Ideal Observers and Optimal ROC Hypersurfaces in *N*-Class Classification

Darrin C. Edwards,
D. C. Edwards is with the Department of Radiology, the University of Chicago, Chicago, IL 60637 USA

Charles E. Metz, and
C. E. Metz is with the Department of Radiology, the University of Chicago, Chicago, IL 60637 USA

Matthew A. Kupinski
M. A. Kupinski is with the Optical Sciences Center, University of Arizona, Tucson, AZ 85721 USA

Abstract

The likelihood ratio, or ideal observer, decision rule is known to be optimal for two-class classification tasks in the sense that it maximizes expected utility (or, equivalently, minimizes the Bayes risk). Furthermore, using this decision rule yields a receiver operating characteristic (ROC) curve which is never above the ROC curve produced using any other decision rule, provided the observer's misclassification rate with respect to one of the two classes is chosen as the dependent variable for the curve (i.e., an "inversion" of the more common formulation in which the observer's true-positive fraction is plotted against its false-positive fraction). It is also known that for a decision task requiring classification of observations into *N* classes, optimal performance in the expected utility sense is obtained using a set of $N-1$ likelihood ratios as decision variables. In the *N*-class extension of ROC analysis, the ideal observer performance is describable in terms of an $(N^2 - N - 1)$-parameter hypersurface in an $(N^2 - N)$-dimensional probability space. We show that the result for two classes holds in this case as well, namely that the ROC hypersurface obtained using the ideal observer decision rule is never above the ROC hypersurface obtained using any other decision rule (where in our formulation performance is given exclusively with respect to between-class error rates rather than within-class sensitivities).

Index Terms

ROC analysis; ideal observer; *N*-class classification

I. Introduction

IN a two-class classification task, observations $\vec{\mathbf{x}}$ are randomly drawn from a distribution of "signals" (or "positive" or "abnormal" observations) and a distribution of "noise" (or "negative" or "normal" observations). The probability density functions (pdfs) are $p_{\vec{x}}(\vec{x}|\,\pi_s)$ for the signal observations and $p_{\vec{x}}(\vec{x}|\pi_n)$ for the noise observations. (We use boldface type to denote statistically variable quantities, and arrows above quantities to denote them as vectors.) It is well known that using the likelihood ratio, defined as

Correspondence to: Darrin C. Edwards.
e-mail: d-edwards@uchicago.edu.

$$\mathbf{LR} \equiv \frac{p_{\vec{x}}(\vec{\mathbf{x}}|\pi_s)}{p_{\vec{x}}(\vec{\mathbf{x}}|\pi_n)} \qquad (1)$$

as the decision variable yields the optimal classification performance achievable in an ideal observer sense [1], [2]. In particular, the ideal observer is obtained when one requires that expected utility be maximized (or equivalently, that Bayes risk be minimized). Here $\mathbf{LR}$ is to be interpreted as a function of the underlying random variable $\vec{\mathbf{x}}$ describing the data. We denote the pdf of $\mathbf{LR}$ as $p_{\mathrm{LR}}(\mathrm{LR})$, distinct from $p_{\vec{x}}(\vec{x})$ above. The receiver operating characteristic (ROC) curve for this decision variable is defined as the locus of all sensitivity and specificity pairs, or equivalently as the parametric curve (FPF(LR_c), TPF(LR_c)), where

$$\mathrm{TPF}(\mathrm{LR}_c) \equiv \int_{\mathrm{LR}_c}^{\infty} p_{\mathrm{LR}}(\mathrm{LR}|\pi_s)d\mathrm{LR} \qquad (2)$$

is the true-positive fraction (TPF), or sensitivity; and where

$$\mathrm{FPF}(\mathrm{LR}_c) \equiv \int_{\mathrm{LR}_c}^{\infty} p_{\mathrm{LR}}(\mathrm{LR}|\pi_n)d\mathrm{LR} \qquad (3)$$

is the false-positive fraction (FPF), or one minus the specificity. Here, LR_c is a threshold applied to the decision variable $\mathbf{LR}$.

