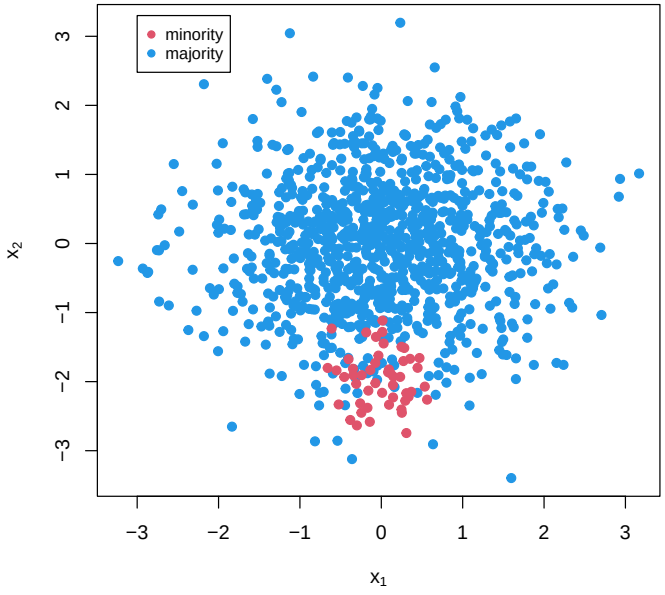
Advanced Machine Learning

Introduction to Imbalanced Learning

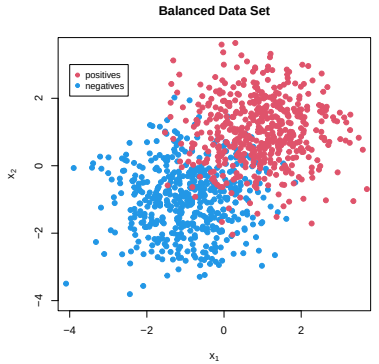


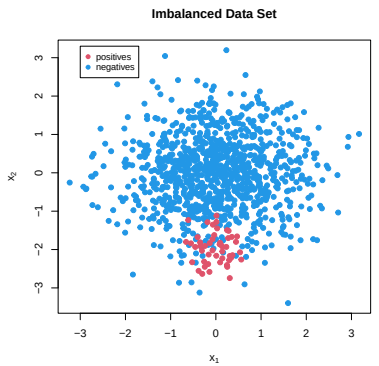
Learning goals

- Know what is meant by imbalanced data sets and how they occur in practice
- Understand why accuracy is not an optimal performance measure for imbalanced data sets
- Get an overview of the prevalent techniques for dealing with imbalanced data sets

IMBALANCED DATA SETS

- Class imbalance refers to the case in which the occurrences of the classes are significantly different. Usually one or more classes are underrepresented in the data.
- Such situations are common in practice, and classical classification algorithms sometimes exhibit undesirable predictive behavior without adaptation to this imbalance.





- In binary classification (i.e., $\mathcal{Y} = \{-1, +1\}$) the minority (underrepresented) class is usually the positive class ($y = +1$), while the majority (overrepresented) class is the negative ($y = -1$). This is because the positive class is oftentimes the more important one in real-world applications.

