

# Understanding Multiclass Extensions of ROC/AUC and their Relationships

Bowen Ke

(PennKey: bowenke; Email: bowenke@seas.upenn.edu)

Hui Lyu

(PennKey: huilyu; Email: huilyu@seas.upenn.edu)

Simeng Sun

(PennKey: simsun; Email: simsun@seas.upenn.edu)

## Abstract

There are three multiclass extensions of ROC curve/AUC: pairwise, one-vs-rest and VUS. However, the relationships between them are not well understood. This project has derived comparable mathematical expressions for each of them for theoretical comparison, and also conducted two experiments for empirical analysis of their pros and cons.

## 1 Introduction

A *Receiver operating characteristics* (ROC) curve is a 2-dimensional plot where y-axis is the true positive rate (TPR) and x-axis is the false positive rate (FPR). It can be generated by changing the threshold  $t$  for a binary classifier  $h_t(x) = \text{sign}(f(x) - t)$  where  $f$  is a scoring model  $f: \mathcal{X} \rightarrow \mathbb{R}$ . The *area under the ROC curve* (AUC) is a single number measuring the overall performance of the scoring model. Though AUC for binary problems is well-defined and effective in real cases, the multiclass extensions are not well studied and explained. There are in general three kinds of ROC/AUC multiclass extensions, each of them has different settings and lacks either statistical explanation or relationship with binary cases. Therefore, we aim to understand the relationships among the existing ROC/AUC multiclass extensions and compare them theoretically and empirically.

Our contributions are as follows: (1) We unified binary AUC and multiclass AUC extensions using class score density function. (2) We compared multiclass AUC extensions with binary AUC theoretically and analyzed their pros and cons (3) Based on our derived expression, we conducted simulation and explored how class separability and class imbalance affects each kind of multiclass extensions.

## 2 Related work

### 2.1 Preliminaries

The ROC curve can be plotted by varying the threshold given the output of a scoring model  $f : \mathcal{X} \rightarrow \mathbb{R}$ . Each operating point is a pair of  $(FPR, TPR)$  and corresponds to a specific confusion matrix. The area under the curve is a scalar performance measure for evaluating classifiers. Closely related to the Gini Coefficient, it is the probability a randomly chosen positive instance having higher rank than a randomly chosen negative instance [1].

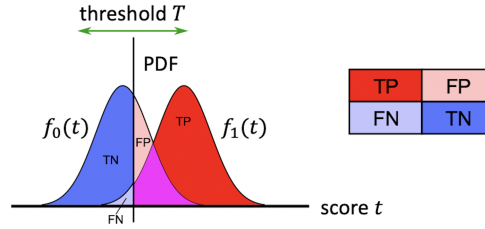


Figure 1: Class score density for binary problem.  $f_1(t)$  is the score distribution for positives and  $f_0(t)$  is the score distribution for negatives

Both true positive rate and false positive rate can be related to threshold in a form of integral  $TPR = \int_T^\infty f_1(t)dt$ ,  $FPR = \int_T^\infty f_0(t)dt$ , where the score of positives follows the probability density function  $f_1(\cdot)$ , that of negatives follows  $f_0(\cdot)$ . The densities are jointly decided by the classifier and the data set. There is previous work estimating such distribution using Gaussian Mixture Model [5]. A more generic depiction is shown in Figure 1<sup>1</sup>. True positive rate is therefore the area under  $f_1(t)$  when  $t$  is greater than certain threshold  $T$ .  $T$  tending to infinity corresponds to the point  $(0, 0)$  in ROC space.

### 2.2 Multiclass AUC

There are in general 3 kinds of multiclass AUC extensions, two of which are approximations to the real exact number that characterizes the performance of a multiclass classifier.

