Limits on the discrimination possible with discrete valued data, with application to medical risk prediction

D. R. Lovell, C. R. Dance, M. Niranjan,
R. W. Prager and K. J. Dalton

CUED/F-INFENG/TR.243

January 16, 1996

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
United Kingdom

Limits on the discrimination possible with discrete valued data, with application to medical risk prediction

D. R. Lovell, C. R. Dance, M. Niranjan, R. W. Prager and K. J. Dalton*
{drl,crd,niranjan,rwp}@eng.cam.ac.uk, kjd5@cam.ac.uk

January 16, 1996

Abstract

We describe an upper bound on the *accuracy* (in the ROC sense) attainable in two-alternative forced choice risk prediction, for a specific set of data represented by discrete features. By accuracy, we mean the probability that a risk prediction system will correctly rank a randomly chosen high risk case and a randomly chosen low risk case.

We also present methods for estimating the maximum accuracy we can expect to attain using a given set of discrete features to represent data sampled from a given population.

These techniques allow an experimenter to calculate the maximum performance that could be achieved, without having to resort to applying specific risk prediction methods. Furthermore, these techniques can be used to rank discrete features in order of their effect on maximum attainable accuracy.

Keywords: risk prediction, receiver operating characteristic, accuracy.

1 Introduction

In real life, it is often impossible to predict the outcome of an event with certainty. In these circumstances it is necessary to speak in terms of the probability, or *risk*, of a particular outcome. We are interested in estimating the risk of a specific outcome on the basis of *discretely* valued information.

Any overlap between the classes of outcome associated with discrete features limits the extent to which we can discriminate between those classes. Here, we restrict our attention to the two class problem and establish a theoretical limit on the level of discrimination attainable when identical discrete features are associated with different outcomes. This upper bound allows the experimenter to measure the inherent separability of the data, which may then be compared with the degree of discrimination attained by specific prediction systems.

In this report, discrimination is expressed in terms of the area under the *receiver operating characteristic* (ROC) curve [1]: a quantity referred to as *accuracy*. First, we show how to determine the maximum accuracy attainable on a specific data set with a given set of discrete features. Next, we extend this result so we can estimate the maximum accuracy we could *expect* to obtain with a given set of discrete features. This estimate indicates how useful features are in discriminating between classes, and is especially applicable when data is partitioned into training and test sets.

Once we have shown how to estimate the maximum expected accuracy attainable with a set of discrete features, we present a simple backwards stepwise deletion algorithm which can be used to rank features in terms of the effect they have on attainable accuracy.

The next section gives a formal description of the problem and introduces some of the terminology needed to tackle it. In Section 3 we review the ROC curve and its relation to risk prediction. Sections 4 and 5 present techniques to calculate maximum attainable accuracy for a particular data set, and to estimate the maximum accuracy we can expect to achieve on new data.

Readers who want a concrete example of the methods we discuss in this report *without much mathematical detail* should skip straight to Appendix D.

*The first four authors may be contacted at the Cambridge University Engineering Department, Trumpington St., Cambridge CB2 1PZ, UK. K. J. Dalton is with the Cambridge University Department of Obstetrics and Gynaecology, Rosie Maternity Hospital, Cambridge CB2 2SW, UK.

2 Definition of the problem

This report uses language borrowed from the medical informatics domain, since the results we describe were developed to help tackle the problem of risk prediction in pregnancy [2]. In this context, the two classes we are concerned with describe the outcome of a particular case, that is, *adverse* or *benign*.

It is conventional to label a case expected to have adverse outcome as "positive", and a case expected to have benign outcome as "negative"; hence the terms *false positive* and *false negative* to describe incorrectly classified cases.

Now to the formal description of the problem. Consider a data set in which each case consists of a vector of discrete valued features (or *predictor variables*), $\boldsymbol{x} = (x_1, \ldots, x_N)$, and an outcome, y, which may be either adverse or benign. If feature x_i can take on $|x_i|$ possible values, and X is the set of all distinct feature vectors, the number of elements in X is

$$|X| = \prod_{i=1}^{N} |x_i|.$$

Cases with identical feature vectors (but not necessarily identical *outcomes*) are said to belong to the same *bin*. In more formal terms, if $(\boldsymbol{x}_i, y_i)$ is the i^{th} case in our data set, and $X_j \in X$, then the j^{th} bin, B_j, is defined by the set function

$$\begin{aligned} B_j &= B_j(X_j) \\ &= \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i = X_j\}. \end{aligned}$$

In this context, the purpose of a risk prediction system is to estimate the probability of adverse outcome associated with each bin. Assigning risk in this way imposes an order, or ranking, on the bins. Ideally, this ranking would allow us to discriminate perfectly between cases with different outcomes. That is, all bins ranked below a certain risk value would contain only benign outcome cases, and all bins above that threshold would contain only adverse outcome cases.

In practice, however, we have to deal with bins that contain both benign and adverse outcome cases. What is the optimal way to assign risk (*i.e.*, order bins) in this situation? To answer this, we need a means to assess how effectively a particular ranking of bins discriminates between adverse and benign cases.

3 Risk prediction and the ROC curve

Risk prediction is not an end in itself. Risk is predicted so that appropriate action may be taken. In the simplest situation, once the probability of an adverse outcome has been predicted for a particular case, a decision must be taken as to what the outcome will actually be. In effect, this is equivalent to setting a *risk threshold*, or *cut-point*, above which all outcomes are assumed to be adverse and beneath which all outcomes are assumed to be benign.

