Understanding Multiclass Extensions of ROC/AUC and their Relationships

Bowen Ke, Hui Lyu, Simeng Sun

{bowenke, huilyu, simsun}@seas.upenn.edu
University of Pennsylvania

November 29th, 2018

Introduction

Objective

To understand relationships between ROC/AUC multiclass extension methods and their pros and cons

Setup

- General multiclass classification (not restricted to ordinal multiclass problems)
- Each instance gets a c-dimensional score vector $S(\mathbf{x}) = (s_1(\mathbf{x}), \ldots, s_c(\mathbf{x}))^\top$ and is classified to $\hat{y} = \mathrm{argmax}_i s_i(\mathbf{x}), \mathcal{C} = \{1, \ldots, c\}$
- Denote $f_{ij}(t)$ as the probability density function of the i^{th} entry of $S(\mathbf{x})$ for every instance $\mathbf{x}$ of class j, $i, j \in \mathcal{C}$

Our contributions

- Unify binary AUC and multiclass extensions with $f_{ij}(t)$
- Find relationship between binary AUC and multiclass extensions
- Compare multiclass extensions theoretically and empirically

Background

Revisit binary ROC/AUC:

$f : \mathcal{X} \to \mathbb{R}$

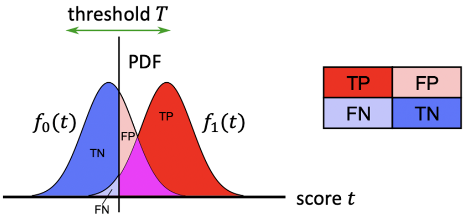
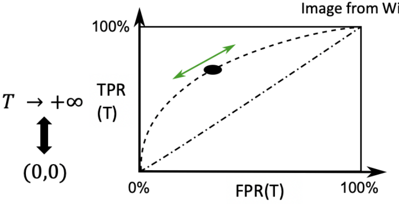


Image from Wikipedia



$$TPR(T) = \int_T^{\infty} f_1(t)dt \qquad FPR(T) = \int_T^{\infty} f_0(t)dt$$

Multiclass AUC extensions:

1. **Pairwise:** $\#\binom{c}{2}$ binary AUC
   $AUC_{pairwise} = \frac{2}{c(c-1)} \sum_{i<j} A(i,j)$
   $A(i,j) = \frac{1}{2}(A(i|j) + A(j|i))$
2. **One-vs-Rest:** $\#c$ binary AUC
   $AUC_{1-vs-rest} = \sum_{c_i \in C} AUC(c_i)p(c_i)$
   $p(c_i) = \frac{\sum_{y_i} I(y_i=c_i)}{m}$
3. **VUS** (Volume Under Hyper-Surface):
   Off-diagonal entries of a confusion matrix corresponds to a single operating point in $c(c-1)$ -dimensional space

Challenge:

- Each with different settings
- Either lack of statistical explanation or relationship with binary AUC

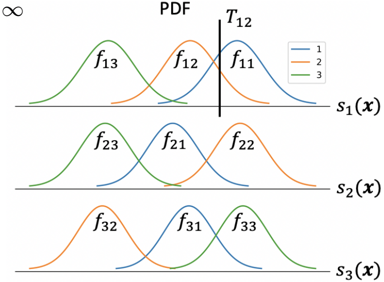
Our Proof and Derivation for Binary AUC and Pairwise AUC

Binary AUC $(= A(1|0))$:

$$\text{AUC} = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{I}(T' > T)f_1(T')f_0(T)dT'dT = P(X_1 > X_0)$$

Pairwise: $A(i,j) = \frac{1}{2}(A(i|j) + A(j|i)) = \frac{1}{2}\Big(P\big(s_i(\mathbf{x^i}) > s_i(\mathbf{x^j})\big) + P\big(s_j(\mathbf{x^j}) > s_j(\mathbf{x^i})\big)\Big)$

$$\text{AUC} = \frac{2}{c(c-1)}\sum_{i<j} A(i,j) = \frac{2}{c(c-1)}\sum_{i<j}\frac{1}{2}\Big(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbf{I}(T_{ij} > T_{ij}`)f_{ii}(T_{ij}`)f_{ij}(T_{ij})dT_{ij}`dT_{ij}$$
$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbf{I}(T_{ji} > T_{ji}`)f_{jj}(T_{ji}`)f_{ji}(T_{ji})dT_{ji}`dT_{ji}\Big)$$