#### 2.2.1 Pairwise

One approximation approach is the Pairwise AUC [2]. It measures what the authors called “separability” of each pair of classes and originates directly from the statistical meaning of binary AUC. Suppose there are total  $c$  classes, for each of the  $\binom{c}{2}$  pair of classes, they use  $A(c_i, c_j)$  as a measure of the separability for class  $c_i$  and  $c_j$ . Pairwise AUC is the unweighted average over  $A(c_i, c_j)$  of each pair of classes.

<sup>1</sup>Image from wikipedia: [https://upload.wikimedia.org/wikipedia/commons/thumb/4/4f/ROC\\_curves.svg/600px-ROC\\_curves.svg.png](https://upload.wikimedia.org/wikipedia/commons/thumb/4/4f/ROC_curves.svg/600px-ROC_curves.svg.png)

$$\text{AUC}_{\text{pairwise}} = \frac{2}{c(c-1)} \sum_{c_i < c_j} A(c_i, c_j) \quad (1)$$

The separability of each pair is an average of  $A(c_i|c_j)$  and  $A(c_j|c_i)$ , where  $A(c_i|c_j)$  is the probability that a randomly drawn instance from class  $c_i$  has a higher estimated probability of belong to class  $c_i$  than a randomly drawn instance from class  $c_j$ .

### 2.2.2 One-vs-Rest

Another kind of multiclass AUC extension adopts an one-vs-rest strategy [4]. This kind of AUC is a weighted average of  $c$  AUCs.  $\text{AUC}(c_i)$  is the area under a ROC curve which is plotted by viewing only  $c_i$  as positive and the union of the rest  $c-1$  classes as negative.  $p(c_i)$  is the prevalence of class  $c_i$  among total  $m$  instances.

$$\text{AUC}_{1\text{-vs-rest}} = \sum_{i=1}^c \text{AUC}(c_i)p(c_i) \quad (2)$$

$$p(c_i) = \frac{\sum_{j=1}^m \mathbb{I}(y_j = c_i)}{m} \quad (3)$$

### 2.2.3 Volume under hyper-surface

The most exact generalization of multiclass AUC is the volume under hyper-surface (VUS). Pairwise and one-vs-rest are approximations to the exact VUS. VUS is a direct extension of binary AUC where we change the y-axis  $TPR$  to false negative rate ( $FNR$ ). The optimal classifier for this kind of ROC variant has the minimum possible area under the curve. Consider all off-diagonal entries in a confusion matrix, it maps to an operating point in  $c(c-1)$ -dimensional space. Extending directly from the binary ROC variant, the volume under the hyper-surface formed by varying a set of thresholds is expected to be the minimum for optimal classifiers. When there are 3 classes, it becomes immediately difficult to analyze and visualize since the surface is in 6-dimensional space. This is the reason why people reduce the true extension to binary cases where good interpretability and visualization preserve.

Besides, there is another kind of VUS which is an approximation to the real VUS where only the diagonal entries of a confusion matrix are considered and therefore corresponds to a point in a  $c$ -dimensional space [3]. The real VUS is computed by transforming to a constraint satisfaction problem. The maximum and minimum VUS can be computed by enforcing constraint subject to normalized confusion matrix. However, the authors didn't provide any analytical expression for the exact VUS.

### 3 Problem settings and notations

We consider general multiclass classification problems. We do not require the classes to satisfy certain order. The number of classes is denoted by  $c$ . We consider scoring model where there are total  $c$  scoring functions  $s_1, \dots, s_c$ , each instance gets a  $c$ -dimensional score vector  $S(\mathbf{x}) = (s_1(\mathbf{x}), \dots, s_c(\mathbf{x}))^\top$ . We define  $f_{ij}(\cdot)$  to be probability density function of the  $i^{\text{th}}$  entry of  $S(\mathbf{x})$  for every instance  $\mathbf{x}$  belonging to class  $j$ .

### 4 Theoretical analysis

In this section, we rederive the integral form of binary AUC and show how it can be unified with multiclass AUC extensions using the score density function  $f_{ij}(\cdot)$  and the thresholds  $T_{ij}$  (introduced later).