Clearly, this *two-alternative forced choice* approach [3] can lead to misclassifications and different cut-points may be selected, depending on the costs of false negative and false positive decisions. Of course, the choice of cut-point does not affect the risk assigned to each case by the prediction system (*i.e.*, the ranking imposed by the system). The ROC curve provides a way to measure — independent of cut-point — how well a particular ranking separates adverse and benign sub-populations.

An ROC curve plots the true and false positive rates obtained over all values of cut-point. Medical informaticians refer to the axes of the ROC as sensitivity *vs* (1 - specificity) [2]; statisticians refer to the axes as the power *vs* the size of a test [4]. Whatever the nomenclature, at a particular cut-point, the values plotted along each axis are defined as

$$\begin{aligned} \text{True positive rate} &= \text{sensitivity, power of test} \\ &= \frac{\text{Number of of true positives}}{\text{Total number of adverse outcomes}} \qquad (1) \end{aligned}$$

$$\begin{aligned} \text{False positive rate} &= \text{(1 - specificity), size of test} \\ &= \frac{\text{Number of of false positives}}{\text{Total number of benign outcomes}}. \qquad (2) \end{aligned}$$

After applying a classifier to a particular data set, how can we assess the classifier's performance from the ROC curve we obtain? Hanley and McNeil [5] state that "the most common quantitative index describing an ROC curve is the area under it". This is due to Green and Swets' result [3] which implies that, for an infinite number of cases, the area under the ROC curve is equal to the probability that a risk prediction system will correctly rank a randomly chosen case with adverse outcome, and a randomly chosen case with benign outcome.

The area under an ROC curve is referred to as *accuracy* and Bamber [6] describes its relationship to the Mann-Whitney U statistic [7]. If X and Y are two sets of continuous observations, the U statistic is defined to be the total number of (x, y) pairs, for all $x \in X$ and $y \in Y$, in which $x < y$. Bamber shows that for the two-alternative forced choice classification of X and Y

$$\text{accuracy} = U/(|X|\,|Y|), \tag{3}$$

where $|X|$ and $|Y|$ are the number of observations in X and Y respectively.

For discrete observations, the U statistic is defined as

$$U = \sum_{\forall x \in X} \sum_{\forall y \in Y} u(x, y), \tag{4}$$

where

$$u(x, y) = \begin{cases} 1 & \text{if } x > y, \\ \frac{1}{2} & \text{if } x = y, \\ 0 & \text{if } x < y. \end{cases} \tag{5}$$

In the context of this paper, the U statistic is the total number of (adverse outcome, benign outcome) case pairs in which the adverse outcome case is ranked at higher risk than the benign outcome case. Clearly, the more discriminative the risk prediction system, the more (adverse, benign) outcome pairs will be correctly ranked and the greater the U statistic of that system.

We now have a way to measure how well a risk prediction system can discriminate between cases with adverse and benign outcome. In the next section we explore how accuracy is limited when cases with identical feature vectors map to different classes.

4 Maximum attainable accuracy

In Section 2, we showed how cases with identical discrete features could be grouped together into bins. A risk prediction system assigns a certain risk to each bin and it is convenient to visualize this as though we were placing the bins in an ordered list, with higher risk bins to the right of lower risk bins (see Figures 1 and 2).

Since the order of the bins depends on the risk assigned by the prediction system, we ask: which ordering of bins maximizes accuracy? It can be shown (Theorem 1, Appendix B) that accuracy is greatest when a risk prediction system ranks bins in ascending order of estimated probability of adverse outcome. For the i^{th} bin, this quantity is defined as

$$\widehat{\Pr}_i(\text{adverse}) = p_i/(p_i + q_i),$$

where B_i contains p_i adverse and q_i benign outcomes.

The results of applying this prescription to the data in Figure 1 are shown in Figures 3 and 4. Note that there may be more than one ordering which maximizes accuracy since bins with equal estimated probability of adverse outcome may be interchanged without violating the ascending order of the list.

After proposing this principle, we found that the concept of ordering data to attain maximum discrimination had been put forward (though not proved) in a slightly different context by Hanley and McNeil [8]. One issue that Hanley and McNeil did not tackle the question of *generalization.* It is feasible to associate each case in our data set with a unique value of a discrete feature (see Appendix D for an example). In this situation, there are as many bins as there are cases and each bin contains a single case. It is trivial to find an ordering which completely discriminates between adverse and benign cases but it is most unlikely that such an ordering would work well on new data. In the next section, we suggest ways to estimate how well discrete features *generalize.*

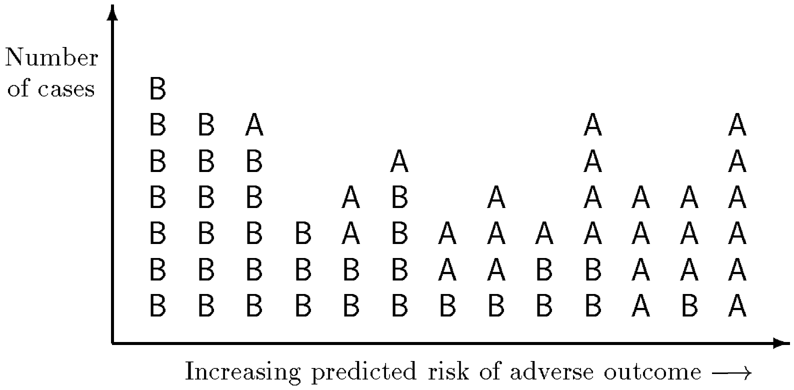


Figure 1: An example of number of cases *vs* risk predicted by some system. Adverse and benign cases are indicated by A and B, respectively. Because we are predicting risk on the basis of discrete features, cases are clumped together in bins. We can think of the risk associated with each bin as imposing an ordering in which predicted risk increases from left to right.

