

RECAP: PERFORMANCE MEASURES FOR BINARY CLASSIFICATION

Advanced Machine Learning

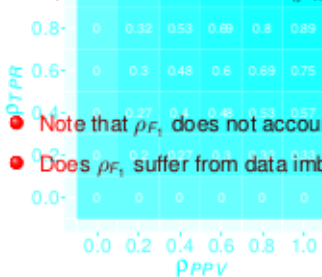
- We encourage readers to first go through [Chapter 10.8 in I2ML](#).
- In binary classification ($\mathcal{Y} = \{-1, +1\}$):

Imbalanced Learning: Performance Measures

		True Class y		
		+	-	
Classification \hat{y}	+	TP	FP	$\rho_{PPV} = \frac{\#TP}{\#TP + \#FP}$
	-	FN	TN	$\rho_{PNV} = \frac{\#TN}{\#FN + \#TN}$
		$\rho_{TPR} = \frac{\#TP}{\#TP + \#FN}$	$\rho_{TNR} = \frac{\#TN}{\#FP + \#TN}$	$\rho_{ACC} = \frac{\#TP + \#TN}{TOTAL}$



- F_1 score balances Recall (ρ_{TPR}) and Precision (ρ_{PPV}):



Learning goals

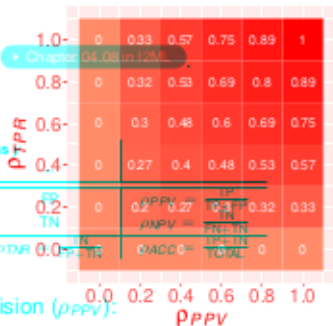
- $\rho_{F_1} = 2 \frac{\rho_{PPV} \cdot \rho_{TPR}}{\rho_{PPV} + \rho_{TPR}}$ Know performance measures beyond accuracy
- Note that ρ_{F_1} does not account for TN. Know their advantages over accuracy for imbalanced data
- Does ρ_{F_1} suffer from data imbalance like accuracy does?
 - Know extensions of these measures for multiclass settings

F₁ SCORE IN BINARY CLASSIFICATION

- We encourage readers to first go through [Chapter 04.08 in 12M](#)
- In binary classification ($\mathcal{Y} = \{-1, +1\}$):

F_1 is the **harmonic mean** of ρ_{PPV} & ρ_{TPR} .
 → Property of harmonic mean: tends more towards the **lower** of two combined values.

\hat{y}	-	+
True Class	FN	TP
	$\rho_{TPR} = \frac{TP}{TP+FN}$	$\rho_{PPV} = \frac{TP}{TP+FP}$



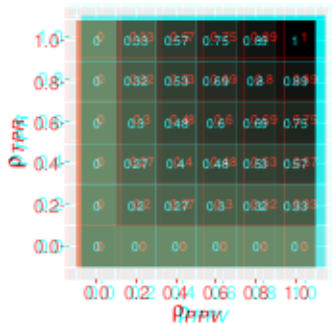
- F_1 score balances Recall (ρ_{TPR}) and Precision (ρ_{PPV}):
- A model with $\rho_{TPR} = 0$ or $\rho_{PPV} = 0$ has $\rho_{F_1} = 0$.
 $\rho_{F_1} = 2 \cdot \frac{\rho_{PPV} \cdot \rho_{TPR}}{\rho_{PPV} + \rho_{TPR}}$
- Always predicting "negative": $\rho_{TPR} = \rho_{F_1} = 0$
- Always predicting "positive": $\rho_{PPV} = \rho_{F_1} = 0$
- Does ρ_{F_1} suffer from data imbalance like accuracy does?
 → small when $n_+ (= TP + FN = TP)$ is small.
- Hence, F_1 score is more robust to data imbalance than accuracy.

F_β IN BINARY CLASSIFICATION

- F_1 puts equal weights to $\frac{1}{\rho_{PPV}}$ & $\frac{1}{\rho_{TPR}}$ because $F_1 = \frac{2}{\frac{1}{\rho_{PPV}} + \frac{1}{\rho_{TPR}}}$.