It is clear from this standard formulation that any monotonic transformation of the decision variable $\mathbf{LR}$ will yield the same ROC curve [2]. The decision variable $\mathbf{LR}$, or any monotonic transformation of this decision variable, will produce an ROC curve which is optimal in the sense that at any given FPF, no ROC curve produced using a different decision variable can have a TPF higher than that of the ideal observer. This optimality can be demonstrated in two ways: by showing that the observer which satisfies the Neyman–Pearson criterion (maximizing TPF at any given FPF) is in fact the ideal observer [1]; or by showing that maximizing expected utility results in an ROC curve with this property [3].

Note that we can just as well specify the observer's performance using the parametric pair (FPF(LR_c), FNF(LR_c)) where

$$\mathrm{FNF}(\mathrm{LR}_c) \equiv \int_{-\infty}^{\mathrm{LR}_c} p_{\mathrm{LR}}(\mathrm{LR}|\pi_s)d\mathrm{LR} \qquad (4)$$

is the false-negative fraction (FNF), or one minus the sensitivity. In this equivalent formulation, the ideal observer decision variable will produce an ROC curve which is optimal in the sense that at any given FPF, no ROC curve produced using a different decision variable (i.e., one which is not a monotonic transformation of $\mathbf{LR}$) can have a FNF lower than that of the ideal observer. Similarly, the Neyman-Pearson criterion is now satisfied by *minimizing* FNF at any given FPF. This is the formulation which we will adopt throughout this paper; despite being somewhat unusual for the two-class classification case, it has certain notational advantages when considering more than two classes.

The ideal observer can be extended to the case of *N* classes [1]. Since the dimensionalities of the probability spaces and of the decision variable vectors involved increase rapidly with the number of classes, we will review this extension in detail for the *N*-class classification

problem in the next section. In the remaining sections, we show that the result for two-class classification holds in the *N*-class extension of ROC analysis as well; namely, that an ROC hypersurface obtained by using the ideal observer decision rule is never above the ROC hypersurface obtained using any other decision rule. Here "above" has the simple interpretation that a function $f(\vec{x})$ is above $g(\vec{x})$ at a particular point $\vec{x}_0$ if $f(\vec{x}_0) > g(\vec{x}_0)$.

II. Maximization of Expected Utility

We seek to classify observations $\vec{\mathbf{x}}$ as coming from one of *N* classes, which we label π_i where *i* varies from 1 to *N*. (Informally, class π_N may be thought of as representing "normal" cases, while the other $N - 1$ classes represent various types of "abnormality," as in, for example, a medical diagnostic task. The actual properties of the classes are, however, irrelevant to the analysis here.) For any observation $\vec{\mathbf{x}}$, we may define the actual class (the "truth") to which it belongs as **t**, and the class to which it is assigned (the "decision") as **d**, where **t** and **d** can take on any of the values π_i, …, π_i, … π_N, the labels of the various classes. A given observation $\vec{x}$ thus arises from one of *N* conditional pdfs $p_{\vec{x}}(\vec{x}/\mathbf{t} = \pi_i)$, and is classified according to a set of *N* rules of the form

$$\text{decide } d = \pi_i \text{ iff } \vec{x} \in Z_i \tag{5}$$

where the sets Z_i partition the domain of $\vec{x}$. In effect, the sets Z_i define a classifier.

The performance of this classifier is described by N^2 conditional probabilities of the form $P(\mathbf{d} = \pi_i \mid \mathbf{t} = \pi_j)$, which represent the various probabilities of classifying an observation as belonging to class π_i given that it is actually drawn from the distribution of class π_j. From the definition of conditional probability [4]

$$\sum_{i=1}^{N} P(\mathbf{d} = \pi_i \mid \mathbf{t} = \pi_j) = 1 \tag{6}$$

meaning that only $N^2 - N$ of these N^2 probabilities are needed to completely describe the behavior of a particular classifier. (This generalizes the familiar fact that, for two classes, we do not need to consider explicitly the true negative fraction (TNF), or the false negative fraction (FNF), since TNF = 1 − FPF and FNF = 1 − TPF.)