IMBALANCED DATA SETS: EXAMPLES

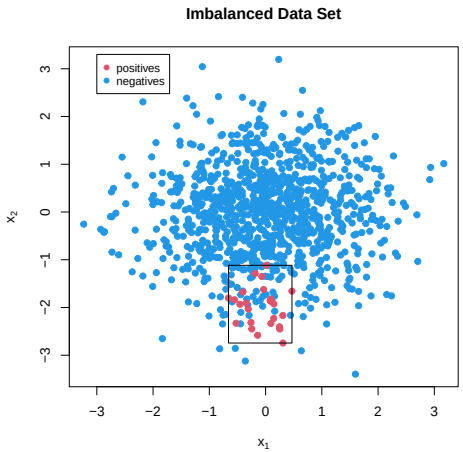
- Some real-life examples, in which we have to deal with imbalanced data sets:
  - Medicine — If we are interested in predicting whether a patient has a serious disease, we often encounter imbalanced class distributions, since not many people have the disease.
  - Information Retrieval — We want to filter out the relevant items (e.g. documents, websites, . . .) for a user query. However, the number of relevant items is very small compared to the size of available items in the entire database.
  - Tracking criminals — We want to detect terrorists in social networks in order to prevent fatal terrorist attacks. However, there are only a few terrorists and most people are righteous people.
  - (Extreme) Weather prediction — We want to predict whether an extreme weather event (e.g., tornado, hurricane, . . .) will occur. (Fortunately) the weather is mostly ”normal” and we have only few data points available for predicting such events.
  - . . .
- Note that the positive class in all examples is the more important one: present disease, relevant documents, terrorists, extreme weather event, . . .
- 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 CLASSICAL CLASSIFIERS

- As usual in classification learning settings, we want a classifier which classifies as many instances as possible correctly, i.e., one which has a high accuracy, preferably 100%.
- What can be observed in practice when dealing with imbalanced data sets is that classical classifiers have
  - a good accuracy on the majority class(es),
  - a poor accuracy on the minority class(es).
- The reason for this is essentially that they are biased towards the majority class(es), as predicting the majority class pays off in terms of accuracy.
- This problem can be illustrated by means of the **accuracy paradox**.

ACCURACY PARADOX: EXAMPLE

- Consider the following setting:
  - $p(\mathbf{x} \mid -1) \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
  - $p(\mathbf{x} \mid +1) \sim \mathcal{N}_2\left(\begin{pmatrix}0\\-2\end{pmatrix}, \begin{pmatrix}0.1 & 0\\0 & 0.1\end{pmatrix}\right)$
  - $n_+ = 25$ and $n_- = 1000$



- We compare two models $f_1(\mathbf{x}) \equiv -1$ and $f_2(\mathbf{x}) = 2 \cdot \mathbb{1}_{[\mathbf{x} \in [-0.66, 0.47] \times [-2.74, -1.12]]} - 1$.
- f_1 has an overall accuracy of ≈ 0.976, while f_2 has an overall accuracy of ≈ 0.951.
- However, f_1 has not a single correct classification for the positive class, i.e., an accuracy of 0 for the positive class, while f_2 classifies **all** positives correctly, i.e., an accuracy of 1 for the positive class.

ACCURACY PARADOX: CONSEQUENCES

- Focusing only on the overall accuracy can have serious consequences in real-life applications. Take the disease prediction example:
  - Assume that only 0.5% of the patients have the disease,
  - Always predicting “no disease” has an accuracy of 99.5% , but this would mean that every patient is send back home
    $\rightsquigarrow$ a very bad doctor!
- Accuracy is a questionable performance measure for imbalanced learning settings and consequently we need more informative performance measures for such cases.
- Ideally the performance measure should be such that the learning is biased towards the minority class(es), but of course we should not overdo it.

DEALING WITH IMBALANCED DATA SETS

- There are four prevalent techniques how to deal with imbalanced data sets:
  1. Algorithm-level/internal approaches — Modify existing classifiers (e.g. SVMs, decision trees, . . .) such that they are biased towards the minority class(es).
     $\rightsquigarrow$ Special knowledge of the classifier, but sometimes also on the application domain is needed, i.e., one needs to understand why the classifier fails for the underlying learning scenario.
  2. Data-level/external approaches — (Re-)Balance the classes by (re-)sampling from the data space.
     $\rightsquigarrow$ No modification of a "classical" classifier is needed, as the data itself is pre-processed such that accuracy as the used performance measure is fine.
  3. Cost-sensitive learning — Introducing different costs for misclassification and incorporating these costs into the learning process.
     $\rightsquigarrow$ This approach is somewhere between the data-level approaches due to adding costs to instances and the algorithm-level approaches due to (possibly) modifying learning algorithms to consider costs.
  4. Ensemble-based approaches — Combining ensemble learning algorithms with one of the three techniques above.