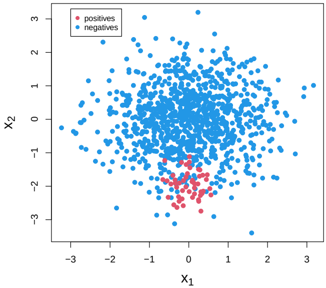
Advanced Machine Learning

Introduction to Imbalanced Learning

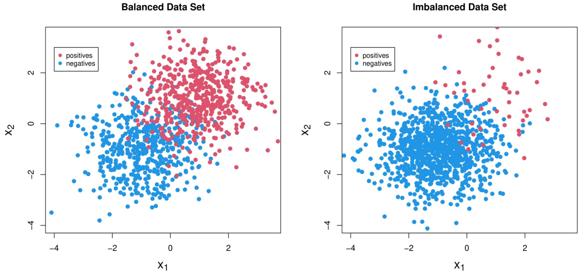


Learning goals

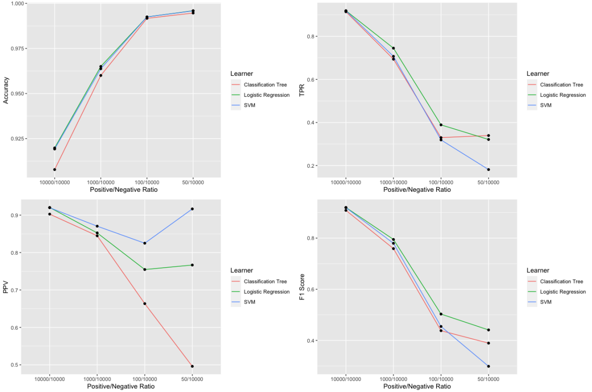
- What are imbalanced data sets
- Disadvantage of accuracy on imbalanced data sets
- Overview of techniques for handling imbalanced data sets

IMBALANCED DATA SETS

- Class imbalance: the occurrences of the classes are significantly different.
- Consequence: undesirable predictive behavior.
- Example:
  - Sampling from two Gaussian distributions



IMBALANCED DATA SETS: BENCHMARK



- Train classifiers on four datasets with threefold cv. Hold the negative class constant at 10000 examples. For the positive class, consider class sizes of 10000, 1000, 100 and 50.

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

- In binary classification, the minority class is usually the positive class, while the majority class is the negative.
- The positive class is oftentimes the more important one in real-world applications.
- Recall that imbalanced data sets can also be a source of bias related to the concept of fairness in ML, e.g. more data on white recidivism outcomes than for blacks.

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:
  - a **good** accuracy on the **majority** class(es),
  - a **poor** accuracy 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 the overall accuracy can have serious consequences.

ISSUES WITH EVALUATING CLASSIFIERS

- Example:
  - Assume that only 0.5% of the patients have the disease,
  - Always predicting “no disease” $\rightsquigarrow$ accuracy of 99.5%
    $\rightsquigarrow$ Every patient is sent back home!

- 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

TRAINING CLASSIFIERS ON IMBALANCED DATASETS

Approach	Main idea	Remark
Algorithm-level	Bias classifiers towards minority	Special knowledge about classifiers is needed
Data-level	Re-balance the classes by resamping	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	-