Some researchers have suggested that in, e.g., a three-class classification task [5], [6], the set of three "sensitivities" ($P(\mathbf{d} = \pi_i \mid \mathbf{t} = \pi_i)$ in our notation) provides a complete description of observer performance. This is incorrect in general, because it ignores the $N^2 - N$ misclassification probabilities, not all of which are determined uniquely by the "sensitivities" when $N > 2$ unless particular restrictions are imposed on the observer's behavior. Complete quantification of the trade-offs available among the probabilities of various kinds of misclassification error is important in medical diagnosis, where different misclassification errors often have substantially different clinical consequences. Moreover, restrictions concerning the observer's behavior are inappropriate when considering ideal observers, human observers, or automated observers (such as automated schemes for computer-aided diagnosis) designed to approximate ideal or human observer behavior.

Hypothetically, a "perfect" classifier would have $P(\mathbf{d} = \pi_i \mid \mathbf{t} = \pi_j) = \delta_{ij}$; that is, data drawn from class π_j would always be assigned to class π_j, and never assigned to any class π_i for $i \neq j$. Since the data $\vec{\mathbf{x}}$ are random, and since the pdfs of the data will overlap in any nontrivial classification task, this is not possible even in principle. We therefore seek a classifier which

performs optimally given the random nature of the data. Ideal observer decision theory requires that a decision alternative be selected only if its expected utility is greater than the expected utility of any other possible decision. That is, for a given observation $\vec{x}$

$$\begin{array}{c} \text{decide } d=\pi_1 \text{ iff} \\ E\{U_{\pi_1}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\}>E\{U_{\pi_j}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \;\; \{2 \le j \le N\} \end{array} \tag{7}$$

$$\begin{array}{c} \text{decide } d=\pi_2 \text{ iff} \\ E\{U_{\pi_2}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \ge E\{U_{\pi_1}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \;\; \text{and} \\ E\{U_{\pi_2}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\}>E\{U_{\pi_j}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \;\; \{3 \le j \le N\} \end{array} \tag{8}$$

$$\begin{array}{c} \text{decide } d=\pi_i \text{ iff} \\ E\{U_{\pi_i}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \ge E\{U_{\pi_j}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \;\; \{j<i\} \;\; \text{and} \\ E\{U_{\pi_i}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\}>E\{U_{\pi_j}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \;\; \{j>i\} \end{array} \tag{9}$$

$$\begin{array}{c} \text{decide } d=\pi_N \text{ iff} \\ E\{U_{\pi_N}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \ge E\{U_{\pi_j}(\vec{\mathbf{x}},\mathbf{t})\lceil\vec{x}\} \;\; \{j<N\} \end{array} \tag{10}$$

where the random variable $U_{\pi_i}(\vec{\mathbf{x}}, \mathbf{t})$ is the utility of assigning an observation $\vec{\mathbf{x}}$, actually drawn from class $\mathbf{t}$, to class π_i[1]. (Since we are free to define the "utility" as the negative of the Bayes risk, it should be clear that maximizing the expected utility is equivalent to minimizing the Bayes risk.) Note that the choice of which decision to make when the expected utilities of a given pair of decisions are equal (informally, where "equality" symbols appear in the above relations) is arbitrary, since in this case the overall expected utility is unchanged regardless of which decision is made.

For a particular observation $\vec{x}$ actually drawn from class π_j, the utility $U_{\pi_i}(\vec{\mathbf{x}}\ \mathbf{t})$ is just a number, which we denote $U_{i|j}$. The expectation values in (7) through (10) can then be evaluated, yielding decision rules of the form

$$\begin{array}{c} \text{decide } d=\pi_i \text{ iff} \\ \displaystyle\sum_{k=1}^{N} U_{i|k}P(\mathbf{t}=\pi_k\lceil\vec{x}) \ge \sum_{k=1}^{N} U_{j|k}P(\mathbf{t}=\pi_k\lceil\vec{x}) \;\; \{j<i\} \\ \text{and} \\ \displaystyle\sum_{k=1}^{N} U_{i|k}P(\mathbf{t}=\pi_k\lceil\vec{x})> \sum_{k=1}^{N} U_{j|k}P(\mathbf{t}=\pi_k\lceil\vec{x}) \;\; \{j>i\}. \end{array} \tag{11}$$

