radar problem is the following:

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.0	0.1	<u>0.3</u>	<u>0.6</u>
H_0	<u>0.4</u>	<u>0.3</u>	0.2	0.1

← underlines indicate the ML decision rule

The reader is invited to verify that for the ML decision rule, $p_{false\_alarm} = 0.2 + 0.1 = 0.3$ and $p_{miss} = 0.0 + 0.1 = 0.1$.

There is another way to express the ML decision rule. Note that for two positive numbers a and b, the statement $a > b$ is equivalent to the statement that $\frac{a}{b} > 1$. Thus, the ML rule can be rewritten in a form called a *likelihood ratio test* (LRT) as follows. Define the likelihood ratio $\Lambda(k)$ for each possible observation k as the ratio of the two conditional probabilities:

$$\Lambda(k) = \frac{P[X = k|H_1]}{P[X = k|H_0]}$$

The ML rule is thus equivalent to deciding that H_1 is true if $\Lambda(X) > 1$ and deciding H_0 is true if $\Lambda(X) < 1$. The ML rule can be compactly written as

$$\Lambda(X) \begin{cases} > 1 & \text{decide } H_1 \text{ is true} \\ < 1 & \text{decide } H_0 \text{ is true} \end{cases}$$

We shall see that the other two decision rules can also be expressed as LRT's, but with the threshold 1 changed to different values. An LRT with threshold τ can be written as

$$\Lambda(X) \begin{cases} > \tau & \text{decide } H_1 \text{ is true} \\ < \tau & \text{decide } H_0 \text{ is true} \end{cases}$$

Note that if the threshold τ is increased, then there are fewer observations that lead to deciding H_1 is true. Thus, as τ increases, p_{false_alarm} decreases and p_{miss} increases. Why would one want to fiddle with these error probabilities instead of accepting what the ML rule gives us? Well, bear in mind that in some instances (such as medical testing for AIDS, cancer or pregnancy), false alarms may have many disastrous consequences, and one may well wish to reduce p_{false_alarm} even though it means increasing p_{miss}. Conversely, vigilant watchers not intending to be burnt alive by falling enemy bombs may wish to reduce p_{miss} even if it means living with the consequences of an increased p_{false_alarm}. Of course, one could reduce p_{miss} to 0 by choosing a threshold of 0 but then $p_{false\_alarm} = 1$, which is no fun. So, how much can we reduce p_{miss} subject to the constraint that $p_{false\_alarm} \leq a$ where a is some number we can live with? We will not give a specific answer but merely note that there is a well-known result in statistics called the Neyman-Pearson lemma to the effect that a decision rule that minimizes p_{miss} subject to the constraint that $p_{false\_alarm} \leq a$ for any given a is an LRT for some appropriately chosen value of threshold. That is, a decision rule that in some sense "minimizes" both error probabilities is an LRT.

As noted above, the ML rule is an LRT with threshold $\tau = 1$.

3.2 Maximum a posteriori probability (MAP) decision rule

The next decision rule we discuss requires the computation of joint probabilities such as $P[\{X = 1\} \cap H_1]$. For brevity we write this probability as $P[H_1, X = 1]$. Such probabilities cannot be deduced from the likelihood matrix alone. Rather, it is necessary for the system designer to assume some values for $P[H_0]$ and $P[H_1]$. Let the assumed value of $P[H_i]$ be denoted by π_i, so that $\pi_0 = P[H_0]$ and $\pi_1 = P[H = 1]$. The probabilities π_0 and π_1 are called *prior* probabilities, because they are the probabilities assumed prior to when the observation is made.

Together the conditional probabilities listed in the likelihood matrix and the prior probabilities determine the joint probabilities $P[H_i, X = k]$, because $P[H_i, X = k] = \pi_i P[X = k|H_i]$. The *joint probability matrix* is the matrix of joint probabilities $P[H_i, X = k]$. For our radar example, suppose $\pi_0 = 0.8$ and $\pi_1 = 0.2$. Then the joint probability matrix is given by

	$X = 0$	$X = 1$	$X = 2$	$X = 3$
H_1	0.00	0.02	0.06	0.12
H_0	0.32	0.24	0.16	0.08

Note that the row for H_i of the joint probability matrix is π_i times the corresponding row of the likelihood matrix. Since the row sums for the likelihood matrix are one, the sum for row H_i of the joint probability matrix is π_i. Therefore the sum of all entries in the joint probability matrix is one. The joint probability matrix can be viewed as a Venn diagram.