F_β puts β^2 times as weight to $\frac{1}{\rho_{TPR}}$.
 → Property of harmonic mean: tends more towards the lower of two combined values.

$$F_\beta = \frac{\beta^2}{\frac{\beta^2}{1+\beta^2} \cdot \frac{1}{\rho_{TPR}} + \frac{1}{1+\beta^2} \cdot \frac{1}{\rho_{PPV}}} = (1 + \beta^2) \cdot \frac{\rho_{PPV} \cdot \rho_{TPR}}{\beta^2 \rho_{PPV} + \rho_{TPR}}$$



- $\beta > 1$ model with $F_\beta \approx \rho_{TPR}$; $\rho_{PPV} = 0$ has $\rho_{F_1} = 0$.
- Always predicting "negative": $\rho_{TPR} = \rho_{F_1} = 0$
- Always predicting "positive":
 $\rho_{TPR} = 1 \Rightarrow \rho_{F_1} = 2 \cdot \rho_{PPV} / (\rho_{PPV} + 1) = 2 \cdot n_+ / (n_+ + n)$,
 \rightsquigarrow small when $n_+ (= TP + FN = TP)$ is small.
- Hence, F_1 score is more robust to data imbalance than accuracy.

G SCORE AND G MEAN

- G score uses geometric mean: $\rho_G = \frac{1}{2} \left(\rho_{PPV} + \frac{1}{\rho_{TPR}} \right)$
- ρ_G puts equal weights to ρ_{PPV} & $\frac{1}{\rho_{TPR}}$

because $\rho_G = \sqrt{\rho_{PPV} \cdot \frac{1}{\rho_{TPR}}}$

- F_2 puts β^2 times of weight to $\frac{1}{\rho_{TPR}}$:
towards the **lower** of the two combined values.

$$\rho_{F_2} = \frac{1}{\beta^2 \rho_{PPV} + \rho_{TPR}}$$

- Geometric mean is **larger** than harmonic mean: $\rho_G = \sqrt{\rho_{PPV} \cdot \frac{1}{\rho_{TPR}}}$

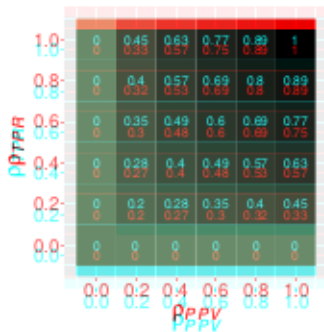
- Closely related is the G mean:

- $\beta \gg 1 \rightsquigarrow F_\beta \approx \frac{1}{\rho_{TPR}}$

- $\beta \ll 1 \rightsquigarrow F_\beta \approx \rho_{PPV}$ $\rho_{Gm} = \sqrt{\rho_{TNR} \cdot \rho_{TPR}}$

It also considers TN.

- Always predicting "negative": $\rho_G = \rho_{Gm} = 0 \rightsquigarrow$ Robust to data imbalance!



BALANCED ACCURACY

- G score uses geometric mean:

- Balanced accuracy (BAC) balances

- ρ_{TNR} and ρ_{TPR} :

Geometric mean tends more towards the lower of the two combined values.

$$\rho_{BAC} = \frac{\rho_{TNR} + \rho_{TPR}}{2}$$

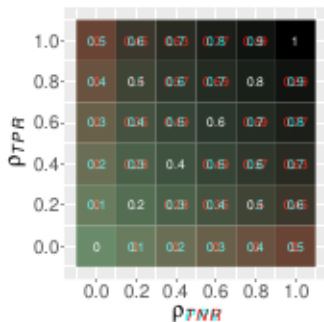
- Geometric mean is larger than harmonic mean.

- If a classifier attains high accuracy on both classes or the data set is almost balanced, then $\rho_{BAC} \approx \rho_{ACC}$.

$$\rho_{Gm} = \sqrt{\rho_{TNR} \cdot \rho_{TPR}}$$

It also considers TN.

- However, if a classifier always predicts "negative" for an imbalanced data set, i.e. $n_+ \ll n_-$, then $\rho_{BAC} \ll \rho_{ACC}$. It also considers TN.
- Always predicting "negative": $\rho_{Gm} = \rho_{Gm} = 0 \rightarrow$ Robust to data imbalance!



MATTHEWS CORRELATION COEFFICIENT

- Recall: Pearson correlation coefficient (PCC):

- Balanced accuracy (BAC) balances ρ_{TNR} and ρ_{TPR} :

$$\rho_{BAC} = \frac{\rho_{TNR} + \rho_{TPR}}{2}$$

- View "predicted" and "true" classes as two binary random variables.

