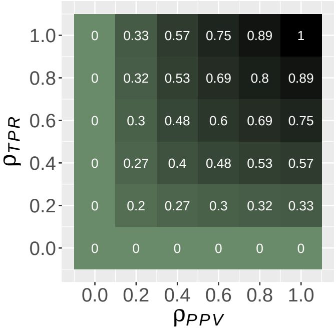
Advanced Machine Learning

Imbalanced Learning: Performance Measures



Learning goals

- Get to know alternative performance measures for accuracy
- See their advantages over accuracy for imbalanced data sets
- Understand the extensions of these measures for multiclass settings

CONFUSION MATRIX

- The confusion matrix gives an overview over the errors as well as correct classifications in a tabulated form. Most performance/evaluation measures can be computed from the confusion matrix.
- In binary classification (i.e., $\mathcal{Y} = \{-1, +1\}$):

		True Class *y*	
		$+$	$-$
Classification	$+$	True Positive (TP)	False Positive (FP)
$\hat{y}$	$-$	False Negative (FN)	True Negative (TN)

- In multiclass classification (i.e., $\mathcal{Y} = \{1, \dots, g\}$):

		True Class *y*			
		1	2	...	*g*
Classification	1	True 1's	False 1's for 2's	...	False 1's for *g*'s
	2	False 2's for 1's	True 2's	...	False 2's for *g*'s
$\hat{y}$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
	g	False *g*'s for 1's	False *g*'s for 2's	...	True *g*'s

PERFORMANCE MEASURES FOR BINARY CLASSIFICATION

We first focus on the binary classification setting and review the relevant performance measures.

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

- True positive rate (**Recall**) ρ_{TPR}: fraction of correctly classified 1s over all 1s.

⇝ Population counterpart: $\mathbb{P}(\hat{y} = 1 \mid y = 1)$

- True negative rate ρ_{TNR}: fraction of correctly classified -1s over all -1s.

⇝ Population counterpart: $\mathbb{P}(\hat{y} = -1 \mid y = -1)$

- Positive predictive value (**Precision**) ρ_{PPV}: fraction of correctly classified 1s over all 1 classifications.

⇝ Population counterpart: $\mathbb{P}(y = 1 \mid \hat{y} = 1)$

- Negative predictive value ρ_{NPV}: fraction of correctly classified -1s over all -1 classifications.

⇝ Population counterpart: $\mathbb{P}(y = -1 \mid \hat{y} = -1)$

- Accuracy ρ_{ACC}: fraction of correct classifications.

⇝ Population counterpart: $\mathbb{P}(\hat{y} = y)$

F_1 SCORE IN BINARY CLASSIFICATION

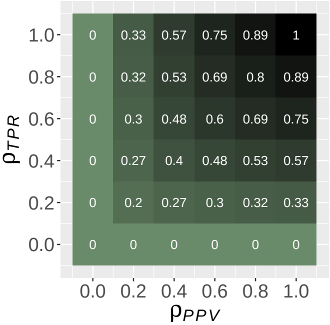
- It is difficult to achieve high **positive predictive value** and high **true positive rate** simultaneously:
    - A classifier predicting more positive will be more sensitive (higher ρ_{TPR}), but it will also tend to give more *false* positives (lower ρ_{TNR}, lower ρ_{PPV}).
    - A classifier that predicts more negatives will be more precise (higher ρ_{PPV}), but it will also produce more *false* negatives (lower ρ_{TPR}).
- The F_1 **score** ρ_{F_1} balances two conflicting goals:
    1. Maximizing positive predictive value
    2. Maximizing true positive rate
- ρ_{F_1} is the harmonic mean of ρ_{PPV} and ρ_{TPR}:

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

- Note ρ_{F_1} does not account for the number of true negatives.

F_1 SCORE IN BINARY CLASSIFICATION

F_1 score for different combinations of ρ_{PPV} & ρ_{TPR}.
$\rightarrow$ Tends more towards the lower of the two combined values.



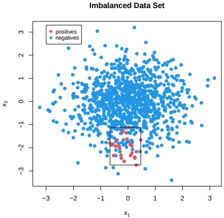
- A model with $\rho_{TPR} = 0$ (no positive instance predicted as positive) or $\rho_{PPV} = 0$ (no true positives among the predicted) has $\rho_{F_1} = 0$.
- Always predicting "negative": $\rho_{F_1} = 0$.

$\rightsquigarrow$ No "F_1 score paradox"!

- Always predicting "positive": $\rho_{F_1} = 2 \cdot \rho_{PPV}/(\rho_{PPV} + 1) = 2 \cdot n_+/(n_+ + n)$, which will be small when the size of the positive class n_+ is small.

