RECAP: PERFORMANCE MEASURES FOR BINARY CLASSIFICATION

- We encourage readers to first go through ▸ Chapter 04.08 in I2ML.
- In binary classification ($\mathcal{Y} = \{-1, +1\}$):

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

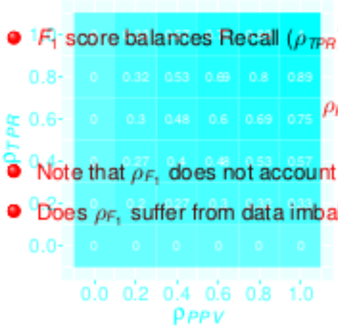
- F_1 score balances Recall (ρ_{TPR}) and Precision (ρ_{PPV}):

$$\rho_{F_1} = 2 \cdot \frac{\rho_{PPV} \cdot \rho_{TPR}}{\rho_{PPV} + \rho_{TPR}}$$

- Note that ρ_{F_1} does not account for TN!
- Does ρ_{F_1} suffer from data imbalance like accuracy does?

Advanced Machine Learning

Imbalanced Learning: Performance Measures



Learning goals

- Know performance measures beyond accuracy
- Know their advantages over accuracy for imbalanced data
- Know extensions of these measures for multiclass settings

F_1 SCORE IN BINARY CLASSIFICATION

RECAP: PERFORMANCE MEASURES FOR BINARY CLASSIFICATION

- We encourage readers to first go through ▸ Chapter 04.08 in I2ML.
- In binary classification ($\mathcal{Y} = \{-1, +1\}$):

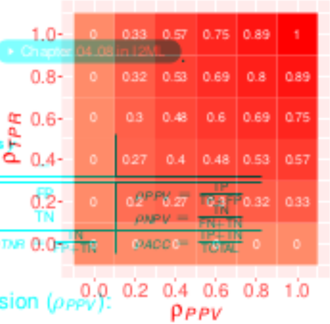
		True Class			
		+	−		
$\hat{y}$	+	TP	FP		$\rho_{PPV} = \frac{TP}{TP+FP}$
	−	FN	TN		$\rho_{NPV} = \frac{TN}{FN+TN}$
		$\rho_{TPR} = \frac{TP}{TP+FN}$	$\rho_{TNR} = \frac{TN}{FP+TN}$		$\rho_{ACC} = \frac{TP+TN}{TOTAL}$

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

$$\rho_{F_1} = 2 \cdot \frac{\rho_{PPV} \cdot \rho_{TPR}}{\rho_{PPV} + \rho_{TPR}}$$

- Note that ρ_{F_1} does not account for TN.
- Does ρ_{F_1} suffer from data imbalance like accuracy does?

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

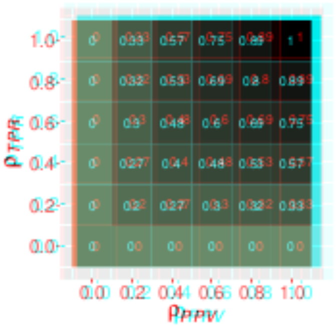


- A model with $\rho_{TPR} = 0$ or $\rho_{PPV} = 0$ has $\rho_{F_1} = 0$.
- Always predicting "negative": $\rho_{TPR} = \rho_{F_1} = 0$
- Always predicting "positive": $\rho_{TPR} = 1 \Rightarrow \rho_{F_1} = 2 \cdot \rho_{PPV}/(\rho_{PPV} + 1) = 2 \cdot n_+/(n_+ + n)$,
  ⇝ small when $n_+ (= TP + FN = TP)$ 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 of weight to $\frac{1}{\rho_{TPR}}$:

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



- $\beta \gg 1 \rightarrow F_\beta \approx \rho_{TPR}$
- $\beta \ll 1 \rightarrow F_\beta \approx \rho_{PPV}$.

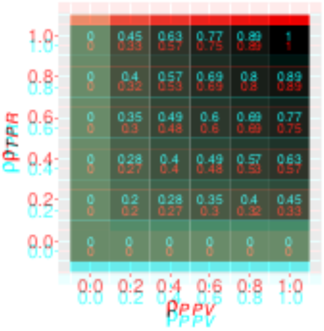
F_1 SCORE IN BINARY CLASSIFICATION

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

- A model with $\rho_{TPR} = 0$ or $\rho_{PPV} = 0$ 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: $\rho_G = \sqrt{\rho_{PPV} \cdot \rho_{TPR}}$
- Geometric mean tends more towards the **lower** of the two combined values.
- Geometric mean is **larger** than harmonic mean.



- Closely related is the G mean: $\rho_{Gm} = \sqrt{\rho_{TNR} \cdot \rho_{TPR}}$.

  It also considers **TN**.

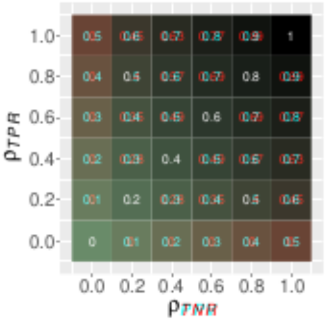
- Always predicting "negative": $\rho_G = \rho_{Gm} = 0 \rightsquigarrow$ Robust to data imbalance!

- 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 of weight to $\frac{1}{\rho_{TPR}}$:
$$F_\beta = \frac{1}{\frac{\beta^2}{1+\beta^2}\cdot\frac{1}{\rho_{TPR}} + \frac{1}{1+\beta^2}\cdot\frac{1}{\rho_{PPV}}} = (1+\beta^2)\cdot\frac{\rho_{PPV}\cdot\rho_{TPR}}{\beta^2\rho_{PPV} + \rho_{TPR}}$$
- $\beta \gg 1 \rightsquigarrow F_\beta \approx \rho_{TPR}$;
- $\beta \ll 1 \rightsquigarrow F_\beta \approx \rho_{PPV}$.