We apply Bayes's rule to obtain

$$P(\mathbf{t}=\pi_k\lceil\vec{x})=\frac{p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_k)P(\mathbf{t}=\pi_k)}{p_{\vec{x}}(\vec{x})} \tag{12}$$

and then multiply the resulting equations by $p_{\vec{x}}(\vec{x})$, which is independent of the summation index k and nonnegative for any feasible observation. The decision rules are now

$$\text{decide } d = \pi_i \text{ iff } \sum_{k=1}^{N} U_{i|k} P(\mathbf{t}=\pi_k) p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_k)$$

$$\geq \sum_{k=1}^{N} U_{j|k} P(\mathbf{t}=\pi_k) p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_k) \quad \{j<i\}$$

$$\text{and } \sum_{k=1}^{N} U_{i|k} P(\mathbf{t}=\pi_k) p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_k)$$

$$> \sum_{k=1}^{N} U_{j|k} P(\mathbf{t}=\pi_k) p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_k) \quad \{j<i\}. \tag{13}$$

By defining $N-1$ likelihood ratios as

$$\text{LR}_i \equiv \frac{p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_i)}{p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_N)} \tag{14}$$

for $i < N$, we can divide the decision rules in (13) by $p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_N)$ and rearrange terms to obtain

$$\text{decide } d = \pi_i \text{ iff } \sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t}=\pi_k) \text{LR}_k$$

$$\geq (U_{j|N} - U_{i|N}) P(\mathbf{t}=\pi_N) \quad \{j<i\} \tag{15}$$

$$\text{and } \sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t}=\pi_k) \text{LR}_k$$

$$> (U_{j|N} - U_{i|N}) P(\mathbf{t}=\pi_N) \quad \{j<i\}. \tag{16}$$

Expression (15), with equality holding, defines a set of $(N/2)$ $(N-1)$ hyperplanes which partition the decision variable space, i.e., the domain of the vector LR whose components are defined by (14), into N subsets. (A portion of each hyperplane determines a boundary between two classes.) The partitioning is determined by the parameters

$$\gamma_{ijk} \equiv (U_{i|k} - U_{j|k}) P(\mathbf{t}=\pi_k) \tag{17}$$

with i, j, and k varying from 1 to N, and $j \neq i$. These parameters are not independent, however, because

$$\gamma_{ijk} + \gamma_{jlk} = \gamma_{ilk}. \tag{18}$$

For a given k, all the other parameters may be derived from γ_{1jk}, with j varying from 2 to N. This leaves a total of $N^2 - N$ parameters over all j and k.

A further constraint may be obtained by noting that the hyperplanes represented by the $N-1$ equations

$$\sum_{k=1}^{N-1} \gamma_{1jk} \text{LR}_k = -\gamma_{1jN} \quad \{2 \leq j \leq N\} \tag{19}$$

are unchanged if we multiply all of these equations by a single scalar, such as $1/|\gamma_{1NN}|$. This reduces the total number of parameters which determine the partitioning of the decision variable space to $N^2 - N - 1$. Note that we cannot multiply different equations by different scalars, as this would change the relative orientations of the remaining decision boundary hyperplanes derived via (18).

Thus, an N-class ideal observer is described by an $(N^2 - N - 1)$-parameter ROC hypersurface in an $(N^2 - N)$-dimensional probability space. These probabilities will be given by expressions of the form

$$
\begin{aligned}
&P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j) \\
&= \int \cdots \int_{\overrightarrow{\mathrm{LR}} \in Z_i^*} p_{\overrightarrow{\mathrm{LR}}}(\overrightarrow{\mathrm{LR}}|\mathbf{t}=\pi_j)d^{N-1}\overrightarrow{\mathrm{LR}}
\end{aligned}
\tag{20}
$$

where the region Z_i^* is the set of all $\overrightarrow{\mathrm{LR}}$ such that $\mathbf{d} = \pi_i$ is decided, as defined in (15) and (16).