#### 4.1 Binary AUC

According to the integral form of TPR and FPR and the geometry meaning of AUC, we have the following expression for binary AUC:

$$\text{AUC}_{\text{binary}} = \int_{-\infty}^{\infty} \text{TPR}(T) \text{FPR}'(T) dT \quad (4)$$

$$= \int_{-\infty}^{\infty} \left( \int_T^{\infty} f_1(T') dT' \right) (-f_0(T)) dT \quad (5)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T' > T) f_1(T') f_0(T) dT' dT \quad (6)$$

The integration in equation 4 is from infinity to negative infinity since we integrate  $\text{FPR}$  from 0 to 1, which is equivalent to varying threshold from  $\infty$  to  $-\infty$ . Intuitively, it can be interpreted as putting two thresholds on the score distribution, one for each  $f_i(\cdot)$  and both vary from  $-\infty$  to  $\infty$ . Then AUC is the area where the threshold for  $f_1(t)$  is larger than that for  $f_0(t)$ . The upper left area above the ROC curve is thus characterized by the indicator function  $\mathbb{I}(T' \leq T)$ .

#### 4.2 Pairwise

Pairwise AUC, which has been introduced in Section 2.2.1 is the unweighted “separability” of each pair of classes. The separability between class  $i$  and class  $j$  is the average of  $A(i|j)$  and  $A(j|i)$ . Extending their explanation for CPE model, in our setting,  $A(i|j)$  is the probability that the  $i^{\text{th}}$  score of a random instance from class  $i$  is greater than the  $i^{\text{th}}$  score of an instance randomly drawn from class  $j$ ,  $\forall j \neq i$ . Notice that  $\text{AUC}_{\text{binary}} = A(1|0)$ , which builds a connection between binary AUC and pairwise multiclass AUC. Figure 2 shows 6 possible pairs for a 3-class classification problem.  $A(1|3)$  represents the case in which we select class 1 as positive and class 3 as negative.

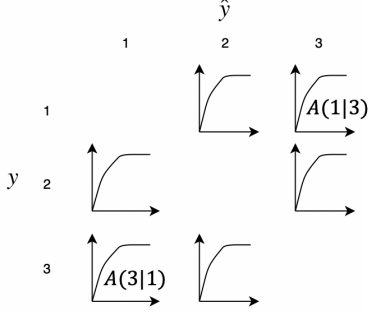


Figure 2: Multiclass pairwise choices

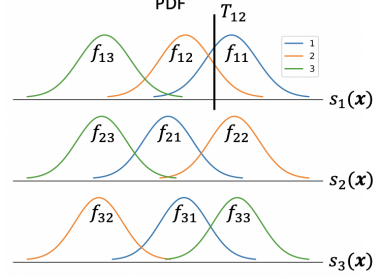


Figure 3: One possible  $f_{ij}(\cdot)$  distribution for 3-class classification

Recall we require that a classifier associates each instance with a  $c$ -dimensional score vector  $S(\mathbf{x}) = (s_1(\mathbf{x}), \dots, s_c(\mathbf{x}))^\top$ . To transform their notion of “separability” into our setting, we have:

$$A(i, j) = \frac{1}{2}(A(i|j) + A(j|i)) = \frac{1}{2}\left(P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})) + P(s_j(\mathbf{x}^{(j)}) > s_j(\mathbf{x}^{(i)}))\right) \quad (7)$$

where  $\mathbf{x}^{(i)}$  represents an instance from class  $i$ ,  $s_i(\mathbf{x}^{(i)})$  denotes the  $i^{\text{th}}$  score associated with that instance.

