School of Engineering, Computer Science
and Mathematics,
University of Exeter. EX4 4QF. UK.
http://www.ex.ac.uk/secsm

A short note on the efficient random sampling of the multi-dimensional pyramid between a simplex and the origin lying in the unit hypercube

Jonathan E. Fieldsend
School of Engineering, Computer Science and Mathematics,
University of Exeter, Exeter, EX4 4QF, UK
J.E.Fieldsend@exeter.ac.uk

24th August 2005

Abstract

When estimating how much better a classifier is than random allocation in Q-class ROC analysis, we need to sample from a particular region of the unit hypercube: specifically the region, in the unit hypercube, which lies between the $Q-1$ simplex in $Q(Q-1)$ space and the origin.

This report introduces a fast method for randomly sampling this volume, and is compared to rejection sampling of uniform draws from the unit hypercube. The new method is based on sampling from a Dirichlet distribution and shifting these samples using a draw from the Uniform distribution. We show that this method generates random samples within the volume at a probability $\approx 1/(Q(Q-1))$, as opposed to $\approx (Q-1)^{Q(Q-1)}/(Q(Q-1))!$ for rejection sampling from the unit hypercube.

The vast reduction in rejection rates of this method means comparing classifiers in a Q-class ROC framework is now feasible, even for large Q.

1 Introduction

In our recent work, we have introduced a method for extending the binary class Receiver Operating Characteristic (ROC) analysis (Hanley and McNeil [1982]; Zweig and Campbell [1993]) to the multi-class domain (Everson and Fieldsend [2005]; Fieldsend and Everson [2005]). Rather than considering the true and false positive rates, we considered the multi-class ROC surface to be the solution of the multi-objective optimisation problem in which these misclassification rates are simultaneously optimised. Srinivasan [1999] has discussed a similar formulation of multi-class ROC, showing that if classifiers for Q classes are considered to be points with coordinates given by their $Q(Q-1)$ misclassification rates, then optimal classifiers lie on the convex hull of these points.

In the evaluation of this multi-class ROC analysis, we developed a straightforward generalisation of the Gini coefficient which quantified the superiority of a classifier's performance to random allocation. In order to estimate this value, the need arises to generate uniformly distributed samples from a specific region of the unit hypercube, namely the region corresponding to that lying between the $Q-1$ simplex in $Q(Q-1)$ space and the origin, which defines classifier misclassification rate space which is better than random allocation. The cost of sampling this region becomes prohibitively large at even relatively small Q when rejection sampling from the unit hypercube is used. Here we introduce a method of uniform sampling from the space which is vastly less expensive.

2 Comparing classifiers

In two class problems the area under the ROC curve (AUC) is often used to compare classifiers. As clearly explained by Hand and Till [2001], the AUC measures a classifier's ability to separate two classes over the range of possible costs and is linearly related to the Gini coefficient.

By analogy with the AUC, we can use the volume of the $Q(Q-1)$-dimensional hypercube that is dominated by elements of the ROC surface for classifier A as a measure of A's performance. In binary and multi-class problems alike its maximum value is 1 when A classifies perfectly. If the classifier allocates at random, the ROC surface is the simplex in $D = Q(Q-1)$-dimensional space with vertices at length $Q-1$ along each coordinate vector. The volume of the unit hypercube dominated by this can be derived as follows: As the volume of the pyramidal region between the origin and the simplex with vertices at a distance L along each coordinate vector is $\frac{L^D}{D!}$. The volume lying between the origin and the random allocation simplex is, therefore:

$$\frac{(Q-1)^D}{D!}. \tag{1}$$

Only part of this volume lies in the unit hypercube however, as the corners (excluding that at the origin) relate to infeasible regions where classification rates are > 1. Each of these D corner regions is also a pyramidal volume, but with sides of length $Q-2$. The total volume of the region between the origin and the random allocation simplex which *also* lies in the unit hypercube is therefore

$$\frac{(Q-1)^D}{D!} - \frac{D(Q-2)^D}{D!}. \tag{2}$$

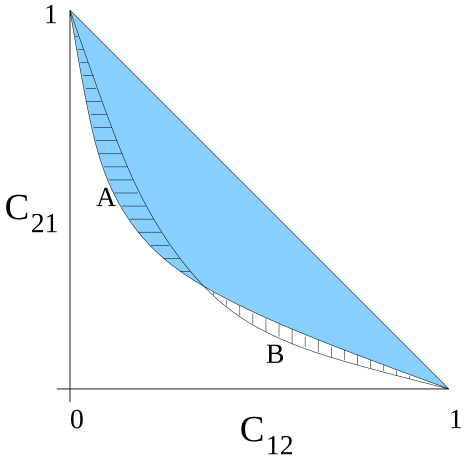

Figure 1. Illustration of the G and δ measures where $Q = 2$. Shaded area denotes $G(A)$, horizontally dashed area denotes $\delta(A, B)$, vertically dashed area denotes $\delta(B, A)$.