BALANCED ACCURACY

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

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



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

G SCORE AND G MEAN

- G score uses geometric mean:

$$\rho_G = \sqrt{\rho_{PPV} \cdot \rho_{TPR}}$$

- Geometric mean tends more towards the **lower** of the two combined values.
- Geometric mean is **larger** than harmonic mean.
- Closely related is the G mean:

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

It also considers **TN**.

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

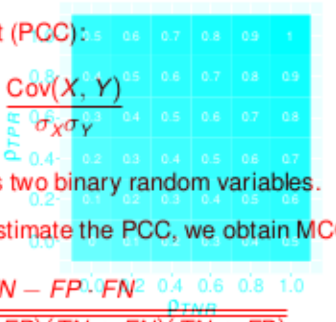
MATTHEWS CORRELATION COEFFICIENT

- Recall: Pearson correlation coefficient (PCC):

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

- 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 balanced, then ...
- 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 $\rho_{BAC} \ll \rho_{ACC}$. It also considers TN.

MATTHEWS CORRELATION COEFFICIENT

- Recall: Pearson correlation coefficient (PCC):

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

- View "predicted" and "true" classes as two binary random variables.
- Using entries in confusion matrix to estimate the PCC, we obtain MCC:

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

- In contrast to other metrics:
  - MCC uses all entries of the confusion matrix;
  - MCC has value in $[-1, 1]$.

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

- $\rho_{MCC} \approx 1 \rightsquigarrow$ 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.
- $\rho_{MCC} \approx -1 \rightsquigarrow$ reversed classification, i.e., switch labels.
- Previous measures requires defining positive class. But MCC does not depend on which class is the positive one.

MULTICLASS CLASSIFICATION

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

- n_{ij}: the number of i instances classified as j.
- $n_i = \sum_{j=1}^{g} n_{ij}$ the total number of i instances.
- **Class-specific** metrics:
  - True positive rate (**Recall**): $\rho_{TPR_i} = \frac{n_{ii}}{n_i}$
  - True negative rate $\rho_{TNR_i} = \frac{\sum_{j \neq i} n_{jj}}{n - n_i}$
  - Positive predictive value (**Precision**) $\rho_{PPR_i} = \frac{n_{ii}}{\sum_{j=1}^{g} n_{ji}}$.

MATTHEWS CORRELATION COEFFICIENT

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

- $\rho_{MCC} \approx 1 \rightsquigarrow$ 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.
- $\rho_{MCC} \approx -1 \rightsquigarrow$ reversed classification, i.e., switch labels.
- Previous measures requires defining positive class. But MCC does not depend on which class is the positive one.

MACRO F_1 SCORE

- Average over classes to obtain a single value:

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

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:

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

- Problem: each class equally weighted $\rightsquigarrow$ class sizes are not considered.
- How about applying different weights to the class-specific metrics?

MULTICLASS CLASSIFICATION

		True Class y			
		1	2	...	g
Classification	1	n_{11} (True 1's)	n_{12} (False 1's for 2's)	...	n_{1g} (False 1's for g's)
$\hat{y}$	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)

- n_{ij}: 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:
  - True positive rate (**Recall**): $\rho_{TPR_i} = \frac{n_{ii}}{n_i}$
  - True negative rate $\rho_{TNR_i} = \frac{\sum_{j \neq i} n_{jj}}{n - n_i}$
  - Positive predictive value (**Precision**) $\rho_{PPV_i} = \frac{n_{ii}}{\sum_{j=1}^{g} n_{ij}}$.

WEIGHTED MACRO F_1 SCORE

- 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 $\sum_{i=1}^{g} w_i = 1$.

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

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

- where $METRIC_i$ is a class-specific metric such as PPV_i, TPR_i of class i.
- Example: $w_i = \frac{n - n_i}{(g-1)n}$ are suitable weights.
- Weighted macro F_1 score:

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

$$\rho_{wmF_1} = 2 \cdot \frac{\rho_{wmPPV} \cdot \rho_{wmTPR}}{\rho_{wmPPV} + \rho_{wmTPR}}$$

- Problem: each class equally weighted → 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 $\frac{\sum_{i=1}^{g} TP_i}{\sum_{i=1}^{g} TP_i + FN_i}$
- MCC can be extended to:

$$\rho_{MCC} = \frac{n \sum_{i=1}^{g} n_{ii} - \sum_{i=1}^{g} \hat{n}_i n_i}{\sqrt{(n^2 - \sum_{i=1}^{g} \hat{n}_i^2)(n^2 - \sum_{i=1}^{g} n_i^2)}},$$

where $\hat{n}_i = \sum_{j=1}^{g} n_{ji}$ 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.

- 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 $\sum_{i=1}^{g} w_i = 1$.

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

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

- Example: $w_i = \frac{n - n_i}{(g-1)n}$ are suitable weights.
- Weighted macro F_1 score:

$$\rho_{wmF_1} = 2 \cdot \frac{\rho_{wmPPV} \cdot \rho_{wmTPR}}{\rho_{wmPPV} + \rho_{wmTPR}}$$

- 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 MEASURES TO USE?

- Since different measures focus on other characteristics ⇝ 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. ⇝ 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.

$$\rho_{MCC} = \frac{n\sum_{i=1}^{g} n_{ii} - \sum_{i=1}^{g} \hat{n}_i n_i}{\sqrt{(n^2 - \sum_{i=1}^{g} \hat{n}_i^2)(n^2 - \sum_{i=1}^{g} n_i^2)}},$$

where $\hat{n}_i = \sum_{j=1}^{g} n_{ij}$ 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.