Further, we describe the probabilities with  $f_{ij}$  and thresholds  $T_{ij}$ :

$$P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T_{ij} > T'_{ij}) f_{ii}(T'_{ij}) f_{ij}(T_{ij}) dT'_{ij} dT_{ij} \quad (8)$$

$$P(s_j(\mathbf{x}^{(j)}) > s_j(\mathbf{x}^{(i)})) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T_{ji} > T'_{ji}) f_{jj}(T'_{ji}) f_{ji}(T_{ji}) dT'_{ji} dT_{ji} \quad (9)$$

where  $T_{ij}$  denotes the threshold on the  $i^{\text{th}}$  score axis for distinguishing instances belonging to class  $i$  and instances belonging to class  $j$ . Figure 3 generically demonstrates one possible distribution of  $f_{ij}(\cdot)$  for a 3-class classification problem.

Based on Equation 1, we derive pairwise AUC using  $f_{ij}(\cdot)$  and  $T_{ij}$ :

$$\text{AUC}_{\text{pairwise}} = \frac{2}{c(c-1)} \sum_{i < j} A(i, j) \quad (10)$$

$$= \frac{2}{c(c-1)} \sum_{i < j} \left[ \frac{1}{2} \left( P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})) + P(s_j(\mathbf{x}^{(j)}) > s_j(\mathbf{x}^{(i)})) \right) \right] \quad (11)$$

$$= \frac{2}{c(c-1)} \sum_{i < j} \frac{1}{2} \left( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T_{ij} > T'_{ij}) f_{ii}(T'_{ij}) f_{ij}(T_{ij}) dT'_{ij} dT_{ij} \right. \quad (12)$$

$$\left. + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T_{ji} > T'_{ji}) f_{jj}(T'_{ji}) f_{ji}(T_{ji}) dT'_{ji} dT_{ji} \right) \quad (13)$$

To further understand the meaning of pairwise multiclass AUC, we find that

$$\text{AUC}_{\text{pairwise}} = \frac{1}{c(c-1)} \sum_{i \neq j} P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})) \quad (14)$$

It is an arithmetic mean (unweighted) of all the  $c(c-1)$  positive-negative pairs of binary AUC and thus should be insensitive to class imbalance.

### 4.3 One-vs-Rest

Based on the discussion in Section 2.2.2, we derive the expression for  $\text{AUC}(c_i)$  within our setting.

$$\text{AUC}(c_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T' > T) f_{ii}(T') \frac{[\sum_{j \neq i} f_{ij}(T)]}{c-1} dT' dT = \frac{1}{c-1} \sum_{j \neq i} P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})) \quad (15)$$

where  $\frac{[\sum_{j \neq i} f_{ij}(T)]}{c-1}$  represents the distribution of the rest  $c-1$  classes. The summation over  $f_{ij}(\cdot)$  is normalized by  $c-1$  to maintain the property of a PDF. For one-vs-rest method, one can choose any class as positive, and the rest  $c-1$  classes as negative, which implies  $c$  pairs of binary AUC in total. Finally, we get

$$\text{AUC}_{1\text{-vs-rest}} = \sum_{i=1}^c \text{AUC}(c_i) p(c_i) \quad (16)$$

$$= \sum_{i=1}^c p(c_i) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}(T' > T) f_{ii}(T') \frac{[\sum_{j \neq i} f_{ij}(T)]}{c-1} dT' dT \quad (17)$$

$$= \sum_{i=1}^c p(c_i) \frac{1}{c-1} \sum_{j \neq i} P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})) \quad (18)$$

where  $p(c_i) = \frac{\sum_{i=1}^m \mathbb{I}(y_i = c_i)}{m}$ . Therefore, 1-vs-rest multiclass AUC is a weighted arithmetic mean of all the  $c$  pairs of binary AUC.