We denote this region by P. When $Q = 2$ the second term in equation 2 is zero so that the total volume (area) between the origin and the random allocation simplex is just 1/2. This corresponds to the area under the diagonal in a conventional ROC plot (although binary ROC plots are usually made in terms of true positive rates versus false positive rates for one class, the false positive rate for the other class is just 1 minus the true positive rate for the other class). However, when $Q > 2$, the volume not dominated by the random allocation simplex is very small; even when $Q = 3$, the volume not dominated is ≈ 0.0806. We therefore define $G(A)$ to be the analogue of the Gini coefficient in two dimensions, namely the proportion of the volume of the D-dimensional unit hypercube that is dominated by elements of the ROC surface, but is not dominated by the simplex defined by random allocation (as illustrated by the shaded area in Figure 1 for the $Q = 2$ case). In binary classification problems this corresponds to twice the area between the ROC curve and the diagonal. In multi-class problems $G(A)$ quantifies how much better A is than random allocation. It can be simply estimated by Monte Carlo sampling of this volume in the unit hypercube.

2.1 Rejection sampling from the unit hypercube

A simple manner to generate uniform samples, $\mathbf{x}$, of this volume is to generate uniform samples from the unit hypercube $\mathbf{x} \sim \mathcal{U}(0,1)^{Q(Q-1)}$ and reject those values where $\sum \mathbf{x}_i \geq Q-1$ or where any element $x_i > 1$ (as shown in Algorithm 1). We know from Equation 2 the probability of generating an acceptable sample, and Figure 2 shows this for $Q = 2, \ldots, 10$.

Algorithm 1 Uniform rejection sampling from the unit hypercube.

```
Inputs:
    $N$    Number of random samples from region
    $Q$    Number of classes
    $X$    Set of $N$ random samples from region
$n = 0$
$X = \{\}$
while $n \neq N$
    $\mathbf{x} := \mathcal{U}(0,1)^{Q(Q-1)}$
    if $\sum \mathbf{x}_i < Q-1 \wedge x_i \leq 1 \; \forall i$
        $X := X \cup \mathbf{x}$
        $n := n+1$
    end
end
```

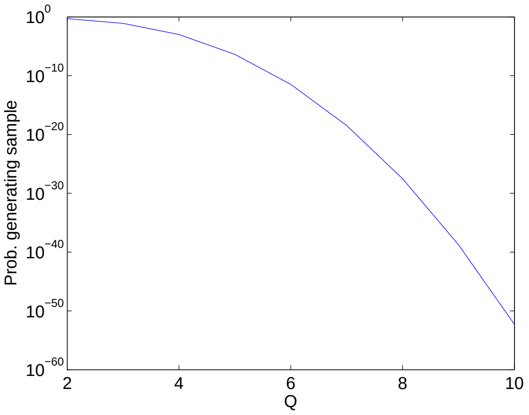

Figure 2. The probability of generating a random value in the unit hypercube which is not dominated by the simplex defined by random allocation.

As can be seen, this value becomes vanishingly small as Q increases, meaning that rejection sampling from the hypercube becomes infeasible for even relatively small Q. This necessitates a less wasteful generation of random samples before Q-class ROC classifier comparison becomes computationally feasible. This can be achieved through the combined use of a Dirichlet and Uniform distribution, which we now discuss.

2.2 Using a transformation of the Dirichlet

First let us concern ourselves with the generation of samples below the simplex. The Dirichlet distribution generates samples from the simplex where $\sum \mathbf{y}_i = 1$

$$p(\mathbf{y}) = Dir(\mathbf{y} \,|\, \alpha_1, \ldots, \alpha_D) \tag{3}$$

$$= \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{D-1} y_i\right)^{\alpha_D - 1} \prod_{i=1}^{D-1} y_i^{\alpha_i - 1} \tag{4}$$

where the index i labels the $D = Q(Q-1)$ elements. The $\alpha_i \geq 0$ determine the density of the samples; by setting all the $\alpha_i = 1$ the simplex is sampled uniformly with respect to Lebesgue measure. Figure 3 shows 10000 samples generated from the Dirichlet distribution where $D = 2, 3$.