III. The Neyman-Pearson Criterion

In the preceding section, we showed that when classifying observations drawn from N classes, the ideal observer, which maximizes expected utility, uses $N - 1$ likelihood ratio decision variables to form its decision rule as stated in (15) and (16). However, this does not directly tell us anything about the actual performance of the ideal observer in terms of its correct and incorrect classification rates, i.e., the probabilities in (20). We wish first to generalize concepts familiar from two-class ROC analysis [7] to the N-class case under consideration, which will enable us to extend the concept of the Neyman-Pearson observer to N classes as well.

The performance of any observer which classifies observations into N classes is determined by the N^2 probabilities $P(\mathbf{d} = \pi_i | \mathbf{t} = \pi_j)$ $\{1 \le i \le N, 1 \le j \le N\}$, where $\mathbf{d} = \pi_i$ indicates a decision that the observation belongs to class i and $\mathbf{t} = \pi_j$ indicates that the observation is actually drawn from class j. As shown above in (6), only $N^2 - N$ of these probabilities are necessary to completely describe the classifier's performance. Without loss of generality, we will eliminate the N probabilities $P(\mathbf{d} = \pi_1 | \mathbf{t} = \pi_1),\ldots, P(\mathbf{d} = \pi_i | \mathbf{t} = \pi_i),\ldots, P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_N)$, i.e., the within-class sensitivities. (In the two-class case, with class π_2 representing "negative" observations, this would be equivalent to using FNF and FPF, or one minus sensitivity and one minus specificity, to specify an observer's performance.)

The $N^2 - N$ probabilities are functions of a set of parameters defined by the observer's decision rule. Given these functions, the observer's performance is given by an ROC hypersurface in which one of the probabilities is an implicit function of the other $N^2 - N - 1$ probabilities; without loss of generality, we take $P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_1)$ (an analogue of the FNF in the two-class case) to be the dependent variable defining our ROC hypersurface.

In the preceding section, it was shown that the ideal observer's performance depends on $N^2 - N - 1$ parameters as well as the N conditional pdfs of $\overrightarrow{\mathbf{x}}$. It may be that a particular observer (not necessarily ideal) uses a decision rule that depends on more than $N^2 - N - 1$ parameters; note that in this case, however, we may define a "simpler" observer whose performance is always given by the set of aggregated parameters which minimize $P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_1)$ at a given set of fixed values of the other $N^2 - N - 1$ probabilities. The performance of this simplified observer is thus effectively determined by only $N^2 - N - 1$ parameters. Furthermore, in this work we will not consider observers which use discrete decision

variables, or which use decision rules dependent on fewer than $N^2 - N - 1$ parameters. Thus, in what follows, we explicitly assume that the performance of any observer under consideration is continuous and dependent on exactly $N^2 - N - 1$ free parameters.

By definition, the Neyman-Pearson observer is that which minimizes the dependent variable (in our case $P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_1)$) at a fixed set of values of the other probabilities; this condition is known as the Neyman-Pearson criterion. Note that the Neyman-Pearson criterion for two classes is typically stated as the minimization of, e.g., $P(\mathbf{d} = \pi_2 | \mathbf{t} = \pi_1)$ for values of $P(\mathbf{d} = \pi_1 | \mathbf{t} = \pi_2)$ less than or equal to a given fixed value. However, the inequality in this formulation is relevant only to the case of discrete decision variables [1] for which arbitrary values of $P(\mathbf{d} = \pi_1 | \mathbf{t} = \pi_2)$ may not in fact be achievable; as stated in the preceding paragraph, we are not considering observers which use discrete decision variables, and thus we will adopt the simplified version of the Neyman-Pearson criterion appropriate for continuous decision variables.