### 4.4 Volume Under Hyper-Surface

In this section, we present a tentative form for VUS within our setting. For binary AUC, we can transform the ROC curve by changing  $y$ -axis  $FNR$  and the minimum area is expected for optimal classifiers. Naturally extending this to multiclass, we look into the tuples of all off-diagonal entries of a confusion matrix. Given confusion matrix in table 1, we can transform the tuple  $(b, c, d, f, g, h)$  to a point in ROC space  $(\frac{b}{m_1-c}, \frac{c}{m_1-b}, \frac{d}{m_2-f}, \frac{f}{m_2-d}, \frac{g}{m_3-h}, \frac{h}{m_3-g})$  where  $m_1, m_2, m_3$  is the number of instances belonging to class 1, 2, 3 respectively. We will denote each dimension of an operating point in ROC space as false rate  $FR(\cdot)$ .

		$\hat{y}$		
		1	2	3
$y$	1	a	b	c
	2	d	e	f
	3	g	h	i

Table 1: A confusion matrix for 3-class multiclass classification.

Inspired by the form of binary AUC which we showed in section 4.1 that the function  $\mathbb{I}(T' > T)$  indicates the area in 2D space that should be under the ROC curve. Generalizing to VUS, the

volume has a similar form where a function indicating which region is below the surface and integrate this function over each axis. However, instead of integrating over a single threshold as what we did for binary case, each axis of the multiclass ROC space should, instead, be a hyper space defined by multiple thresholds. Specifically, the dimension corresponding to  $\frac{b}{m_1-c}$  has the form

$$D_{21}(\mathbf{T}^{(21)}) = \int_{FR(T_{13})=0}^{FR(T_{13})=1} \int_{FR(T_{12})=0}^{FR(T_{12})=1} \int_{FR(T_{31})=0}^{FR(T_{31})=1} \int_{FR(T_{21})=0}^{FR(T_{21})=1} \mathbb{I}(g(T_{21}, T_{31}, T_{12}, T_{13})) dFR(T_{21})dFR(T_{31})dFR(T_{12})dFR(T_{13})$$

Where  $\mathbf{T}^{(21)}$  is a set of thresholds  $\{T_{21}, T_{31}, T_{12}, T_{13}\}$ ,  $g(T_{21}, T_{31}, T_{12}, T_{13})$  inside the indicator function characterizes the space under the ROC surface for dimension  $D_{21}$ . Unlike in binary space where only one threshold defines each axis, a set of  $2c - 2$  thresholds jointly define one axis for multiclass ROC space. The thresholds that defines the entry  $b$  in the confusion matrix should be  $T_{21}, T_{31}, T_{12}, T_{13}$  which has the same meaning as we defined above. Generalizing to  $c$  class problems, the set of thresholds for the entry where  $y = i$  and  $\hat{y} = j$  is the  $2c - 2$  thresholds that differentiates class  $i$  from the rest of classes and class  $j$  from the rest of classes. To generalize to problems of  $c$  classes, each dimension can have the form

$$D_{ij}(\mathbf{T}^{(ij)}) = \underbrace{\int \dots \int_{FR_{ik}(T_{ik})=0}^{FR_{ik}(T_{ik})=1}}_{2c-2} \underbrace{\mathbb{I}(g(\{T_{ik}, T_{kj} \mid \forall k \in \mathcal{C}, i \neq k, j \neq k\}))}_{\mathbb{I}_{ij}(\mathbf{T}^{(ij)})} \underbrace{\prod_{i \neq k, j \neq k} dFR_{ik}(T_{ik})dFR_{kj}(T_{kj})}_{d\mathbf{F}_{ij}} \quad (19)$$

Using  $\mathbb{I}_{ij}(\mathbf{T}^{(ij)})$  to simplify the indicator function and  $d\mathbf{F}_{ij}$  to express the object being integrated, we derive the expression for exact VUS,

$$\text{AUC}_{\text{VUS}} = \underbrace{\int \dots \int_{2c-2} \mathbb{I}_{12}(\mathbf{T}^{(12)})}_{2c-2} \underbrace{\int \dots \int_{2c-2} \mathbb{I}_{13}(\mathbf{T}^{(13)})}_{2c-2} \dots \underbrace{\int \dots \int_{2c-2} \mathbb{I}_{c(c-1)}(\mathbf{T}^{(c(c-1))})}_{2c-2} d\mathbf{F}_{c(c-1)} \dots d\mathbf{F}_{13}d\mathbf{F}_{12} \quad (20)$$

## 4.5 Comparison

We summarize the comparison result of the three AUC multiclass extensions as a table shown in Table 2.