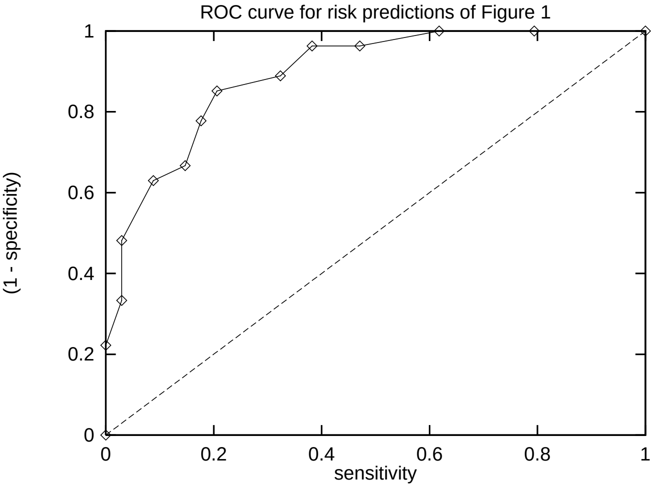


Figure 2: This ROC curve plots the sensitivity and specificity obtained over all possible cut-points for the data in Figure 1. The area under the curve is 0.8927, the probability that the prediction system would rate a randomly chosen adverse case at a higher risk than a randomly chosen benign case. The dotted line is the “chance” ROC curve, *i.e.*, the one that would be obtained if there was 0.5 probability of correctly ranking an adverse and benign outcome pair.

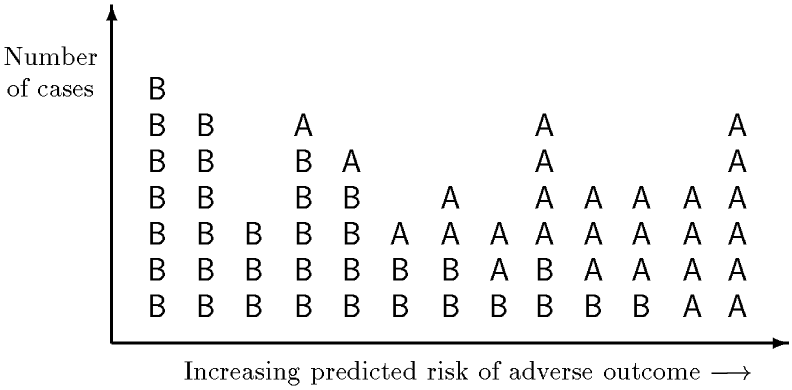


Figure 3: This is one of the orderings of the bins in Figure 1 that maximizes our ability to discriminate between adverse and benign cases.

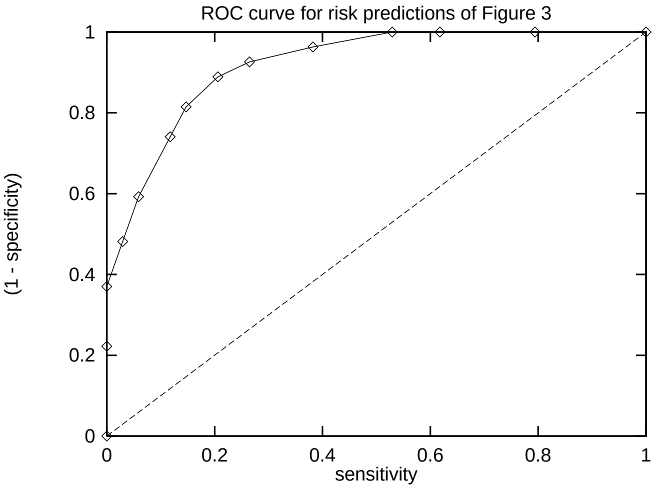


Figure 4: The accuracy of this ROC curve is 0.9199, the maximum attainable by any system for the data in Figures 1 and 3.

5 Finding the expected maximum accuracy

To find the expected maximum accuracy attainable with a given set of discrete features and data drawn from a specific population requires two sets of data. We must:

1. Find the ordering of bins, L^1, which maximizes the accuracy attainable on the first data set.

2. Apply that ordering to the corresponding bins in second data set to give L^2, and measure the accuracy obtained.

This scheme assumes the aim of training a risk prediction system is to maximize its training set accuracy. It could be argued that the aim should be to maximize the system's *test set* accuracy. However, we can easily obtain a bound on test set accuracy by ignoring the training set and applying maximum accuracy ordering to the test set. This approach tells us nothing about how well our features generalize, which is what we wish to find out.

One problem with this scheme is how to order bins in the second data set that do not appear in the first data set. Suppose B_i^1 and B_i^2 are corresponding bins in our two data sets. If $p_i^1 = q_i^1 = 0$ but $p_i^2 + q_i^2 \neq 0$, we say that bin B_i^2 is *unassigned.* There are two meaningful ways to insert an unassigned bin B_i^2 into the ordering of bins from the second data set:

1. Insert B_i^2 randomly into L^2. The accuracy obtained with L^2 will be an estimate of the *average* accuracy we can expect with our discrete features.

2. Insert B_i^2 so as to maximize $U(L^2)$, the Mann-Whitney U statistic for L^2. The accuracy obtained with L^2 will be an estimate of the *maximum* accuracy we can expect with our discrete features.

When there is more than one unassigned bin, method 2 raises the question: if we insert each unassigned bin into the position which maximizes $U(L^2)$ *for that bin only,* do we obtain the maximum $U(L^2)$ once *all* bins have been inserted? Theorems 2 and 3 prove that the answer to this question is "yes".