In order to optimize classification performance in a Neyman-Pearson sense, we wish to minimize $P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_1)$ at a particular point in the domain of the probability space, i.e.,

$$P(\mathbf{d} = \pi_i | \mathbf{t} = \pi_j) = \alpha_{ij} \quad \{j \neq i\} \tag{21}$$

where $1 \le i \le N$ and $1 \le j \le N$, and where the term for $i = N, j = 1$ is excluded. The α_{ij} are fixed but arbitrary, apart from the obvious constraints $0 \le \alpha_{ij} \le 1$ and $0 \le \Sigma_{i \ne j} \alpha_{ij} \le 1$. We construct the function

$$
\begin{aligned}
F &\equiv P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_1) \\
&+ \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda_{ij}(P(\mathbf{d} = \pi_i | \mathbf{t} = \pi_j) - \alpha_{ij})
\end{aligned}
\tag{22}
$$

where the case $i = N, j = 1$ is excluded from the sum. When the constraints $P(\mathbf{d} = \pi_i | \mathbf{t} = \pi_j) = \alpha_{ij}$ are all satisfied, minimizing F is equivalent to minimizing $P(\mathbf{d} = \pi_N | \mathbf{t} = \pi_1)$. This is achieved, using the method of Lagrange multipliers [1], by first minimizing F and then finding values of λ_{ij} which satisfy the constraints.

Rearranging terms in (22) and expressing the probabilities explicitly via the definition in (5), we obtain

$$F = \quad -\sum_{\substack{i=1}}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda_{ij}\alpha_{ij} + \int_{Z_N} p(\vec{x}|\mathbf{t}=\pi_1)d^n\vec{x}$$

$$+\sum_{i=1}^{N} \int_{Z_i} \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda_{ij}p(\vec{x}|\mathbf{t}=\pi_j)d^n\vec{x}$$

$$= \quad -\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda_{ij}\alpha_{ij} \qquad (23)$$

$$+\int_{Z_N} \left[p(\vec{x}|\mathbf{t}=\pi_1)+\sum_{j=2}^{N-1} \lambda_{Nj}p(\vec{x}|\mathbf{t}=\pi_j) \right] d^n\vec{x}$$

$$+\sum_{i=1}^{N-1} \int_{Z_i} \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda_{ij}p(\vec{x}|\mathbf{t}=\pi_j)\, d^n\vec{x}$$

where n is the dimensionality of the observations $\vec{x}$, and where the case $i = N, j = 1$ is again excluded from any sum in which it would otherwise occur.

Clearly, F is minimized by choosing the class partitions Z_i so that each observation $\vec{x}$ is assigned to the term in (23) that has the smallest integrand. That is, for $i < N$

$$\text{decide } d = \pi_i \text{ iff } \sum_{\substack{k=1 \\ k \neq i}}^{N} \lambda_{ik}p(\vec{x}|\mathbf{t}=\pi_k)$$

$$< \sum_{\substack{k=1 \\ k \neq j}}^{N} \lambda_{jk}p(\vec{x}|\mathbf{t}=\pi_k) \;\; \{j \neq i\}$$

$$\text{and } \sum_{\substack{k=1 \\ k \neq i}}^{N} \lambda_{ik}p(\vec{x}|\mathbf{t}=\pi_k) \qquad (24)$$

$$< p(\vec{x}|\mathbf{t}=\pi_1)+\sum_{k=2}^{N-1} \lambda_{Nk}p(\vec{x}|\mathbf{t}=\pi_k)$$

and

$$\text{decide } d = \pi_N \text{ iff } p(\vec{x}|\mathbf{t}=\pi_1)+\sum_{k=2}^{N-1} \lambda_{Nk}p(\vec{x}|\mathbf{t}=\pi_k)$$

$$< \sum_{\substack{k=1 \\ k \neq j}}^{N} \lambda_{jk}p(\vec{x}|\mathbf{t}=\pi_k) \qquad (25)$$

where $1 \leq j \leq N-1$, and where the restrictions on the sums are the same as before. Cases for which particular integrands are equal may be decided in an arbitrary but consistent manner.