## 5 Experiments

We empirically explore how *class separability* and *class imbalance* affect the multiclass AUC extensions. For *class separability*, we compute both pairwise and one-vs-rest AUCs under different separability settings. For *class imbalance*, we are interested in how one-vs-rest multiclass AUC reacts under different class prior probability assumptions.

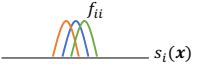
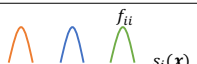
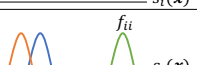
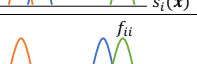
Method	Multiclass AUC	Pros	Cons
Pairwise	$\frac{1}{c(c-1)} \sum_{i \neq j} P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})), O(c^2 n \log n)$	insensitive to class distribution and error costs	hard to visualize the computed surface, not scalable
One-vs-Rest	$\sum_{i=1}^c p(c_i) \frac{1}{c-1} \sum_{j \neq i} P(s_i(\mathbf{x}^{(i)}) > s_i(\mathbf{x}^{(j)})), O(cn \log n)$	curves can be easily generated and visualized	sensitive to class distributions and error costs
VUS	(haven't figured out due to the presence of $g(\mathbf{T})$ )	ideal multiclass extension, exact number to measure the performance	multiple definitions, not scalable, hard to visualize

Table 2: Comparison of three AUC multiclass extensions.

## 5.1 Class separability

### 5.1.1 Setup

Although our experiments are set to 3-class settings, it can be easily generated to problems with more classes. For 3-class problem, we have 9 PDF  $f_{ij}(\cdot), \forall i, j \in \{1, 2, 3\}$ , assuming each of which follows normal distribution with unit variance. Only the means are varied since variance also affects separability [2]. We consider 4 cases of different class separability for 3-class problem ([o],[a],[b],[c] as explained in Figure 4(a)). There are 3 score axes for 3-class classification, each of which corresponds to any separability case. Hence, there are 20 scenarios in total as demonstrated in Figure 4(b). We assume 3 classes are evenly distributed ( $p(c_i) = \frac{1}{3}, \forall i$  for one-vs-rest AUC).

Separability Scenario	Sketch Map	Means of PDF	Separability Case	One-vs-Rest AUC	Pairwise AUC	Separability Case	One-vs-Rest AUC	Pairwise AUC
All close: case [o]		$\mu_j = 1, \mu_k = 1.25, \mu_l = 1.5$	[o][o][o]	0.6042	almost the same as one-vs-rest	[a][a][a]	0.9595	almost the same as one-vs-rest
Uniformly separable: case [a]		$\mu_j = 1, \mu_k = 3, \mu_l = 5$	[o][o][a]	0.7226		[a][a][b]	0.9698	
True class is separable with others: case [b]		$\mu_j = 1, \mu_k = 2, \mu_l = 5$	[o][o][b]	0.7329		[a][a][c]	0.9327	
			[o][o][c]	0.6958		[a][b][b]	0.9801	
True class is close to another one class: case [c]		$\mu_j = 1, \mu_k = 4, \mu_l = 5$	[o][a][a]	0.8411		[a][b][c]	0.9429	
			[o][a][b]	0.8513		[a][c][c]	0.9058	
			[o][a][c]	0.8142		[b][b][b]	0.9904	
			[o][b][b]	0.8616		[b][b][c]	0.9532	
			[o][b][c]	0.8245		[b][c][c]	0.9161	
			[o][c][c]	0.7874		[c][c][c]	0.8790	

(a)

(b)

Figure 4: (a) Separability scenarios for 3 classes; (b) AUC for each separability case. (Pairwise AUC is omitted as it is almost identical to One-vs-Rest AUC, with an average L1 distance of  $2 \times 10^{-7}$ .)