Algorithm 2 Random sampling from the region of interest using a combination of Dirichlet and Uniform distributions.

```
Inputs:
    $N$    Number of random samples from region
    $Q$    Number of classes
    $X$    Set of $N$ random samples from region
    $\mathbf{1}$    A $Q(Q-1)$-dimensional vector of ones
$n = 0$
$X = \{\}$
while $n \neq N$
    $\mathbf{y} := Dir(\mathbf{1})$
    $z := \mathcal{U}(0, \frac{1}{Q(Q-1)})$
    $\boldsymbol{\theta} := (Q-1)(\mathbf{y} - z\mathbf{1})$
    if $\theta_i \leq 1 \; \forall i \wedge \theta_i \geq 0 \; \forall i$
        $X := X \cup \mathbf{x}$
        $n := n + 1$
    end
end
```

If we now generate a uniform sample $z \sim \mathcal{U}(0, 1/D)$, then $\mathbf{y} - z\mathbf{1}$, where $\mathbf{1}$ is a D-dimensional vector of 1s, creates a transformed sample, $\boldsymbol{\theta}$, uniformly distributed in a hyper-prism. Figure 4 shows samples generated for Figure 3 shifted in this fashion.

A simple check of sign of the elements of $\boldsymbol{\theta}$ then allows us to extract those lying in the pyramid between the origin and the simplex (as shown in Figure 5).

In order to transform these points to ones lying beneath the $Q-1$ simplex the elements are simply multiplied by $Q-1$. As this doesn't affect the assignment to whether a point lies in the pyramid or not, the illustrations using unit simplex (Figures 3-5) are general. Algorithm 2 steps through this process

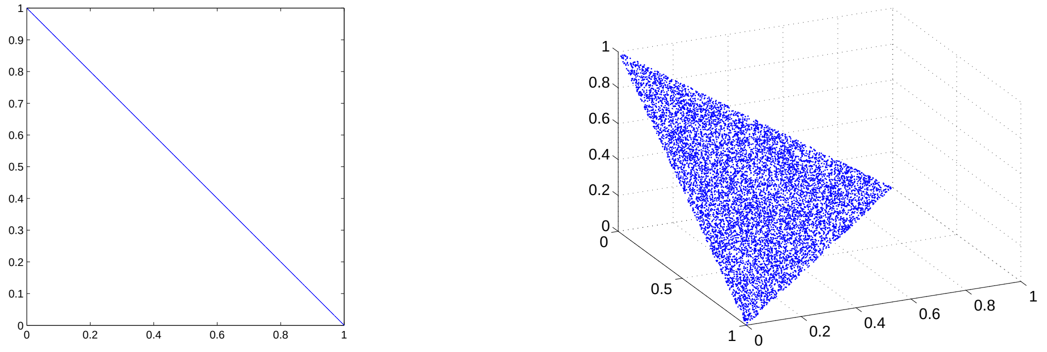

Figure 3. 10000 samples generated from the Dirichlet distribution. *Left:* $D = 2$; *Right:* $D = 3$.

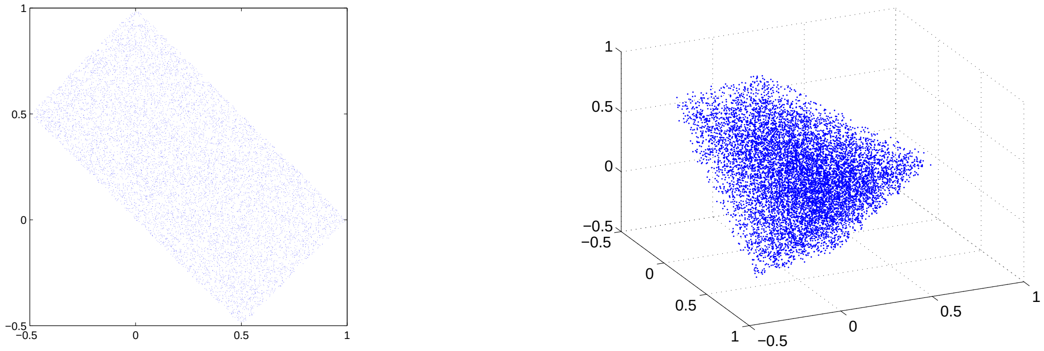

Figure 4. 10000 samples generated from the Dirichlet distribution, shifted by the subtraction of uniformly distributed values. *Left:* $D = 2$; *Right:* $D = 3$.