We now divide (24) and (25) by $p(\vec{x} | \mathbf{t} = \pi_N)$ and rearrange terms to obtain (for $i < N$)

$$\text{decide } d = \pi_i \text{ iff } \sum_{k=1}^{N-1} (\lambda_{jk} - \lambda_{ik})\text{LR}_k > \lambda_{iN} - \lambda_{jN}$$
$$\text{and } \sum_{k=1}^{N-1} (\lambda_{Nk} - \lambda_{ik})\text{LR}_k > \lambda_{iN} \tag{26}$$

$$\text{decide } d = \pi_N \text{ iff } \sum_{k=1}^{N-1} (\lambda_{jk} - \lambda_{Nk})\text{LR}_k > -\lambda_{jN} \tag{27}$$

where $1 \leq j \leq N-1$, and where $\lambda_{N1} \equiv 1$. Note that some of the λ_{ij} above may actually be zero, because we have for clarity removed the explicit restrictions on the sums.

Comparison of (26) and (27) with (15) and (16) reveals that the observer which satisfies the Neyman-Pearson criterion is in fact an ideal observer. That is, given the inherent arbitrariness in the definition of utility, we are free to define

$$U_{i|j} \equiv \frac{\lambda_{ij}}{P(\mathbf{t}=\pi_j)} \tag{28}$$

$$U_{N|N} \equiv 0 \tag{29}$$

yielding inequalities in (15) and (16) identical to those in (26) and (27); furthermore, equalities in (26) and (27) may be decided consistently with those in (15) and (16). The known result for two-class classification tasks, that the ideal observer achieves optimal performance in an ROC sense [1], thus holds for N-class classification tasks as well.

IV. Expected Utility and Optimal Performance: An Alternative View

We have shown that in an N-class classification task, the observer which maximizes expected utility, and that which maximizes performance in an ROC or Neyman-Pearson sense, are the same, namely the ideal observer. It is interesting to ask whether there is a more direct connection between maximization of expected utility and optimization of performance in an ROC sense. Such a connection was explored by one of us [3] for the two-class classification task; in this section we extend those results to the N-class classification task.

Consider again the expected utility as given in (7) through (10), but now in general form rather than with respect to the utility of particular decisions. That is

$$E\{\mathbf{U}\} = \sum_{i=1}^{N} \sum_{j=1}^{N} U_{i|j} P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j) P(\mathbf{t}=\pi_j). \tag{30}$$

As in Section III, we simplify this expression by eliminating N of the N^2 conditional probabilities (namely, the within-class sensitivities) to obtain

$$E\{\mathbf{U}\}= \sum_{j=1}^{N} U_{j|j}P(\mathbf{t}=\pi_j)$$
$$+\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (U_{i|j} - U_{j|j})P(\mathbf{t}=\pi_j)P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j). \tag{31}$$

Although this expression appears cumbersome, it can be regarded quite simply as a linear relation between the expected utility $E\{\mathbf{U}\}$ and the $N^2 - N$ probabilities which have been chosen to represent observer performance; the coefficients in this linear relation are determined by the various decision utilities and the *a priori* class probabilities.

Equation (31) can be rearranged to give $P(\mathbf{d} = \pi_N| \mathbf{t}= \pi_1)$ as a function of the other conditional probabilities (i.e., a locus in ROC space) and of $E\{\mathbf{U}\}$. The result is

$$P(\mathbf{d}=\pi_N|\mathbf{t}=\pi_1)$$
$$=\sum_{i=2}^{N-1} \frac{(U_{i|1}-U_{1|1})}{(U_{1|1}-U_{N|1})}P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_1)$$
$$+\sum_{i=1}^{N} \sum_{\substack{j=2 \\ j \neq i}}^{N-1} \frac{(U_{i|j}-U_{j|j})\,P(\mathbf{t}=\pi_j)}{(U_{1|1}-U_{N|1})P(\mathbf{t}=\pi_1)}P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j) \tag{32}$$
$$+\frac{\left[\sum_{j=1}^{N}U_{j|j}P(\mathbf{t}=\pi_j)\right]-E\{\mathbf{U}\}}{(U_{1|1}-U_{N|1})P(\mathbf{t}=\pi_1)}.$$

For what follows it will be necessary to assume that $U_{1\,|\,1} > U_{N\,|\,1}$; if this were not the case, an increase in expected utility would correspond to an increase in the incorrect-classification probability $P(\mathbf{d} = \pi_N| \mathbf{t} = \pi_1)$ (or $P(\mathbf{d} = \pi_N| \mathbf{t} = \pi_1)$ would be undefined).