- Using entries in confusion matrix to estimate the PCC, we obtain MCC:

$$\rho_{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

- If a classifier attains high accuracy on both classes or the data set is almost

- In contrast to other metrics:

- MCC uses all entries of the confusion matrix;
- MCC has value in $[-1, 1]$.
- However, if a classifier always predicts "negative" for an imbalanced data set, i.e. $n_+ \ll n_-$, then $\rho_{BAC} \ll \rho_{MCC}$. It also considers TN.

0.5	0.6	0.7	0.8	0.9	1
0.8	0.4	0.5	0.6	0.7	0.8
0.6	0.3	0.4	0.5	0.6	0.7
0.4	0.2	0.3	0.4	0.5	0.6
0.2	0.1	0.2	0.3	0.4	0.5
0.0	0.0	0.1	0.2	0.3	0.4



MATTHEWS CORRELATION COEFFICIENT

- Recall: Pearson correlation coefficient (PCC)

$$r_{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + EN)(TN + FP)}}$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- $r_{MCC} \approx 1$ \rightsquigarrow nearly zero error \rightsquigarrow good classification, i.e., strong correlation between predicted and true classes.
- View "predicted" and "true" classes as two binary random variables.
- Using entries in confusion matrix to estimate the PCC, we obtain MCC:
- $r_{MCC} \approx 0$ \rightsquigarrow no correlation, i.e., not better than random guessing.

$$r_{MCC} \approx \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + EN)(TN + FP)}}$$

- In contrast to other metrics:
 - Previous measures requires defining positive class. But MCC does not depend on which class is the positive one;
 - MCC has value in $[-1, 1]$.



MULTICLASS CLASSIFICATION EFFICIENT



Classification \hat{y}	True Class y			
	1	2	...	g
1	n_{11} (True 1's)	n_{12} (False 1's for 2's)	...	n_{1g} (False 1's for g 's)
2	n_{21} (False 2's for 1's)	n_{22} (True 2's)	...	n_{2g} (False 2's for g 's)
...
g	n_{g1} (False g 's for 1's)	n_{g2} (False g 's for 2's)	...	n_{gg} (True g 's)

- $\rho_{MCC} \approx 1 \rightsquigarrow$ nearly zero error \rightsquigarrow good classification, i.e., strong correlation between predicted and true classes.
- $\rho_{MCC} \approx 0 \rightsquigarrow$ no correlation, i.e., not better than random guessing.
- n_{ji} : the number of i instances classified as j .
- $n_i = \sum_{j=1}^g n_{ji}$ the total number of i instances.
- Class-specific metrics:**
 - Previous measures requires defining positive class. But MCC does not depend on which class is the positive one.
 - True positive rate (**Recall**): $\rho_{TPR_i} = \frac{n_i}{n}$
 - True negative rate $\rho_{TNR_i} = \frac{\sum_{j \neq i} n_j}{n - n_i}$
 - Positive predictive value (**Precision**) $\rho_{PPV_i} = \frac{n_i}{\sum_{j=1}^g n_{ji}}$

MACRO F1 SCORE SENSITIZATION

- Average over classes to obtain a single value:

Classification	True Class y			
	1	2	...	g
1	n_{11} (True 1's)	n_{12} (False 1's for 2's)	...	n_{1g} (False 1's for g 's)
2	n_{21} (False 2's for 1's)	n_{22} (True 2's)	...	n_{2g} (False 2's for g 's)
...
g	n_{g1} (False g 's for 1's)	n_{g2} (False g 's for 2's)	...	n_{gg} (True g 's)

$$\rho_{mMETRIC} = \frac{1}{g} \sum_{i=1}^g \rho_{METRIC_i}$$

where $METRIC_i$ is a class-specific metric such as PPV_i , TPR_i of class i .

- With this, one can simply define a **macro F_1** score:

- n_{ji} : the number of i instances classified as j .

- $n_i = \sum_{j=1}^g n_{ji}$ the total number of i instances

$$\rho_{mF_1} = 2 \cdot \frac{\rho_{mPPV} \cdot \rho_{mTPR}}{\rho_{mPPV} + \rho_{mTPR}}$$

- Class-specific** metrics:

- Problem: each class equally weighted \rightsquigarrow class sizes are not considered.

- True positive rate (**Recall**): $\rho_{TPR_i} = \frac{n_{i1}}{n_i}$

- How about applying different weights to the class-specific metrics?

- True negative rate $\rho_{TNR_i} = \frac{\sum_{j=1}^g n_{ji} - n_{i1}}{n - n_i}$

- Positive predictive value (**Precision**) $\rho_{PPV_i} = \frac{n_{i1}}{\sum_{j=1}^g n_{j1}}$



WEIGHTED MACRO F_1 SCORE

- For imbalanced datasets, give **more weights** to **minority** classes.