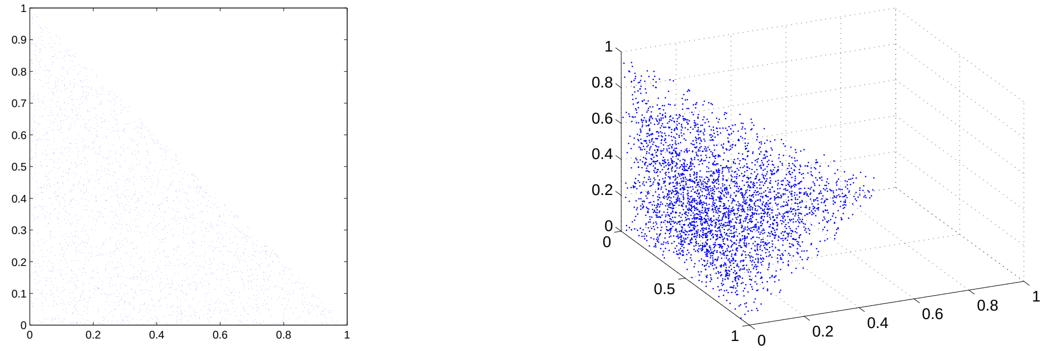

Figure 5. Those samples generated from a Dirichlet distribution and shifted by the subtraction of uniformly distributed values, which lie in the pyramid. *Left:* $D = 2$; *Right:* $D = 3$.

3 Rejection rate of the proposed approach

For any given Q we can easily define the convex hull of the pyramidal region bounded by the simplex and the origin, and also the parallelapiped defined by the sampled volume of the simplex projection.

3.1 Volume ratios

Table 1. Calculated volumes of regions for various Q.

Q	2	3	4	5	6
Pyramid volume	0.5	8.89×10^{-2}	1.11×10^{-3}	4.52×10^{-7}	3.51×10^{-12}
$(Q-1)^{Q(Q-1)}/(Q(Q-1))!$	0.5	8.89×10^{-2}	1.11×10^{-3}	4.52×10^{-7}	3.51×10^{-12}
Sampled volume	1	5.33×10^{-1}	1.33×10^{-2}	9.04×10^{-6}	1.05×10^{-10}
Volume ratio	2	6	12	20	30

Table 1 shows the volumes of these two regions calculated for various Q. As can clearly be seen, the ratio of the two regions' volumes is actually D. The proportion of samples generated in this two step fashion that lie in the pyramid between the origin and the simplex is therefore $1/D$. Using equation 2, the proportion of points generated in this fashion that lie between the $Q-1$ simplex and the origin and *also* lie in the unit cube is:

$$= \frac{1}{D} \frac{\left(\frac{(Q-1)^D}{D!} - \frac{D(Q-2)^D}{D!} \right)}{\left(\frac{(Q-1)^D}{D!} \right)} \tag{5}$$

$$= \frac{1}{D} \left(1 - \frac{D(Q-2)^D}{(Q-1)^D} \right) \tag{6}$$

$$\approx \frac{1}{Q(Q-1)} \quad \text{as } Q \text{ becomes large.} \tag{7}$$

These points can essentially be found from those lying in the pyramidal region by rejecting those with elements $\theta_i > 1$. Figure 6 shows the probability of generating an acceptable sample using this method for $Q = 2, \ldots, 10$.

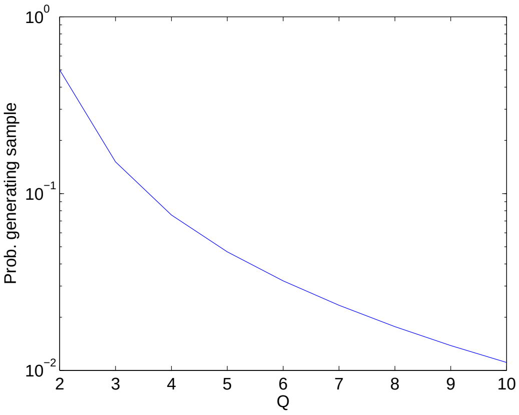

Figure 6. The probability of generating a random value in the unit hypercube which is not dominated by the simplex defined by random allocation, using the method of combing samples from Dirichlet and Uniform distributions.

