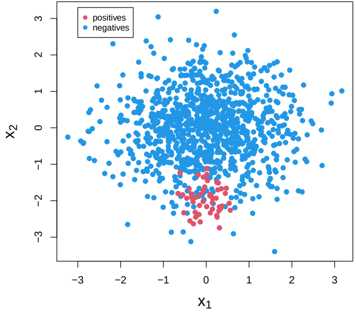
Applied Machine Learning

Imbalanced Data: Problem and Diagnostics



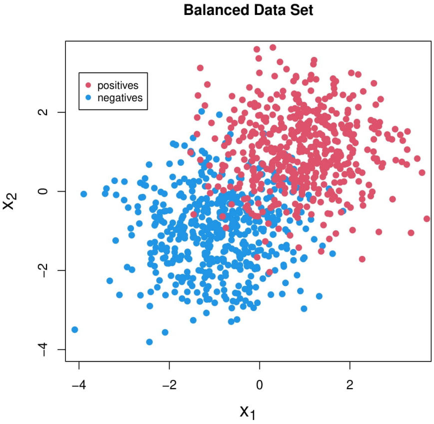
Learning goals

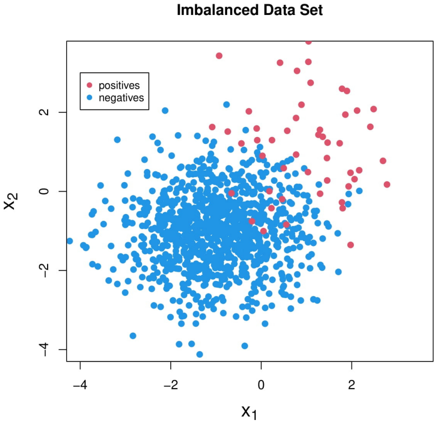
- Know what an imbalanced data set is
- Understand disadvantage of accuracy on imbalanced data
- Learn evaluation metrics suitable for imbalanced data

Introduction to Imbalanced Data

IMBALANCED DATA SETS

- Class imbalance: Ratio of classes is significantly different.
- Consequence: Undesirable predictive behavior for smaller class.
- Example: Sampling from two Gaussian distributions





IMBALANCED DATA SETS: EXAMPLES

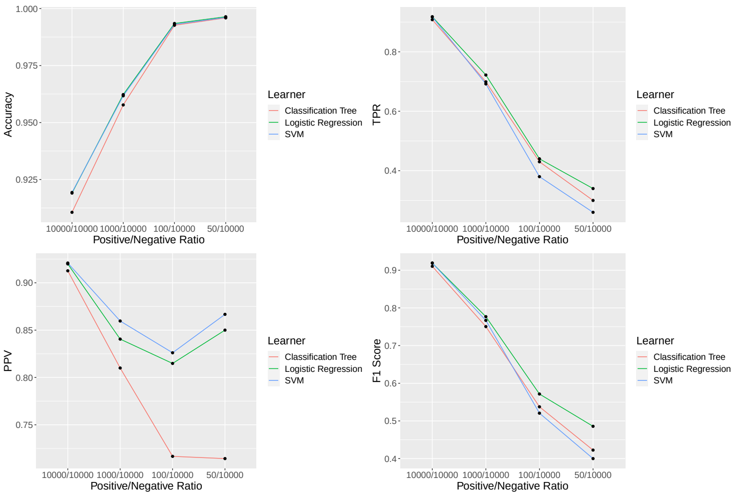
Domain	Task	Majority Class	Minor Class
Medicine	Predict tumor pathology	Benign	Malignant
Information retrieval	Find relevant items	Irrelevant items	Relevant items
Tracking criminals	Detect fraud emails	Non-fraud emails	Fraud emails
Weather prediction	Predict extreme weather	Normal weather	Tornado, hurricane

- Often, the minority class is the more important class.
- Imbalanced data can be a source of bias related to concept of fairness.

ISSUES WITH EVALUATING CLASSIFIERS

- Ideal case: correctly classify as many instances as possible
  $\Rightarrow$ High accuracy, preferably 100%.
- In practice, we often obtain on imbalanced data sets: **good** performance on the **majority** class(es), a **poor** performance on the **minority** class(es).
- Reason: the classifier is biased towards the **majority** class(es), as predicting the majority class pays off in terms of accuracy.
- Focusing only on accuracy can lead to bad performance on minority class.
- Example:
  - Assume that only 0.5% of the patients have a disease,
  - Always predicting “no disease” leads to accuracy of 99.5%

ISSUES WITH EVALUATING CLASSIFIERS



In each scenario, we have 10.000 obs in the negative class. Number of obs in positive class varies between 10.000, 1.000, 100, and 50. Train classifiers with 10-fold stratified cv. Evaluate via aggregated predictions on test set.

