Confidence Intervals
For Multiclass F1 Scores

Adam Vinestock
Idan Cohen

Based on the paper

link

Before we start....

- Introduction
- Time plan
- Q & A

Table of Contents

01

Motivation

Why F1 is useful

- **Given data D - we want to measure the performance of different classification models** $f_1, f_2, f_3 \dots$

We want a single metric - comparing apples to apples

We want to account for both precision and recall

We want to account for imbalanced classes

F1 elegantly sums up the predictive performance of a model by combining two otherwise competing metrics — precision and recall

Comparing point estimates

- **Neglect sampling variability** - overlook the uncertainty of model performance, this can result in unreliable assessments
- **Incomplete picture of the true performance** - without considering confidence intervals, point estimates offer limited practical utility
- **Example** - consider the results of an analysis reported by Dong et al

We need to calculate variance!

Confidence Intervals for F1

- **Binary case** - some statistical methods have been proposed to compute variance estimates for F1 scores (Bootstrapping, binomial distribution assumption)

- **Multiclass case** - this paper addresses the knowledge gap, providing the methods for computing variances of these multiclass F1 scores

02

The Key Components

Performance measures

- **Precision - positive predictive value**

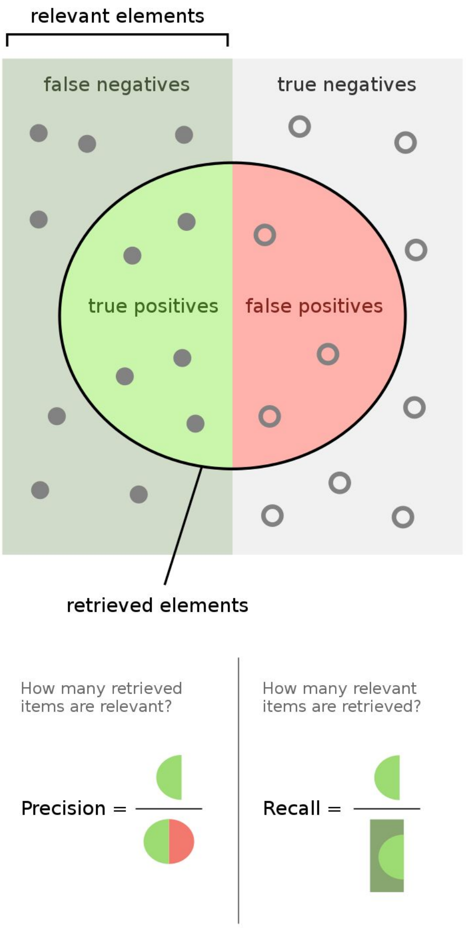
  We want high precision for SPAM detection, predicting bank E-mails as spam is bad

- **Recall - sensitivity**

  We want high recall for cancer detection, better to over predict positives then miss a few cases

- **F1 - the harmonic mean of precision & recall**

$$F_1 = 2 \times \frac{P * R}{P + R}$$



Performance measures

- **How do we evaluate multi-class classification models?**
- **Calculate F1 score for each class** F_{1i}
- **Then use average F1, micro F1 or macro F1**

Notations

- **Given a multiclass classification task, with $r > 2$ classes**
- **We can compute the confusion matrix using the observed data**
- **We denote cell probabilities p_{ij} and marginal probabilities $p_{i\cdot}, p_{\cdot i}$**

Frequencies

		True Classification			
		Class1	Class 2	Class 3	
Prediction	Class 1	2	2	2	
	Class 2	5	70	2	
	Class 3	0	2	15	

Proportions

Prediction	Class 1	0.02	0.02	0.02	$p_{1\cdot} = 0.06$
	Class 2	0.05	0.7	0.02	0.77
	Class 3	0	0.02	0.15	0.17
		$p_{\cdot 1} = 0.07$	0.74	0.19	

$p_{11} = 0.02$

$p_{12} = 0.02$

$p_{13} = 0.02$

...