Three points are worth noting. First, in practice, estimates of expected accuracy obtained with many different sets of data (sampled from the same population) may be averaged together to form a more precise estimate of expected accuracy. Obviously, the larger the sample we use the better our estimates will be.

Second, a common scenario in machine learning involves training a system on one set of data then testing that system on another set of data to assess its ability to generalize (a method known as *hold-out validation*). The techniques we have described allow a bound to be placed on the test set accuracy achievable with a given set of discrete features.

Third, estimates of expected accuracy may also be used to assess the contribution of individual discrete features to the attainable accuracy. We propose the following *backwards deletion* strategy:

$$W_{n+1} = W_n - \operatorname*{argmax}_{x_i \in W_n} \text{accuracy}(W_n - \{x_i\}),$$

where the working set W_n initially contains all the discrete features used to describe the data. The accuracy(W_n) function returns an estimate of the accuracy we can expect to obtain using the features in W_n to represent a given data set. Features are removed from the working set in order of increasing importance to attainable accuracy.

6 Conclusions

The techniques described in this report place upper limits on the discrimination achievable, *by any classification system*, with a particular data set whose records are described by discrete features. This result assumes a two-alternative forced choice setting. We have proposed a method for estimating the expected maximum accuracy attainable when a given set of discrete features is used to describe data sampled from a certain population. We have also put forward a *backwards deletion* strategy to allow us to assess the contribution of individual discrete to the attainable accuracy.

Risk is often predicted on the basis of discrete information. The bounds we have presented allow the experimenter to calculate the degree of discrimination that discrete data permits without having to use a particular risk prediction system. One obvious direction for future investigation is to see how close the accuracy attained by different risk prediction methods (*e.g.*, neural networks, Bayesian belief networks, *etc.*) is to the upper bounds that we can now calculate.

7 Acknowledgements

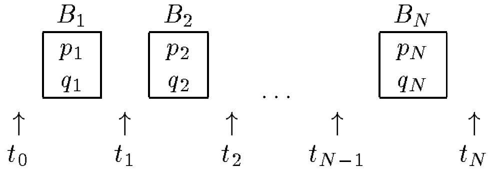
This work was funded by the Quality Assurance in Maternity Care project, EC Biomed Grant Reference 16779-MKI-UO9, led by Dr. Kevin Dalton of the Rosie Maternity Hospital, Cambridge, UK. The first author would like to thank Rudy Lapeer, Richard Shaw and Robert Israel for their suggestions and advice. The first author gratefully acknowledges the support of The British Council in providing travel funding for him to take up the position of research associate with the QAMC project.

A The ROC for a list of bins

Let

$$L = (B_1, B_2, \ldots, B_N)$$

define an ordered list of bins in which bin B_i contains p_i adverse and q_i benign outcomes. We can represent this ordering graphically:



The points $t_0, \ldots, t_N$ indicate thresholds which separate each bin in the ordered list. If we label all items to the right of a threshold as positive, and all items to the left as negative, then the ROC of this list plots the true positive *vs* false positive rates obtained across thresholds $t_0, \ldots, t_N$.

Let the true positive and false positive rates at the i^{th} threshold be written TP_i and FP_i, respectively. From Equations 1 and 2 we have

$$\begin{aligned} TP_i &= \sum_{j>i}^{N} p_j \Big/ \sum_{j=1}^{N} p_j \\ &= \sum_{j>i}^{N} \frac{p_j}{P} \\ FP_i &= \sum_{j>i}^{N} q_j \Big/ \sum_{j=1}^{N} q_j \\ &= \sum_{j>i}^{N} \frac{q_j}{Q}, \end{aligned}$$

where P and Q denote the sum of adverse and sum of benign outcomes, respectively. Note that

$$\begin{aligned} TP_i &= TP_{i+1} + \frac{p_{i+1}}{P}, \\ FP_i &= FP_{i+1} + \frac{q_{i+1}}{Q}. \end{aligned}$$

Using the trapezoidal rule for integration (see Figure 5), the area under the ROC curve can now be written

$$\begin{aligned} \text{Area} &= \sum_{i=0}^{N-1} \tfrac{1}{2}(TP_{i+1} + TP_i)(FP_i - FP_{i+1}) \\ &= \sum_{i=0}^{N-1} \tfrac{1}{2}(2\,TP_{i+1} + \frac{p_{i+1}}{P})\frac{q_{i+1}}{Q} \\ &= \sum_{i=1}^{N} \tfrac{1}{2}(2\,TP_i + \frac{p_i}{P})\frac{q_i}{Q} \\ &= \sum_{i=1}^{N} TP_i \frac{q_i}{Q} + \frac{1}{PQ}\sum_{i=1}^{N} \tfrac{1}{2} p_i q_i \\ &= \frac{1}{PQ}\sum_{i=1}^{N}\sum_{j>i}^{N} p_j q_i + \frac{1}{PQ}\sum_{i=1}^{N} \tfrac{1}{2} p_i q_i \\ &= \frac{1}{PQ}\left(1 \cdot \sum_{i=1}^{N}\sum_{j>i} p_j q_i + \tfrac{1}{2} \cdot \sum_{i=1}^{N} p_i q_i + 0 \cdot \sum_{i=1}^{N}\sum_{j<i} p_j q_i\right) \\ &= \frac{1}{PQ}\sum_{i=1}^{N}\sum_{j=1}^{N} u(B_i, B_j), \end{aligned}$$

where

$$u(B_i, B_j) = \begin{cases} p_j q_i & \text{if } i < j, \\ \frac{1}{2} p_i q_i & \text{if } i = j, \\ 0 & \text{if } i > j. \end{cases} \tag{6}$$

By analogy with Equations 3, 4 and 5, we make the following definition.