ACCURACY PARADOX EXAMPLE: F_1 SCORE

- Recalling our exemplary setting to illustrate the accuracy paradox:
  - $p(\mathbf{x} \mid -1) \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$,
  - $p(\mathbf{x} \mid +1) \sim \mathcal{N}_2\left(\begin{pmatrix}0\\-2\end{pmatrix}, \begin{pmatrix}0.1 & 0\\0 & 0.1\end{pmatrix}\right)$,
  - $n_+ = 25$, $n_- = 1000$,
  - $f_1(\mathbf{x}) \equiv -1$,
  - $f_2(\mathbf{x}) = 2 \cdot \mathbb{1}_{[\mathbf{x} \in [-0.66, 0.47] \times [-2.74, -1.12]]} - 1$.



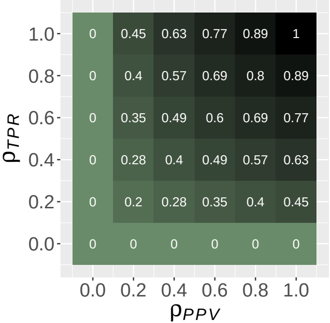
- f_1 has an overall accuracy of ≈ 0.976, while f_2 has an overall accuracy of ≈ 0.951.
- The F_1 score of f_1 is zero, as it has $\rho_{TPR} = 0 =$ (since $TP = 0$), while f_2 has F_1 score of $1/2$, since TP=25, FN = 0 and FP=50, so that $\rho_{PPV} = 1/3$ and $\rho_{TPR} = 1$.

G SCORE

- Instead of the harmonic mean in the F_1 score one can use also the geometric mean, which results in the G score:

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

- Its behavior is similar to the F_1 score for different combinations of ρ_{PPV} & ρ_{TPR}, i.e., it tends more towards the lower of the two combined values.



- Closely related is the G mean, which uses the geometric mean of TPR and TNR:

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

It also takes the true negative rate into account, i.e., also true negatives.

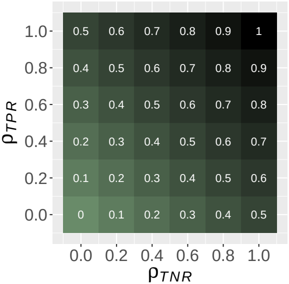
- For the accuracy paradox example we get a G score of ≈ 0.577 and a G mean of ≈ 0.975 for f_2, while f_1 achieves G score and G mean of zero, respectively.
- Always predicting “negative”: $\rho_G = \rho_{Gm} = 0 \rightsquigarrow$ No “*G* score/mean paradox”!

BALANCED ACCURACY

- Finally, one can also replace the geometric mean in the *G* mean by the arithmetic mean, which results in the balanced accuracy (BAC):

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

- It tends more towards the higher of the two combined values.



- If a classifier has good predictive accuracy on both classes or the data set is almost balanced, then ρ_{BAC} is essentially the classical accuracy ρ_{ACC}.
- However, if a classifier always predicts “negative” for an imbalanced data set, i.e. $n_+ \ll n_-$, then ρ_{BAC} is much lower than ρ_{ACC}. It also takes the true negative rate into account, i.e., also true negatives.
- For the accuracy paradox example we get a BAC of 0.975 for f_2, while f_1 achieves a BAC of 0.5. As a reminder, f_1 has an overall accuracy (ρ_{ACC}) of ≈ 0.976, while f_2 has an overall accuracy of ≈ 0.951.

MATTHEUS CORRELATION COEFFICIENT

- Mattheus Correlation Coefficient (MCC) (aka Phi coefficient) is a measure for the correlation between two binary random variables. In the classification setting the "predicted" classes and the "true" classes are the two discrete random variables:

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

- In contrast to all previous measures the MCC uses all entries of the confusion matrix and its value is in $[-1, 1]$ with the following interpretation:
    - $\rho_{MCC} \approx 1 \rightsquigarrow$ good classification, i.e., strong correlation between correct classifications and true classes.
    - $\rho_{MCC} \approx 0 \rightsquigarrow$ no correlation, i.e., the classification is no better than random guessing.
    - $\rho_{MCC} \approx -1 \rightsquigarrow$ reversed classification, i.e., the classifier is essentially switching the labels.
- While for the previous measures it is important which class is the positive one (especially in light of imbalanced data), MCC does not depend on which class is the positive one.
- For the accuracy paradox example we get an MCC of ≈ 0.563 for f_2, while f_1 achieves an MCC of 0.

PERFORMANCE MEASURES FOR MULTICLASS CLASSIFICATION

Now consider the multiclass classification setting and the corresponding confusion matrix:

		True Class y			
		1	2	. . .	g
Classification	1	$n(1,1)$	$n(1,2)$	. . .	$n(1,g)$
		(True 1's)	(False 1's for 2's)	. . .	(False 1's for g's)
	2	$n(2,1)$	$n(2,2)$	. . .	$n(2,g)$
$\hat{y}$		(False 2's for 1's)	(True 2's)	. . .	(False 2's for g's)
	⋮	⋮	⋮	. . .	⋮
	g	$n(g,1)$	$n(g,2)$	. . .	$n(g,g)$
		(False g's for 1's)	(False g's for 2's)	. . .	(True g's)

Here, $n(i,j)$ is the number of j instances classified as i and $n_i = \sum_{j=1}^{g} n(j,i)$ the total number of i instances.