$p_{33} = 0.15$

$$\sum_{i,j} p_{ij} = 1$$

Performance measures

- **Micro-averaged F1 - harmonic mean of** m_iP **,** m_iR

Micro P

$$m_iP = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r}(TP_i + FP_i)} = \frac{\sum_{i=1}^{r} p_{ii}}{\sum_{i=1}^{r} p \cdot i} = \sum_{i=1}^{r} p_{ii}$$

Micro R

$$m_iR = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r}(TP_i + FP_i)} = \frac{\sum_{i=1}^{r} p_{ii}}{\sum_{i=1}^{r} p_{\cdot i}} = \sum_{i=1}^{r} p_{ii}$$

Micro F1

$$m_iF_1 = 2 \times \frac{m_iP \times m_iR}{m_iP + m_iR} = \sum_{i=1}^{r} p_{ii}$$

Performance measures

- **Macro-averaged F1**

First calculate F1 per class

$$F_{1i} = 2 \times \frac{P_i \times R_i}{P_i + R_i} = 2 \times \frac{p_{ii}}{p_{i\cdot} + p_{\cdot i}}$$

The macro-averaged F1 score is defined as the simple arithmetic mean of F_{1i}

$$maF_1 = \frac{1}{r}\sum_{i=1}^{r} F_{1i} = \frac{2}{r}\sum_{i=1}^{r} \frac{p_{ii}}{p_{i\cdot} + p_{\cdot i}}$$

Reminder - MLE, CLT & CI

- **Maximum Likelihood Estimation (MLE)**

  Finds the distribution parameter values that maximize the likelihood given observed data

- **Central Limit Theorem (CLT)**

  The average of many independent observations, regardless of the underlying distribution approximates to a normal distribution

- **Confidence Intervals (CI)**

  Defines a plausible range of values for the true parameter, considering the uncertainty from sampling, based on the observed sample

03

Confidence Intervals For F1

Variance Estimation

- **One can observe the confusion matrix as a random variable with multinomial distribution with sample size n and probabilities**

$$p = (p_{11}, \ldots, p_{1r}, p_{21}, \ldots, p_{2r}, \ldots, p_{r1}, \ldots, p_{rr})^T$$

- **That is**

$$(n_{11}, n_{12}, \ldots, n_{rr}) \sim Multinomial(n; \mathbf{p})$$

- **And using MLE** $\hat{p}_{ij} = \frac{n_{ij}}{n}$

		True Classification		
		Class1	Class 2	Class 3
Prediction	Class 1	2	2	2
	Class 2	5	70	2
	Class 3	0	2	15
		0.07	0.74	0.19

Variance Estimation - Cont.

- **Multivariate central limit theorem for multinomial:**

$$\sqrt{n}\,(\hat{p} - p) \dot{\sim} Normal\left(0, diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right)$$

- **Using this, along with multivariate delta-method we will get that** $\widehat{maF_1}$ **and** $\widehat{miF_1}$ **are approximately normally distributed as:**

$$\widehat{maF_1} \dot{\sim} \mathrm{Normal}(maF_1, Var(\widehat{maF1}))$$

$$\widehat{miF_1} \dot{\sim} \mathrm{Normal}(miF_1, Var(\widehat{miF1}))$$

		True class			
		Class 1	Class 2	Class 3	F1
Prediction	Class 1	0.02	0.02	0.02	0.308
	Class 2	0.05	0.7	0.02	0.927
	Class 3	0	0.02	0.15	0.833

Multivariate Delta method

- **Given random variables** X_n **satisfying** $\sqrt{n}[X_n - \theta] \xrightarrow{D} \mathcal{N}(0, \sigma^2)$
- **For any function g satisfying the property that g'(θ) exists and is non-zero valued we get**

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{D} \mathcal{N}(0, \sigma^2 \cdot [g'(\theta)]^2)$$

- **In our case:**

$$\sqrt{n}\left(\hat{p} - p\right) \dot{\sim} Normal\left(0, diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right)$$

$$\downarrow$$

$$\sqrt{n}\left(\widehat{miF_1} - miF_1\right)$$
$$\dot{\sim} Normal\left(0, \left[\frac{\partial\left(miF_1\right)}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \left[\frac{\partial\left(miF_1\right)}{\partial(\boldsymbol{p})}\right]\right)$$

$$\downarrow$$

$$\widehat{maF_1} \dot{\sim} \mathrm{Normal}(maF_1, Var(\widehat{maF1}))$$

Variance Estimation - Cont.