- $w_1, \dots, w_g \in [0, 1]$ such that $w_i > w_j$ iff $n_i < n_j$ and $\sum_{i=1}^g w_i = 1$.

$$\rho_{mMETRIC} = \frac{1}{g} \sum_{i=1}^g \rho_{METRIC_i}$$

$$\rho_{wmMETRIC} = \frac{1}{g} \sum_{i=1}^g \rho_{METRIC_i} w_i$$

where $METRIC_i$ is a class-specific metric such as PPV_i , TPR_i of class i .

- where $METRIC_i$ is a class-specific metric such as PPV_i , TPR_i of class i .

- Example: $w_i = \frac{n - n_i}{(g-1)n}$ are suitable weights.

- Weighted macro F_1 score: $\rho_{F_1} = 2 \cdot \frac{\rho_{mPPV} \cdot \rho_{mTPR}}{\rho_{mPPV} + \rho_{mTPR}}$

- Problem: each class equally weighted \rightarrow class sizes are not considered.

- How about applying different weights to the class-specific metrics?

- This idea gives rise to a weighted macro G score or weighted BAC.

- Usually**, weighted F_1 score uses $w_i = n_i/n$. However, for imbalanced data sets this would **overweight** majority classes.



OTHER PERFORMANCE MEASURES

- "Micro" versions, e.g. the micro TPR is $\frac{\sum_{i=1}^g TP_i}{\sum_{i=1}^g TP_i + FN_i}$. For imbalanced data sets, give more weights to minority classes.
- $w_1, \dots, w_g \in [0, 1]$ such that $w_i > w_j$ iff $n_i < n_j$ and $\sum_{i=1}^g w_i = 1$.
- MCC can be extended to:

$$PMCC = \frac{\rho_{wmmetric} n \sum_{i=1}^g \frac{1}{n_i} n_i - \rho_{METRIC_i} \hat{n}_i n_i}{\sqrt{(n^2 - \sum_{i=1}^g \hat{n}_i^2)(n^2 - \sum_{i=1}^g n_i^2)}}$$

where $METRIC_i$ is a class-specific metric such as PPV_i , TPR_i of class i .

- where $\hat{n}_i = w \sum_{j=1}^g \frac{n_j n_i}{(g-1)n}$ is the total number of instances classified as i .
- Weighted macro F_1 score:
- Cohen's Kappa or Cross Entropy (see Grandini et al. (2021)) treat "predicted" and "true" classes as two discrete random variables.

$$\rho_{wmmF_1} = \frac{\rho_{wmmPPV} + \rho_{wmmTPR}}{\rho_{wmmPPV} + \rho_{wmmTPR}}$$
- This idea gives rise to a weighted macro G score or weighted BAC.
- Usually, weighted F_1 score uses $w_i = n_i/n$. However, for imbalanced data sets this would **overweight** majority classes.



WHICH PERFORMANCE MEASURE TO USE?

- Since different measures focus on other characteristics \rightsquigarrow No golden answer to this question.
 $\frac{\sum_{i=1}^g TP_i}{\sum_{i=1}^g TP_i + FN}$
- Depends on application and importance of characteristics.
- However, it is clear that accuracy usage is inappropriate if the data set is imbalanced. \rightsquigarrow Use alternative metrics.
$$p_{MCC} = \frac{n \sum_{i=1}^g \hat{n}_i - \sum_{i=1}^g n_i \hat{n}_i}{\sqrt{(n^2 - \sum_{i=1}^g \hat{n}_i^2)(n^2 - \sum_{i=1}^g n_i^2)}}$$
- Be careful with comparing the absolute values of the different measures, as these can be on different "scales", e.g., MCC and BAC, where $\hat{n}_i = \sum_{j=1}^g n_{ij}$ is the total number of instances classified as i .
- Cohen's Kappa or Cross Entropy (see Grandini et al. (2021)) treat "predicted" and "true" classes as two discrete random variables.



WHICH PERFORMANCE MEASURE TO USE?

- Since different measures focus on other characteristics \rightsquigarrow No golden answer to this question.
- Depends on application and importance of characteristics.
- However, it is clear that accuracy usage is inappropriate if the data set is imbalanced. \rightsquigarrow Use alternative metrics.
- Be careful with comparing the absolute values of the different measures, as these can be on different “scales”, e.g., MCC and BAC.

