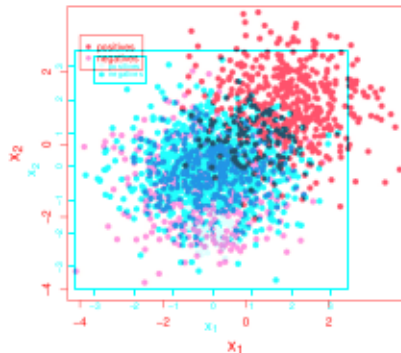


IMBALANCED DATA SETS

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



Balanced Data Set



Imbalanced Data Set

Learning goals

- Know what an imbalanced data set is
- Understand disadvantage of accuracy on imbalanced data
- Know techniques for handling imbalanced data sets

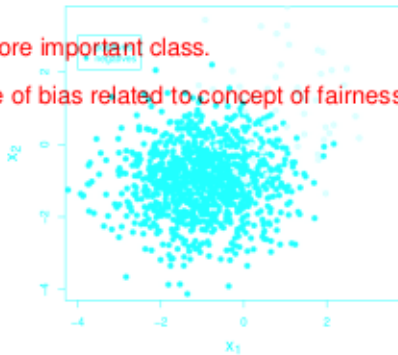
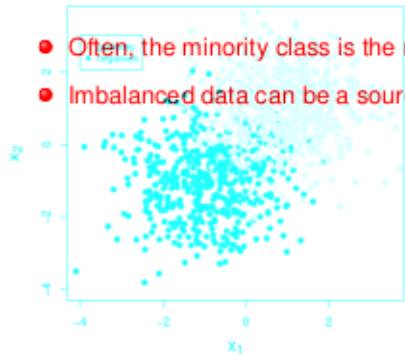


IMBALANCED DATA SETS: EXAMPLES

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
⇒ High accuracy, preferably 100%.

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

- 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 imbalanced data can be a source of bias related to concept of fairness. 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

- Ideal case: correctly classify as many instances as possible
⇒ 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%



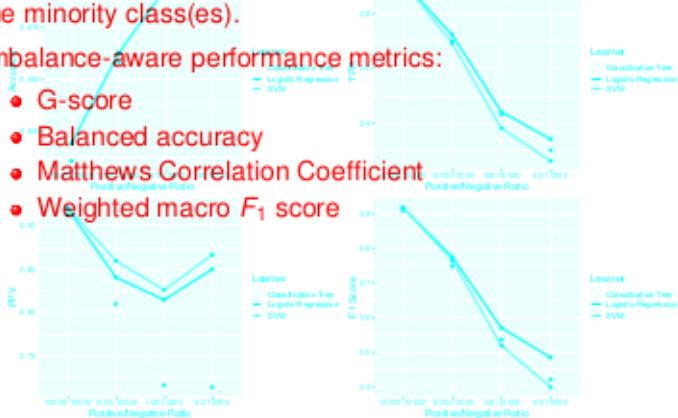
In each scenario, we have 10.000 obs in the negative class. Number of obs in positive class varies between 10.000, 1.000, 100, and 50. Train classifiers with 10-fold stratified cv. Evaluate via aggregated predictions on test set.

POSSIBLE SOLUTIONS TO IMBALANCED CLASSIFIERS

- 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



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)

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	-



- 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	-