Definition 1 We define the Mann-Whitney U statistic for list L as

$$U(L) \;=\; \sum_{i=1}^{N} \sum_{j=i}^{N} u(B_i, B_j),$$

where $u(B_i, B_j)$ is defined according to Equation (6).

Since we are concerned with lists where bins are ranked by assigning a prediction of risk to each bin, we must consider the case where equal rank is assigned to two or more bins. From Definition 1, it is straightforward to show that if bins $B_i, B_{i+1}, \ldots, B_{i+x}$ are assigned equal rank, any permutation of the positions of these bins in list L will not affect the area under the ROC curve provided

$$\widehat{\Pr_i}(\text{adverse}) = \widehat{\Pr_{i+1}}(\text{adverse}) = \cdots = \widehat{\Pr_{i+x}}(\text{adverse}).$$

Furthermore, if this condition holds, bins $B_i, \ldots, B_{i+x}$ may be merged into a single bin B_j where $p_j = p_i + p_{i+1} + \cdots + p_{i+x}$ and $q_j = q_i + q_{i+1} + \cdots + q_{i+x}$. We show this as follows.

Consider three adjacent bins: B_i, B_{i+1} and B_{i+2}. Using the trapezoidal rule, the area under the ROC curve due to these bins is

$$\begin{aligned} \text{Area}(B_i, B_{i+1}, B_{i+2}) &= \tfrac{1}{2}(TP_{i+2} + TP_{i+1})(FP_{i+1} - FP_{i+2}) + \tfrac{1}{2}(TP_{i+1} + TP_i)(FP_i - FP_{i+1}) \\ &= \tfrac{1}{2}\left(2\,TP_i - \frac{2p_{i+1} + p_{i+2}}{P}\right)\frac{q_{i+2}}{Q} + \tfrac{1}{2}\left(2\,TP_i - \frac{p_{i+1}}{P}\right)\frac{q_{i+1}}{Q} \\ &= TP_i \frac{q_{i+1} + q_{i+2}}{Q} - \frac{1}{2PQ}(p_{i+1}q_{i+1} + p_{i+2}q_{i+2} + 2p_{i+1}q_{i+2}). \end{aligned}$$

If bins B_{i+1} and B_{i+2} are merged into a new bin, B_j, such that

$$\begin{aligned} p_j &= p_{i+1} + p_{i+2} \\ q_j &= q_{i+1} + q_{i+2}, \end{aligned}$$

then

$$\begin{aligned} \text{Area}(B_i, B_j) &= \tfrac{1}{2}(TP_i + TP_j)(FP_i - FP_j) \\ &= \tfrac{1}{2}\left(2\,TP_i - \frac{p_{i+1} + p_{i+2}}{P}\right)\frac{q_{i+1} + q_{i+2}}{Q} \\ &= TP_i \frac{q_{i+1} + q_{i+2}}{Q} - \frac{1}{2PQ}(p_{i+1}q_{i+1} + p_{i+2}q_{i+2} + p_{i+1}q_{i+2} + p_{i+2}q_{i+1}). \end{aligned}$$

If this merger has no effect on the area under the ROC curve

$$\text{Area}(B_i, B_{i+1}, B_{i+2}) \;=\; \text{Area}(B_i, B_j),$$

hence,

$$\begin{aligned} 2p_{i+1}q_{i+2} &= p_{i+1}q_{i+2} + p_{i+2}q_{i+1} \\ \frac{p_{i+1}}{q_{i+1}} &= \frac{p_{i+2}}{q_{i+2}} \\ \widehat{\Pr_{i+1}}(\text{adverse}) &= \widehat{\Pr_{i+2}}(\text{adverse}). \end{aligned}$$

This implies that adjacent bins with equal estimated probability of adverse outcome can be merged without affecting the area under the ROC curve.

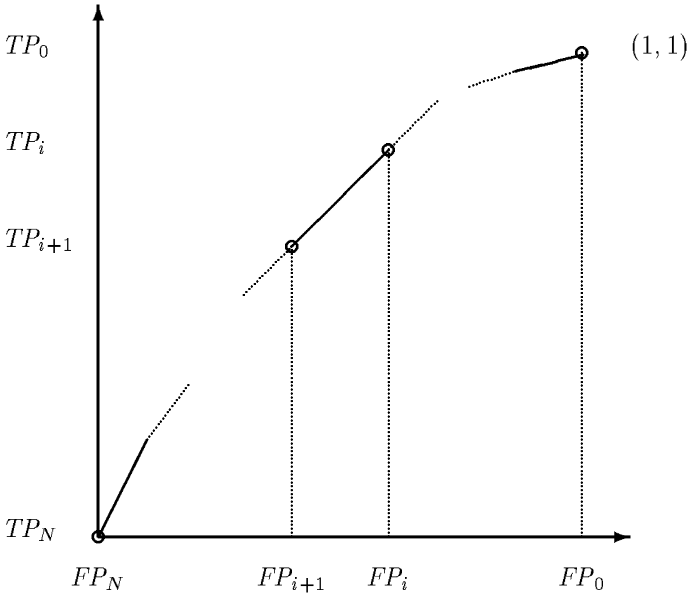


Figure 5: These sections of an ROC curve shows how true and false positive rates alter across the thresholds, $t_0, \ldots, t_N$, which separate the bins, $B_1, \ldots, B_N$, in an ordered list, L. The total area under the curve is the sum of the individual trapezoidal areas underneath the lines connecting adjacent points on the curve.

B Maximum accuracy ordering

By assigning a prediction of risk to each bin, we create an ordered list of bins, L. As discussed in Section 3, the ordering of bins which maximizes the Mann-Whitney U statistic, maximizes accuracy.