Consider a fixed set of decision utilities $U_{i\,|\,j}$ and *a priori* class probabilities $P(\mathbf{t} = \pi_j)$. For this fixed set of values, (32) defines a hyperplane in ROC space. In particular, the operating point achieved by the ideal observer via (15) and (16) using the given decision utilities and *a priori* class probabilities is contained in this hyperplane. The ideal observer achieves a particular value of $E\{\mathbf{U}\}$ when operating at this operating point; an observer using a different decision rule, but still with the same fixed values for the decision utilities and *a priori* class probabilities, would obtain a generally different $E\{\mathbf{U}\}$ and, thus, would have an operating point lying in a hyperplane parallel to that just described. Moreover, because the ideal observer is that which maximizes $E\{\mathbf{U}\}$ for any given set of decision utilities, this second hyperplane would necessarily be above the hyperplane containing the ideal observer's operating point. This implies that no operating points below the hyperplane containing the ideal observer's operating point under consideration are achievable, no matter what decision rule is used.

Now suppose we allow the decision utilities $U_{i\,|\,j}$, or the *a priori* class probabilities, to change slightly. A generally different operating point will be attained by the ideal observer, and this point will lie in a different hyperplane. The remaining arguments of the preceding paragraphs are unchanged, however, and it follows that no decision rule can achieve any operating point lying in the union of the two regions below the hyperplanes containing the two ideal observer operating points in question. Continuing in this fashion for every point in the domain of the ROC space, we find that the ideal observer's performance is given by the concave hull of the set of hyperplanes so defined, and that no observer using any decision rule can achieve a performance (operating point) which lies below this hypersurface.

V. Conclusion

The *N*-class classification task presents many challenges which are daunting when compared with the great successes achieved in analyzing two-class classification tasks. In particular, the number of parameters required by a nontrivial decision rule, and the dimensionalities of the ROC spaces involved, increase rapidly with the number of classes.

Nevertheless, we have shown that certain conclusions regarding ideal observers can in fact be carried over from the two-class task to the *N*-class task. In particular, the two standard definitions of the ideal observer—the observer which maximizes expected utility, and the observer which satisfies the Neyman-Pearson criterion (optimizing ROC performance)—are consistent, as in the two-class case. Furthermore, a direct relation between maximizing expected utility and ROC performance, originally developed for the two-class case, was found to generalize in a straightforward fashion to the *N*-class case.

Whether these promising results can be extended to more practical issues in characterizing observer performance—e.g., fitting an ROC hypersurface to a set of estimated observer operating points—is currently unknown. It is our hope that the present work may at least provide a starting point for addressing these more difficult questions.

Acknowledgments

References

1. Van Trees, HL. Detection, Estimation and Modulation Theory: Part I. New York: Wiley; 1968.

2. Egan, JP. Signal Detection Theory and ROC Analysis. New York: Academic; 1975.

3. Metz, CE. The optimal decision variable," in unpublished lecture notes for the course "Mathematics for Medical Physicists. Dept. Radiology; University of Chicago: 2000.

4. Papoulis, A. Probability, Random Variables, and Stochastic Processes. New York, NY: McGraw-Hill, Inc; 1991.

5. Mossman D. Three-way ROCs. Med Decis Making. 1999; 19:78–89. [PubMed: 9917023]

6. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making. 2000; 20:323–331. [PubMed: 10929855]

7. Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine. 1978; VIII(4):283–298. [PubMed: 112681]