### 5.1.2 Results and discussion

Figure 4(b) shows that when data is balanced among classes, one-vs-rest and pairwise methods produce almost the same AUC values under different separability cases. We also find that, in general, AUC value decreases when more classes are close to each other (more case [o]). For separable cases, AUCs in case [b] are slightly higher than AUCs in case [a]. AUCs in case [c] are



lower than AUCs in case [a]. We have prior knowledge that case [b], case [a], case [c] and case [o] have a decreasing easiness in terms of separability and that easier separable cases should have higher AUC values. The experimental findings meet our expectation, which also implies that both one-vs-rest and pairwise AUC values are reasonable as a good metric.

## 5.2 Class imbalance for one-vs-rest

In this section, we explore how class imbalance affects the one-vs-rest AUC.

### 5.2.1 Setup

Considering 3-class classification, for some PDF  $f_{ij}$ 's, we vary  $p(c_i)$  for one-vs-rest method. We choose entropy as the measurement of class imbalance. Higher entropy means lower extent of class imbalance. Without loss of generality, we fix  $p(c_1) = \frac{1}{3}$ , and vary  $p(c_2)$  and  $p(c_3)$  to achieve different entropy values. The entropy is calculated as  $H = -\sum_{i=1}^3 p(c_i) \log_3 p(c_i)$ . Although pairwise AUC is unchanged under different class distributions, we compute its value as reference.

We consider two separability cases in this experiment. One situation (Situation A) is where the 3 PDFs on each score axis obey the same distributions, and the other situation (Situation B) is when they are different. We choose one representative distribution for each situation:

- Situation A:  $\forall i, j \ f_{ij} \sim N(\mu_{ij}, 1)$ .  $\mu_{11} = 5, \mu_{12} = 3, \mu_{13} = 1, \mu_{21} = 1, \mu_{22} = 5, \mu_{23} = 3, \mu_{31} = 1, \mu_{32} = 3, \mu_{33} = 5$
- Situation B:  $\forall i, j \ f_{ij} \sim N(\mu_{ij}, 1)$ .  $\mu_{11} = 1.5, \mu_{12} = 1.25, \mu_{13} = 1, \mu_{21} = 1, \mu_{22} = 5, \mu_{23} = 3, \mu_{31} = 1, \mu_{32} = 4, \mu_{33} = 5$

### 5.2.2 Results and discussion

Figure 5 and 6 imply the following:

- (1) As the data becomes less imbalanced, one-vs-rest AUC approaches the value of pairwise AUC.
- (2) The value of one-vs-rest AUC can be greater, equal or less than pairwise AUC. When the hard-separable case is more often, such as score axis 3 in Figure 6(a), the one-vs-rest AUC will be lower than pairwise AUC. If the easier-separable case is more prevalent, one-vs-rest AUC will be greater than pairwise AUC.
- (3) When all PDFs for each score axis are distributed in the same way, one-vs-rest is insensitive to class imbalance and have the exactly same value as pairwise AUC.