We write the Mann-Whitney statistic for list L as $U(L)$ and use the notation

$$L = (B_1, B_2, \ldots)$$

to describe a list in which the predicted risk assigned to bin B_i is less than or equal to the predicted risk assigned to B_j, for all $i < j$. For convenience, we use the relations $<$, $=$, and $>$ when one bin is ranked as having less, equal or greater risk than another.

If bin B_i contains p_i adverse and q_i benign outcomes, the estimated probability of adverse outcome for cases corresponding to this bin is

$$\widehat{\Pr}_i(\text{adverse}) = p_i/(p_i + q_i).$$

Theorem 1 *The Mann-Whitney U statistic for an ordered list of bins is maximized when those bins are ranked in ascending order of estimated probability of adverse outcome.*

Proof: Recall that the list $L = (B_1, \ldots, B_N)$ denotes a list in which $B_1 \leq B_2 \leq \cdots \leq B_N$. From Definition 1, the Mann-Whitney statistic for this list can be written

$$\begin{aligned} U(L) &= \sum_{i=1}^{N}\sum_{j>i}^{N} u(B_i, B_j) + \text{constant} \\ &= \sum_{i=1}^{N}\sum_{j>i}^{N} p_j q_i + \text{constant} \\ &= \sum_{i=1}^{N}\sum_{j>i}^{N} q_i q_j \, a_j/(1 - a_j) + \text{constant}. \end{aligned}$$

where $a_j = \widehat{\Pr}_j(\text{adverse})$.

The double summation contains all $q_i q_j$ terms, for $i < j$. Hence, the value of the double summation depends only on the $a_j/(1-a_j)$ ratios. Since $a_j/(1-a_j)$ is a monotonically increasing function of a_j, the double summation will be maximized when $a_i \leq a_j$ for all $i < j$. In other words, the U statistic for the entire list will be maximized when

$$\widehat{\Pr}_1(\text{adverse}) \leq \widehat{\Pr}_2(\text{adverse}) \leq \cdots \leq \widehat{\Pr}_N(\text{adverse}).$$

□

C Maximum accuracy insertion into an unordered list

We wish to insert one or more non-empty bins into a list of bins so as to maximize the Mann-Whitney U statistic of the final list. Before we show how to do this, we need to extend Definition 1 to define the change in Mann-Whitney U statistic caused by the insertion of a new bin.

Definition 2 From Definition 1 the change in Mann-Whitney U statistic caused by inserting bin B_x into list $L = (B_1, \ldots, B_N)$ at position j is

$$\begin{aligned}\Delta U(B_x, j, L) &= U(B_1, \ldots, B_{j-1}, B_x, B_j, \ldots, B_N) - U(B_1, \ldots, B_{j-1}, B_j, \ldots, B_N) \\ &= p_x \sum_{i=1}^{j-1} q_i + \tfrac{1}{2} p_x q_x + q_x \sum_{i=j}^{N} p_i.\end{aligned}$$

With the necessary definitions in place, we can formally present a method to insert several non-empty bins into a list of bins and maximize the Mann-Whitney U statistic of the final list.

Theorem 2 *Let L^1 and L^2 be lists of bins where*

$$\begin{aligned}L^1 &= (B_1^1, \ldots, B_{N_1}^1), \\ L^2 &= (B_1^2, \ldots, B_{N_2}^2).\end{aligned}$$

L^3 is the list obtained when the elements of L^2 are inserted into L^1. $U(L^3)$ is maximized by inserting each bin, B_j^2, into L^1 at position x_j, where

$$x_j = \operatorname*{argmax}_{x_j} \Delta U(B_j^2, x_j, L^1).$$

Proof: By definition

$$\begin{aligned}U(L^3) &= \sum_{i=1}^{N_3} \sum_{j=i}^{N_3} u(B_i^3, B_j^3) \\ &= \sum_{i=1}^{N_3} \sum_{j>i}^{N_3} u(B_i^3, B_j^3) + \tfrac{1}{2} \sum_{i=1}^{N_1} p_i^1 q_i^1 + \tfrac{1}{2} \sum_{i=1}^{N_2} p_i^2 q_i^2 \\ &= \sum_{i=1}^{N_1} \sum_{j>i}^{N_1} u(B_i^1, B_j^1) + 2 \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u(B_i^1, B_j^2) + \sum_{i=1}^{N_2} \sum_{j>i}^{N_2} u(B_i^2, B_j^2) + \tfrac{1}{2} \sum_{i=1}^{N_1} p_i^1 q_i^1 + \tfrac{1}{2} \sum_{i=1}^{N_2} p_i^2 q_i^2.\end{aligned}$$

The two single summations will stay constant irrespective of insertion order. Since we are inserting elements of L^2 into L^1, the order of the L^1 elements remains the same. Thus the first double summation also remains constant.