POSSIBLE SOLUTIONS

- Ideal performance metric: the learning is *properly* biased towards the minority class(es).
- Imbalance-aware performance metrics:
  - G-score
  - Balanced accuracy
  - Matthews Correlation Coefficient
  - Weighted macro F_1 score

POSSIBLE SOLUTIONS

Approach	Main idea	Remark
Algorithm-level	Bias classifiers towards minority	Special knowledge about classifiers is needed
Data-level	Rebalance classes by resampling	No modification of classifiers is needed
Cost-sensitive Learning	Introduce different costs for misclassification when learning	Between algorithm- and data-level approaches
Ensemble-based	Ensemble learning plus one of three techniques above	-

Performance Measures for Imbalanced Data

RECAP: PERFORMANCE MEASURES FOR BINARY CLASSIFICATION

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

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

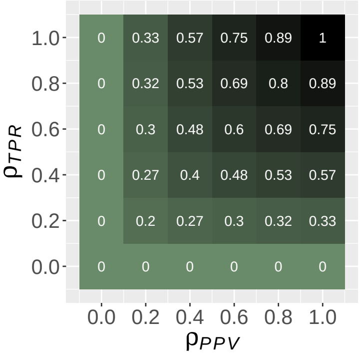
- F_1 score balances Recall (ρ_{TPR}) and Precision (ρ_{PPV}):

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

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

F_1 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.

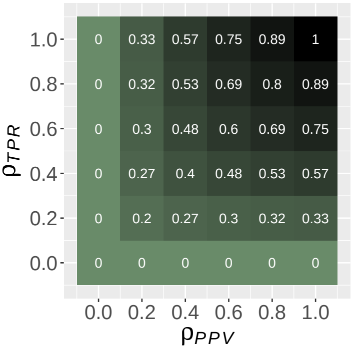


- A model with $\rho_{TPR} = 0$ or $\rho_{PPV} = 0$ has $\rho_{F_1} = 0$.
- Always predicting “negative”: $\rho_{TPR} = \rho_{F_1} = 0$
- Always predicting “positive”: $\rho_{TPR} = 1 \Rightarrow \rho_{F_1} = 2 \cdot \rho_{PPV}/(\rho_{PPV} + 1) = 2 \cdot n_+/(n_+ + n)$,
  $\rightsquigarrow$ small when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

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 $\frac{1}{\rho_{TPR}}$:

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



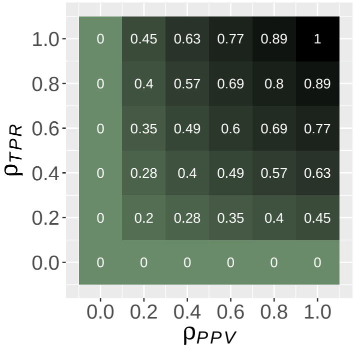
- $\beta \gg 1 \rightsquigarrow F_\beta \approx \rho_{TPR}$;
- $\beta \ll 1 \rightsquigarrow F_\beta \approx \rho_{PPV}$.

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:

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

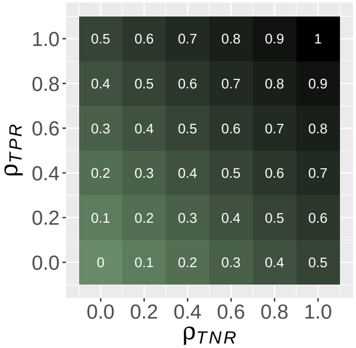
  It also considers **TN**.

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

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.

MATTHEWS CORRELATION COEFFICIENT

- Recall: Pearson correlation coefficient (PCC):

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

- 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 $[-1, 1]$.

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.

MULTICLASS CLASSIFICATION

		True Class y			
		1	2	$\dots$	g
Classification	1	n_{11} (True 1's)	n_{12} (False 1's for 2's)	$\dots$	n_{1g} (False 1's for g's)
$\hat{y}$	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_{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:
  - 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**) $\rho_{PPR_j} = \frac{n_{jj}}{\sum_{i=1}^{g} n_{ji}}$.

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?

WEIGHTED MACRO F_1 SCORE

- For imbalanced data sets, give **more weights** to **minority** classes.
- $w_1, \ldots, w_g \in [0, 1]$ such that $w_i > w_j$ iff $n_i < n_j$ and $\sum_{i=1}^{g} w_i = 1$.

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

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.

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 $\hat{n}_i = \sum_{j=1}^{g} n_{ij}$ 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.
- Area under the ROC curve is also immune against class imbalance.