Conditional probabilities such as $P[H_1|X = 2]$ and $P[H_0|X = 2]$ are called a posteriori probabilities, because they are probabilities that an observer would assign to the two hypotheses after making the observation (in this case observing that $X = 2$). Given an observation, such as $X = 2$, the MAP rule chooses the hypothesis with the larger conditional probability. By Bayes' formula, $P[H_1|X = 2] = \frac{P[H_1,X=2]}{P[X=2]} = \frac{P[H_1,X=2]}{P[H_1,X=2]+P[H_0,X=2]} = \frac{0.06}{0.06+0.16}$ That is, $P[H_1|X = 2]$ is the top number in the column for $X = 2$ in the joint probability matrix divided by the sum of the numbers in the column for $X = 2$. Similarly the conditional probability $P[H_0|X = 2]$ is the bottom number in the column for $X = 2$ divided by the sum of the numbers in the column for $X = 2$. Since the denominators are the same (both denominators are equal to $P[X = 2]$) it follows that whether $P[H_1|X = 2] > P[H_1|X = 2]$ is equivalent to whether the top entry in the column for $X = 2$ is greater than the bottom entry in the column for $X = 2$.

Thus, the MAP decision rule can be specified by underlining the larger entry in each column of the joint probability matrix. For our radar example, the MAP rule is given by the following.

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.00	0.02	0.06	<u>0.12</u>
H_0	<u>0.32</u>	<u>0.24</u>	<u>0.16</u>	0.08

$\leftarrow$ underlines indicate the MAP decision rule

Thus, if the observation is $X = k$, the MAP rule decides hypothesis H_1 is true if $\pi_1 P[X = k|H_1] > \pi_0 P[X = k|H_0]$, or equivalently if $\Lambda(k) > \frac{\pi_0}{\pi_1}$, where Λ is the likelihood ratio defined above. Therefore, the MAP rule is equivalent to the LRT with threshold $\tau = \frac{\pi_0}{\pi_1}$.

Sanity check: Does it make sense that if $\pi_0 > \pi_1$, then the threshold for the MAP rule (in LRT form) is greater than one? Indeed it does, because a larger threshold value in the LRT means there are fewer observations leading to deciding H_1 is true, which is appropriate behavior if $\pi_0 > \pi_1$.

The MAP rule has a remarkable optimality property, as we now explain. The unconditional error probability, which we call p_e, for any decision rule can be written as $p_e = \pi_0 p_{false\_alarm} + \pi_1 p_{miss}$. Suppose a decision rule is specified by underlining one number from each column of the joint probability matrix. Then p_e is the sum of all numbers in the joint probability matrix that are not underlined. From this observation it easily follows that, among all decision rules, the MAP decision rule is the one that minimizes p_e. (For that reason, in some books the MAP decision rule is called the minimum probability of error rule.)

Philosophical digression: We remark that many statisticians vehemently disagree with the notion of assigning probabilities to hypotheses. They claim quite reasonably that the hypotheses are assertions about the state of nature. One of the two hypothesis is true, and the other isn't (we don't know which), but how does probability enter into the picture at all? Furthermore, any assignment of the prior probabilities π_0 and π_1 is a matter of pure guesswork! For example, most U.S. military radars have never detected an echo from a target, so what we can assign as the probability of the hypothesis H_1 that a target is present? But, the same statisticians have no difficulty in assigning probability 1/2 to the event that a toss of a brand-new straight-from-the-mint coin results in a Head! The counter-argument is provided by the Bayesians who claim that it is quite reasonable to assign probabilities to hypotheses even if these merely reflect our beliefs and prejudices, and the above theorem of total probability expression makes perfect sense. In fact, the ML decision rule and the LRTs are special cases of Bayesian decision rules. In particular, the ML rule is the Bayesian decision rule in the case when the two hypotheses are equally likely, and the Bayesians gleefully point out that proponents of the ML rule, while castigating Bayesian prejudices, are in fact indulging their own prejudices in unwittingly assuming that $\pi_0 = \pi_1 = 1/2$ even in cases when such assumptions are untenable. We will not delve further into these philosophical issues because that will take us too far out of our path, but you should be aware that Bayesian decision rules are not accepted universally in statistical circles.