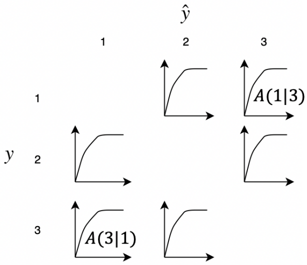


T_{ij} is the threshold to classify label i and label j

$\therefore AUC = \frac{1}{c(c-1)}\sum_{i\neq j} P_i(X_i > X_j)$

— arithmetic mean (equal weight)

Our Derivation for One-vs-Rest and VUS AUC



Pairwise

One-vs-Rest:

$$AUC = \sum_{c_i \in C} AUC(c_i)p(c_i), \; p(c_i) = \frac{\sum_{y_i} \mathbb{I}(y_i = c_i)}{m}$$

$$AUC(c_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{I}(T` > T) f_{ii}(T`) \frac{[\sum_{j \neq i} f_{ij}(T)]}{c-1} dT`dT$$

$$= \frac{1}{c-1} \sum_{j \neq i} P_i(X_i > X_j)$$

$$\therefore AUC = \sum_i p(c_i) \frac{1}{c-1} \sum_{j \neq i} P_i(X_i > X_j)$$ — weighted arithmetic mean

VUS: (Tentative derivation)

Consider FNR (1-TPR) and FPR for binary; Extend to multiclass: c(c-1) error rates

Each confusion matrix (controlled by T_{ij}) corresponds to a point in $\mathbb{R}^{c(c-1)}$

Each dimension $D_{ij}(\mathbf{T}) =$

$$\underbrace{\int \cdots \int_{-\infty}^{\infty}}_{2c-2} \underbrace{\mathbf{I}(g(\{T_{ik}, T_{kj} \mid \forall k \in \mathcal{C}, i \neq k, j \neq k\}))}_{\mathbf{I}_{ij}} \underbrace{\prod_{i \neq k, j \neq k} dFR_{ik}(T_{ik}) dFR_{kj}(T_{kj})}_{dF_{ij}}$$

$$VUS = \int \ldots \int_{-\infty}^{\infty} \mathbf{I}_{12} \int \ldots \int_{-\infty}^{\infty} \mathbf{I}_{13} \cdots \int \ldots \int_{-\infty}^{\infty} \mathbf{I}_{c(c-1)} dF_{c(c-1)} \ldots dF_{13} dF_{12}$$

Results

Theoretical:

Method	Multiclass AUC	Pros	Cons
Pairwise	$\frac{1}{c(c-1)}\sum_{i\neq j} P_i(X_i > X_j)$ $O(C^2 n \log n)$	insensitive to class distribution and error costs	hard to visualize the computed surface, not scalable
One-vs-Rest	$\frac{1}{c-1}\sum_{i} p(c_i)\sum_{i\neq j} P_i(X_i > X_j)$, $O(Cn \log n)$	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

Empirical:

- Measure how class separability affects pairwise & one-vs-rest multiclass AUC
- Measure how class imbalance affects one-vs-rest multiclass AUC

Discussion

Experimental Result

- Pairwise and one-vs-rest AUC perform similarly for multiple class separability combinations when data is balanced.
- Class imbalance does not affect one-vs-rest AUC if f_{ij}s are distributed in a same way in each score axis.
- Please see our report for more detailed experimental results.

Strengths:

- Derive mathematical expressions

Limitations:

- Purely theoretical derivation
- PDF is hard to estimate

Next steps:

- Find the correct representation of VUS in terms of f_{ij} and T_{ij}
- Explore consistent algorithms for these methods

References

- Tom Fawcett. An introduction to ROC analysis.Pattern Recognition Letters, 27(8):861–874, 2006.
- Peter Flach. The many faces of ROC analysis in machine learning.ICML Tutorial, 2004.
- David J. Hand and Robert J. Till. A simple generalisation of the area under the ROC curve for multipleclass classification problems.Machine Learning, 45(2):171–186, 2001.
- Terran Lane. Position paper: Extensions of ROC analysis to multi-class domains. InProceedings ofICML workshop on cost-sensitive learning, 2000.
- Ashwin Srinivasan. Note on the location of optimal classifiers in n-dimensional ROC space. Technical Report, 1999.