The second double summation is maximized by inserting each bin, B_j^2, into L^1 at position x_j, where

$$x_j = \operatorname*{argmax}_{x_j} \Delta U(B_j^2, x_j, L^1).$$

The third double summation is maximized when the elements of L^2 appear in ascending order of $\widehat{\Pr}_i^2$(adverse) in L^3. Theorem 3 proves that the insertion positions which maximize the second double summation make the elements of L^2 appear in ascending order of $\widehat{\Pr}_i^2$(adverse) in L^3. □

Theorem 3 *Let j be the insertion point in list L which maximizes $\Delta U(B_x, j, L)$. If $B_x \leq B_y$ then the position, k, which maximizes $\Delta U(B_y, k, L)$ must be $\geq j$.*

Proof: Since j maximizes $\Delta U(B_x, j, L)$, all insertion points to the left of j would produce a change in U statistic less than or equal to that. Hence,

$$\begin{aligned}\Delta U(B_x, j-n, L) - \Delta U(B_x, j, L) &= q_x \sum_{i=j-1}^{j-n} p_i - p_x \sum_{i=j-1}^{j-n} q_i \\ &\leq 0.\end{aligned}$$

This implies that

$$\frac{\sum\limits_{i=j-1}^{j-n} p_i}{\sum\limits_{i=j-1}^{j-n} q_i} \leq \frac{p_x}{q_x} \leq \frac{p_y}{q_y},$$

since $B_x \leq B_y$. This, in turn, means that

$$\Delta U(B_y, j-n, L) - \Delta U(B_y, j, L) \leq 0,$$

hence there are no positions $k < j$ which produce a change in U statistic greater than $U(B_y, j)$. Therefore, the position, k, which maximizes $\Delta U(B_y, k, L)$ must be $\geq j$.

D A simple example

This section uses a toy problem in perinatal risk prediction to demonstrate the methods described in the body of the report.

D^1 is a set of perinatal data. Each record in the data set describes an individual pregnancy and consists of a vector of discrete features and a single outcome: *adverse* or *benign*.

Each subset of records with identical features is placed in a "bin". Eventually, we obtain a set of these bins, B, where the j^{th} bin

$$B_j \text{ contains} \begin{cases} p_j^1 \text{ records with adverse outcomes} \\ q_j^1 \text{ records with benign outcomes.} \end{cases}$$

A certain risk of adverse outcome is associated with each bin (or, more specifically, each particular combination of features). Suppose we consider three predictors: smoking, drinking and social class. These predictors can take the following discrete values:

smokes:	N or Y (for "no" or "yes")
drinks:	N or Y
class:	L, M, U (for "lower", "middle" or "upper")

Let's say our data set has information 100 pregnant women: their smoking and drinking habits, social class and a particular outcome of their pregnancy (*e.g.*, low birthweight). We might end up with the following bins:

j	class, drinks, smokes	p_j^1	q_j^1
0	L N N	1	4
1	L N Y	1	0
2	L Y N	2	8
3	L Y Y	1	1
4	M N N	1	9
5	M N Y	0	0
6	M Y N	1	9
7	M Y Y	6	14
8	U N N	0	7
9	U N Y	3	7
10	U Y N	4	16
11	U Y Y	2	3

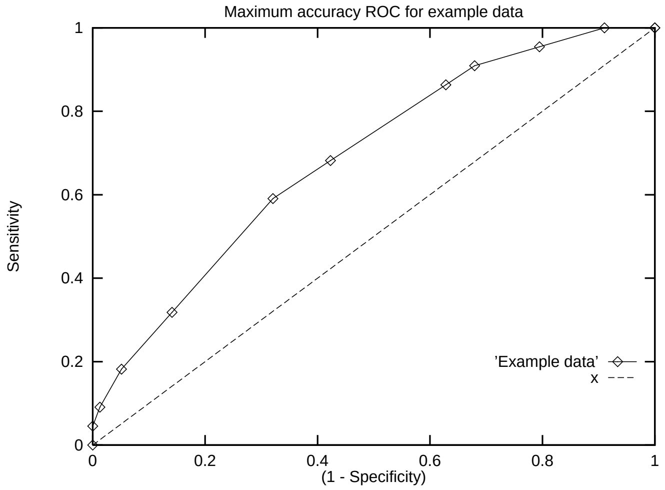
From this information, we might expect a trained classifier to rate a mother who is upper class and drinks and smokes at greater risk than an upper class mother who neither smokes nor drinks. In light of this, the obvious question is "what is the best way to rank these bins?", what ordering of bins would produce maximum accuracy[1]?

It can be shown that the ranking which maximizes accuracy is one which puts bins in ascending order of $\widehat{\Pr}(\text{adverse})$, the estimated probability of adverse outcome. If we apply this prescription to our example, we obtain the following table:

[1] Here we use the term "accuracy" in the ROC sense [1]. That is, the accuracy of a classifier is the probability that it will correctly rank a case that had adverse outcome and a case that had benign outcome. (It is correct to rank the case with adverse outcome as having higher risk.)

j	class, drinks, smokes	p^1_j	q^1_j	$\widehat{\Pr}(\text{adverse})$
5	M N Y	0	0	—
8	U N N	0	7	0.0
4	M N N	1	9	0.1
6	M Y N	1	9	0.1
0	L N N	1	4	0.2
10	U Y N	4	16	0.2
2	L Y N	2	8	0.2
7	M Y Y	6	14	0.3
9	U N Y	3	7	0.3
11	U Y Y	2	3	0.4
3	L Y Y	1	1	0.5
1	L N Y	1	0	1.0

This ordering achieves the maximum possible accuracy (area under ROC curve = 0.689) for that data, given the three features. No classifier can attain better discrimination between adverse and benign outcomes. This is what the corresponding ROC curve looks like:



What would happen if we used an extremely fine grained feature (such as a mother's two digit hospital number) instead of social class, smoking and drinking? Provided no two mothers have the same hospital number, the 100 new bins we would obtain might look like:

j	hosp. no.	p^1_j	q^1_j	$\widehat{\Pr}(\text{adverse})$
0	83	0	1	0.0
1	52	0	1	0.0
2	72	0	1	0.0
3	02	1	0	1.0
4	30	0	1	0.0
⋮	⋮	⋮	⋮	⋮

If we sort these bins in ascending order of $\widehat{\Pr}(\text{adverse})$, the first 78 bins will have $\widehat{\Pr}(\text{adverse}) = 0.0$, the remainder: $\widehat{\Pr}(\text{adverse}) = 1.0$. We will be able to perfectly discriminate between the two groups and the maximum attainable accuracy will be 1.0. This is all very well, provided we only want to work with this limited sample of 100 cases. We are more likely to be interested in the accuracy we can *expect* to achieve on hitherto unseen data.