We can get multiclass counterparts of the binary evaluation measures by making them class specific:

- True positive rate (**Recall**): $\rho_{TPR_i} = \frac{n(i,i)}{n_i}$ the fraction of correctly classified instances i among all i instances.
- True negative rate $\rho_{TNR_i} = \frac{\sum_{j \neq i} n(j,j)}{n - n_i}$ the fraction of correctly classified non-i instances among all non-i instances.
- Positive predictive value (**Precision**) $\rho_{PPR_i} = \frac{n(i,i)}{\sum_{j=1}^{g} n(i,j)}$, the fraction of correctly classified instances i among all i classifications.

MACRO F_1 SCORE

- In order to obtain a single “true positive rate”, “true negative rate”, or “positive predictive value” we can simply average all class-specific rates over all available classes:
  - Macro average true positive rate (macro average precision): $\rho_{mTPR} = \frac{1}{g}\sum_{i=1}^{g} \rho_{TPR_i}$
  - Macro average true negative rate : $\rho_{mTNR} = \frac{1}{g}\sum_{i=1}^{g} \rho_{TNR_i}$
  - Macro positive predictive value (macro average recall): $\rho_{mPPR} = \frac{1}{g}\sum_{i=1}^{g} \rho_{PPR_i}$
- With this, one can simply define a macro F_1 score by using the harmonic mean of the macro average precision and macro average recall:

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

- The problem of this extension for imbalanced data sets is that each class gets an equal weight in the macro average such that the class sizes are not taken into account.

WEIGHTED MACRO F_1 SCORE

- For imbalanced data sets it is thus better to use a weighted average of the class-specific rates with weights giving more weight to minority classes and few weight to majority classes, i.e., $w_1, \ldots, w_g \in [0, 1]$ such that $w_i > w_j$ iff $n_i < n_j$ and $\sum_{i=1}^{g} w_i = 1$. With this, we obtain
  - Macro weighted average true positive rate (macro average precision): $\rho_{wmTPR} = \sum_{i=1}^{g} \rho_{TPR_i} \boldsymbol{w}_i$
  - Macro weighted average true negative rate : $\rho_{wmTNR} = \sum_{i=1}^{g} \rho_{TNR_i} \boldsymbol{w}_i$
  - Macro weighted positive predictive value (macro average recall): $\rho_{wmPPR} = \sum_{i=1}^{g} \rho_{PPR_i} \boldsymbol{w}_i$
- Example: $w_i = \frac{n-n_i}{(g-1)n}$ are suitable weights.
- This leads to the weighted macro F_1 score: $\rho_{wmF_1} = 2 \cdot \dfrac{\rho_{wmPPV} \cdot \rho_{wmTPR}}{\rho_{wmPPV} + \rho_{wmTPR}}$
- Following this idea, it is straightforward to obtain a weighted macro G score/measure or weighted BAC.
- **Usually** the weighted F_1 score is used with weights $w_i = n_i/n$, i.e., the relative frequency of the *i*-th class in the data set. However, for imbalanced data sets this is not appropriate as this would give majority classes even more weight.

OTHER PERFORMANCE MEASURES

- There are also “micro” versions of the F_1 score for the multiclass setting, where, for example, the micro weighted average true positive rate (micro average precision) is $\sum_{i=1}^{g} \rho_{TPR_i} \frac{1}{n}$. However, the micro F_1 score essentially boils down to the accuracy in this case.
- Moreover, the Mattheus Correlation Coefficient (MCC) can be extended to the multiclass setting:

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

where $\hat{n}_i = \sum_{j=1}^{g} n(i,j)$ is the total number of instances classified as i.

- Finally, there are also other performance measures which are based on the idea of treating the "predicted" classes and the "true" classes as two discrete (or categorical) random variables, e.g. Cohen’s Kappa or Cross Entropy (see Grandini et al. (2021)).

WHICH PERFORMANCE MEASURE TO USE?

- As we have seen, there is a plethora of measures available, which gives rise to the question which of them should be used?
- Since all of these measures are focusing on different characteristics of the data, this is a question which in general cannot be answered unambiguously, as this is often depending on the application at hand, which characteristic is more important than another.
- However, it is clear that the usage of accuracy is inappropriate if the data set is imbalanced and another alternative measures should be considered. Usually, the overall picture obtained by the alternative measures will not differ too much.
- Finally, one needs to be careful by comparing the absolute values of the different measures, as these can be on different “scales”, e.g. MCC and BAC.