- **While,**

$$Var\left(\widehat{miF_1}\right) = \left(\sum_{i=1}^{r} p_{ii}\right)\left(1 - \sum_{i=1}^{r} p_{ii}\right)/n$$

$$Var\left(\widehat{maF_1}\right) = \frac{2}{r^2}\left\{\sum_{i=1}^{r} \frac{F_{1i}\left(p_{i\cdot} + p_{\cdot i} - 2p_{ii}\right)}{\left(p_{i\cdot} + p_{\cdot i}\right)^2}\left(\frac{p_{i\cdot} + p_{\cdot i} - 2p_{ii}}{p_{i\cdot} + p_{\cdot i}} + \frac{F_{1i}}{2}\right) + \sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ij}F_{1i}F_{1j}}{\left(p_{i\cdot} + p_{\cdot i}\right)\left(p_{j\cdot} + p_{\cdot j}\right)}\right\}/n$$

- **Now we have all the components - Normally dist & we can calc variance**
- **We can plug-in to compute confidence intervals for** miF_1 **,** maF_1

Confidence Interval Calculation - example

- **Confidence Interval of miF_1 :**

$$\widehat{miF_1} \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{\mathrm{Var}(\widehat{miF_1})}$$

$$\widehat{miF_1} = 2\frac{m_iP \times m_iR}{m_iP + m_iR} = \sum_{i=1}^{r} p_{ii} = 0.87$$

$$Var(\widehat{miF1}) = \frac{(0.02+0.7+0.15)(1-(0.02+0.7+0.15)}{100} = 0.00113$$

⇒ 95% confidence interval of miF_1 :

$$0.87 \pm 1.960 \times \sqrt{0.00113} = (0.804, 0.936)$$

		True class		
		Class 1	Class 2	Class 3
Prediction	Class 1	0.02	0.02	0.02
	Class 2	0.05	0.7	0.02
	Class 3	0	0.02	0.15

Confidence Interval Calculation - example

- **Confidence Interval of maF_1 :**

$$\widehat{maF_1} \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{\mathrm{Var}(\widehat{maF_1})}$$

$$\widehat{maF_1} = \frac{1}{r}\sum_{i=1}^{r} F_{1i} = (0.308 + 0.927 + 0.833)/3 = 0.689$$

$$Var(\widehat{maF1}) = 0.0650^2$$

⇒ 95% confidence interval of maF_1 :

$$0.69 \pm 1.960 \times 0.0650 = (0.562, 0.817)$$

		True class			
		Class 1	Class 2	Class 3	F1
Prediction	Class 1	0.02	0.02	0.02	0.308
	Class 2	0.05	0.7	0.02	0.927
	Class 3	0	0.02	0.15	0.833

04

Conclusion

Conclusion

Single Metric for comparison

A metric that takes into account both precision and recall, offering a more holistic measure of model effectiveness

More Robust & informed decision-making

This statistical approach enables a more accurate and objective assessment of model performance

Questions?

Thank you!

Appendix

Appendix A: Derivation of the distribution and variance of $\widehat{miF_1}$

Go to: ▸

Let $\boldsymbol{p}$ be the ordered elements of a confusion matrix. $\boldsymbol{p} = (p_{11}, \cdots, p_{1r}, p_{21}, \cdots, p_{2r}, \cdots, p_{r1}, \cdots, p_{rr})^T$. Using the multivariate delta-method for $\widehat{p}$, we get

$$\sqrt{n}\left(\widehat{miF_1} - miF_1\right)$$

$$\dot{\sim} Normal\left(0, \left[\frac{\partial\left(miF_1\right)}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial\left(miF_1\right)}{\partial(\boldsymbol{p})}\right]\right). \quad (5)$$