How can we use our knowledge of optimum ordering to obtain the maximum *expected* accuracy attainable with a given set of features? We find a maximum accuracy ordering of our bins using data set D^1, apply this ordering to the bins containing data from another set, D^2, then measure the accuracy given by that second set.

(In practice, we would randomly partition our entire data set, D, into two disjoint subsets, $D^1, D^2 : D^1 \cup D^2 = D$. Furthermore, to get good estimate of our expected maximum accuracy, we would perform this process many times.)

There is one unresolved issue: what do we do with a bin that occurs in the second data set, but not the first? How should we rank this bin among the other bins in the second set? There are two alternatives:

1. insert that bin randomly

2. insert that bin so as to maximize the accuracy attained on the second data set.

The first option will give us an estimate of the average accuracy that could be attained. The second option will give us an estimate of the best accuracy we could hope for.

Let's go back to our original example, this time with some additional data, D^2; 100 new samples in total. Here is our list of bins with datasets D^1 an D^2 shown alongside each other:

j	class, drinks, smokes	p_j^1	q_j^1	$\widehat{\Pr}_j^1$(adverse)	p_j^2	q_j^2	$\widehat{\Pr}_j^2$(adverse)
0	L N N	1	4	0.20	2	8	0.20
1	L N Y	1	0	1.00	1	3	0.25
2	L Y N	2	8	0.20	3	11	0.21
3	L Y Y	1	1	0.50	2	4	0.33
4	M N N	1	9	0.10	0	4	0.00
5	M N Y	0	0	—	1	0	1.00
6	M Y N	1	9	0.10	1	8	0.11
7	M Y Y	6	14	0.30	3	9	0.25
8	U N N	0	7	0.00	1	9	0.10
9	U N Y	3	7	0.30	3	9	0.25
10	U Y N	4	16	0.20	2	12	0.14
11	U Y Y	2	3	0.40	1	3	0.25

Now we sort the bins in ascending order of $\widehat{\Pr}_j^1$(adverse) to obtain:

j	class, drinks, smokes	p_j^1	q_j^1	$\widehat{\Pr}_j^1$(adverse)	p_j^2	q_j^2	$\widehat{\Pr}_j^2$(adverse)
8	U N N	0	7	0.00	1	9	0.10
4	M N N	1	9	0.10	0	4	0.00
6	M Y N	1	9	0.10	1	8	0.11
0	L N N	1	4	0.20	2	8	0.20
10	U Y N	4	16	0.20	2	12	0.14
2	L Y N	2	8	0.20	3	11	0.21
7	M Y Y	6	14	0.30	3	9	0.25
9	U N Y	3	7	0.30	3	9	0.25
11	U Y Y	2	3	0.40	1	3	0.25
3	L Y Y	1	1	0.50	2	4	0.33
1	L N Y	1	0	1.00	1	3	0.25

Notice that we've left bin B_5 out because data set D^1 gives no indication of where we should rank it. As before, this ranking gives an accuracy of 0.689 with D^1 but the accuracy obtained on data set D^2 is somewhat less: 0.642. (It should be clear from the $\widehat{\Pr}_j^2$(adverse) column that set D^2 is not optimally ranked.

What can we do with B_5? We could insert it into the ranking at random but, for the sake of demonstration, we shall insert it to maximize the D^2 accuracy:

j	class, drinks, smokes	p_j^1	q_j^1	$\widehat{\Pr}_j^1$(adverse)	p_j^2	q_j^2	$\widehat{\Pr}_j^2$(adverse)
8	U N N	0	7	0.00	1	9	0.10
4	M N N	1	9	0.10	0	4	0.00
6	M Y N	1	9	0.10	1	8	0.11
0	L N N	1	4	0.20	2	8	0.20
10	U Y N	4	16	0.20	2	12	0.14
2	L Y N	2	8	0.20	3	11	0.21
7	M Y Y	6	14	0.30	3	9	0.25
9	U N Y	3	7	0.30	3	9	0.25
11	U Y Y	2	3	0.40	1	3	0.25
3	L Y Y	1	1	0.50	2	4	0.33
1	L N Y	1	0	1.00	1	3	0.25
5	M N Y	0	0	—	1	0	1.00

Inserting bin 5 in the above position gives a D^2 accuracy of 0.642.

As mentioned earlier, we would need to perform this process many times, with many different data sets to obtain a good estimate of the expected accuracy given by this set of features and this data. Once we have obtained a good estimate of expected accuracy, we are then in a position to determine the expected accuracy we obtain when we remove different features. This *backwards deletion* process gives us a picture of how each feature contributes to the attainable accuracy.

References

[1] J. A. Swets and R. M. Pickett, *Evaluation of diagnostic systems.* New York: Academic Press, 1982.

[2] D. K. James and G. M. Stirrat, eds., *Pregnancy and Risk.* Wiley, 1988.

[3] D. Green and J. Swets, *Signal detection theory and psychophysics.* New York: John Wiley, 1966.

[4] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics.* McGraw-Hill, 3rd ed., 1974.

[5] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[6] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, pp. 387–415, 1975.

[7] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, pp. 50–60, 1947.

[8] J. Hanley and B. McNeil, "Maximum attainable discrimination and the utilization of radiologic examinations," *Journal of Chronic Disease*, vol. 35, pp. 601–611, 1982.