3.3 Bayes' minimum average cost decision rule

Suppose that we are fined one dollar for each wrong decision. There is no fine, but no reward either, for a correct decision. Then our average cost per trial is p_e dollars per trial (which is only pennies a day, as the commercials assert). We can minimize our average cost by using a MAP decision rule. This scenario is said to involve uniform costs (both kinds of errors cost the same). More generally, the fines for different types of errors might be different, e.g. a false alarm costs "only" aviation fuel, overtime pay for the on-duty pilot who has to go investigate the alleged target, wear-and-tear on the fighter plane, etc. whereas a missed detection can be far more expensive and catastrophic. The Bayesian formulation can also be used profitably when different errors have different costs. Let C_{ij} denote the cost of deciding hypothesis H_i is true given that H_j is the true hypothesis. Note that C_{00} and C_{11} are the "costs" of getting the answer right(!) but we include them in our analysis so as to cover the cases when there is a reward for getting the right answer (C_{00} and C_{11} are negative), and the case when there is a fixed cost of doing the experiment, regardless of the decision. We shall insist, however, that $C_{ji} > C_{ii}$, that is, making an error costs more than getting the right answer. As we shall see below, the tests depend only on the cost differences between wrong and right answers, $C_{10} - C_{00}$ and $C_{01} - C_{11}$. Summing $C_{ij}P[\text{decide } H_i, H_j]$ over the four possible values of (i, j) yields the average cost (also called the risk) of any decision rule. The Bayes' minimum average cost rule minimizes the average cost. Such a rule, for each possible observation k, should choose the hypothesis i to minimize the average cost given the observation k. That is, for observation k, the Bayes' decision rule should select H_i such that i minimizes

$$C_{i0}P[H_0|Z = k] + C_{i1}P[H_1|Z = k]$$

By Bayes' formula, this ratio is equal to

$$\frac{C_{i0}\pi_0 P[Z = k|H_0] + C_{i1}\pi_1 P[Z = k|H_1]}{P[Z = k]}.$$

Since the denominator $P[Z = k]$ doesn't depend on the decision i, it can be ignored, so that the Bayes' minimum average cost rule decides H_1 is true if

$$C_{10}\pi_0 P[Z = k|H_0] + C_{11}\pi_1 P[Z = k|H_1] < C_{00}\pi_0 P[Z = k|H_0] + C_{01}\pi_1 P[Z = k|H_1].$$

Rearrangement of this condition shows that the Bayes' minimum average cost rule is the LRT with threshold τ given by

$$\tau = \frac{(C_{10} - C_{00})\pi_0}{(C_{01} - C_{11})\pi_1} \qquad \text{(Bayes' minimum average cost threshold).}$$

As a sanity check, suppose that the net cost of a mistake given H_0 is true, $C_{10} - C_{00}$, is much larger than the net cost of making a mistake given H_1 is true, $C_{01} - C_{11}$. Then taking the costs into account leads to a larger value of the threshold in the LRT test. Why does that make sense?

We remark that Bayesian decision rules for minimizing the average cost are excoriated even more by some statisticians. Not only do they involve a priori probabilities which can be difficult to estimate, and about which rational individuals may reasonably differ, but estimates of the costs of errors can be even more subjective. As trial lawyers are fond of reminding us, "Only you can put a dollar value on the pain and suffering of my client." Thus, be aware that not everyone will agree with you if you use Bayesian decision procedures. But, as mentioned earlier, everyone is secretly a Bayesian! In the spirit of "All men are created equal" they implicitly endow their hypotheses with equal probability and merrily use ML rules without realizing the consequences.

4 Summary

Three popular decision rules for binary hypothesis testing have been discussed:

Maximum likelihood (ML) rule: Decisions correspond to the largest element in each column of the likelihood matrix. No prior probabilities are needed or used. Can implement as an LRT with threshold τ=1.

Maximum a posteriori probability (MAP) rule: Decisions correspond to the largest element in each column of the joint probability matrix. Prior probabilities are needed to compute the matrix. Can implement as an LRT with threshold $\tau = \frac{\pi_0}{\pi_1}$. The MAP rule minimizes the unconditional probability of error, p_e.

Bayes' minimum average cost rule: This rule minimizes the average cost. Both prior probabilities and costs must be specified. Can implement as an LRT with threshold $\tau = \frac{(C_{10}-C_{00})\pi_0}{(C_{01}-C_{11})\pi_1}$

The following relationships among the rules exist. The ML rule is the same as the MAP rule for uniform priors ($\pi_0 = \pi_1$). The Bayes' minimum average cost rule is the same as the MAP rule for uniform costs ($C_{10} - C_{00} = C_{01} - C_{11}$). Any of the three rules can be expressed as an LRT; only the threshold in the LRT depends on what rule is used.

The performance of any decision rule can be summarized by the pair of conditional error probabilities, p_{false_alarm} and p_{miss}. The decision rule determines which numbers in the likelihood matrix to add together to yield p_{false_alarm} or p_{miss}. For example, p_{false_alarm} for the MAP rule is the sum of the entries in the H_0 row of the likelihood matrix that correspond to entries that are not underlined in the H_0 row of the joint probability matrix.