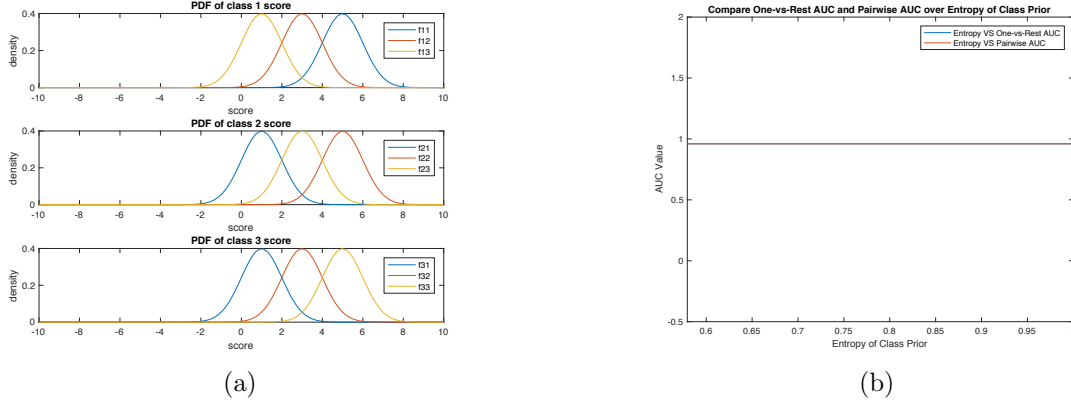


Figure 5: One-VS-Rest AUC and Pairwise AUC plotted over the entropy of the probability distribution of the class, with the same separability setting among classes (Situation A).

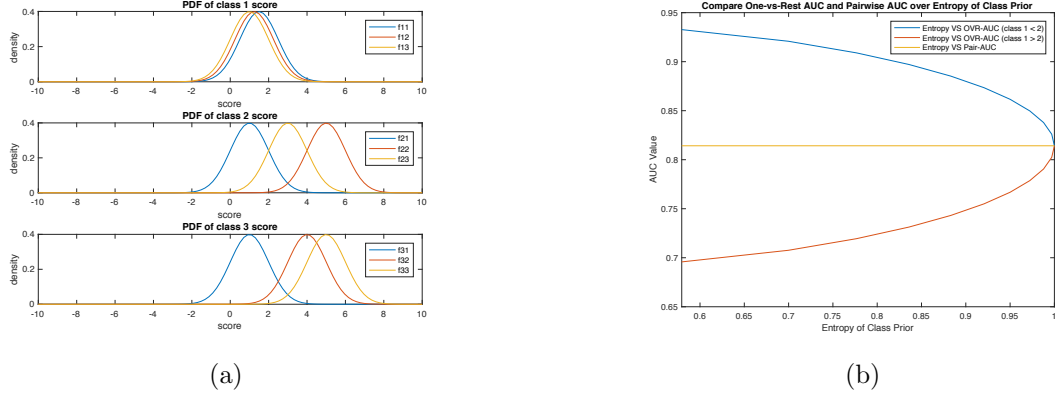


Figure 6: One-VS-Rest AUC and Pairwise AUC plotted over the entropy of the probability distribution of the class, with different separability setting among classes (Situation B).

## 6 Conclusion

**Summary:** We explored the relationships and differences among 3 kinds of multiclass extensions of AUC/ROC (pairwise, one-vs-rest and VUS) theoretically and empirically. We unified binary AUC and the aforementioned extensions using score density function and thresholds. We empirically showed the relationship between pairwise and one-vs-rest AUC and how they are affected by class separability and class imbalance.

**Future work:** One limitation of our work is that the threshold  $T_{ij}$  can only be useful when the density function are given. It is necessary to focus on relating such thresholds to actual multiclass classifiers in the future. Successfully relating the thresholds to real classifiers will enable us to derive the correct form of VUS. What we focused in our experiments is the numerical relationship between pairwise and one-vs-rest AUC. When the number of classes is large, it is also necessary to compare the running time of different multiclass extensions.

## References

- [1] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [2] David J. Hand and Robert J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.
- [3] Thomas Landgrebe and Robert P. W. Duin. A simplified extension of the area under the roc to the multiclass domain. 2006.
- [4] Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52(3):199–215, Sep 2003.
- [5] Amay S.M. Cheam and Paul McNicholas. Modelling receiver operating characteristic curves using gaussian mixtures. *Computational Statistics Data Analysis*, 93, 06 2014.