Note that the second term in (5) rapidly approaches one, so the proportion of samples generated between the random allocation simplex and the origin that lie in the unit hypercube can be taken as $\approx 1/(Q(Q-1))$.

4 Example

If every point on the optimal ROC surface for classifier A is dominated by a point on the ROC surface for classifier B, then classifier B has a superior performance to classifier A. In general, however, neither ROC surface will completely dominate the other: regions of A's surface will be dominated by B and vice versa; in binary problems this corresponds to ROC curves that cross. To quantify the classifier's relative performance we therefore define $\delta(A, B)$ to be the volume of P that is dominated by elements of A and not by elements of B (marked in Figure 1 with horizontal lines). Note that $\delta(A, B)$ is not a metric; although it is non-negative, it is not symmetric. Also if A and B are subsets of the same non-dominated set W, (i.e., $A \subseteq W$ and $B \subseteq W$), then $\delta(A, B)$ and $\delta(B, A)$ may have a range of values depending on their precise composition; see Fieldsend et al. [2003] for more details. Situations like this are rare in practice, however, and measures like δ have proved useful for comparing Pareto fronts.

Table 2. Generalised Gini coefficients and exclusively dominated volume comparisons of the multinomial logistic regression (MLR) and k-nn classifiers. Taken from Fieldsend and Everson [2005].

Measure	Synth.	Vehicle	Image
$G(\text{MLR})$	0.840	≈ 0	≈ 0
$G(k\text{-nn})$	0.920	0.168	0.076
$\delta(\text{MLR}, k\text{-nn})$	0.001	0.000	0.000
$\delta(k\text{-nn}, \text{MLR})$	0.081	0.168	0.076

Table 2 shows the results of comparing the performance of the probabilistic k-nn classifier [Holmes and Adams, 2002] with that of the multinomial logistic regression classifier on Synthetic data, and also on the UCI Image and Vehicle data sets [Blake and Merz, 1998]. These were calculated from 10^5 uniform samples from the region between the random allocation simplex and the origin. Numerically the values obtained using rejection sampling from the unit hypercube for the Synthetic and Vehicle data sets were the same as those presented here. The exorbitant computation cost of rejection sampling for the $Q(Q-1) = 42$ case of the Image dataset meant we did not compare the two techniques for the final set.

5 Conclusion

A method has been introduced for the uniform sampling of the space between the random allocation simplex and the origin for multi-class ROC assessment purposes. This method has been shown to be orders of magnitude more efficient than the simple rejection sampling of the unit hypercube – with its comparable efficiency increasing with the number of classes. This method generates random samples within the volume at a probability $\approx 1/(Q(Q-1))$, as opposed to $\approx (Q-1)^{Q(Q-1)}/(Q(Q-1))!$ for rejection sampling from the unit hypercube.

Matlab code to generate samples from this region is provided from the author's website at `http://www.dcs.ex.ac.uk/people/jefields/`, along with code for proving the sample rejection rate.

Acknowledgements

The author would like to thank Richard Everson for his help and discussion in the generation of this report.

References

C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

R.M. Everson and J.E. Fieldsend. Multi-class ROC analysis from a multi-objective optimisation perspective. Technical Report 421, Department of Computer Science, University of Exeter, April 2005.

J.E. Fieldsend and R.M. Everson. Formulation and comparison of multi-class ROC surfaces. In *Proceedings of the 2nd ROC Analysis in Machine Learning Workshop, part of the 22nd International Conference on Machine Learning (ICML 2005)*, pages 41–48, 2005.

J.E. Fieldsend, R.M. Everson, and S. Singh. Using Unconstrained Elite Archives for Multi-Objective Optimisation. *IEEE Transactions on Evolutionary Computation*, 7(3):305–323, 2003.

D.J. Hand and R.J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 82(143):29–36, 1982.

C.C. Holmes and N.M. Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Journal Royal Statistical Society B*, 64:1–12, 2002.

A. Srinivasan. Note on the location of optimal classifiers in n-dimensional ROC space. Technical Report PRG-TR-2-99, Oxford University Computing Laboratory, Oxford, 1999. URL `ftp://ftp.comlab.ox.ac.uk/pub/Packages/ILP/Papers/AS/roc.ps.gz`.

M.H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–577, 1993.