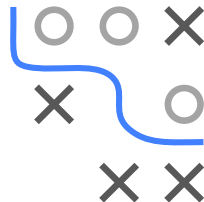
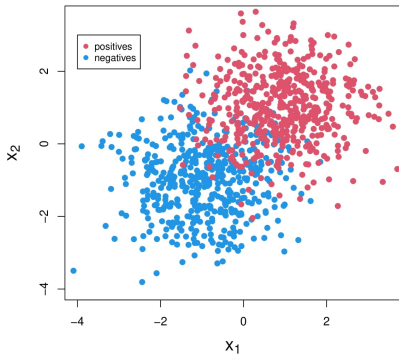


IMBALANCED DATA SETS

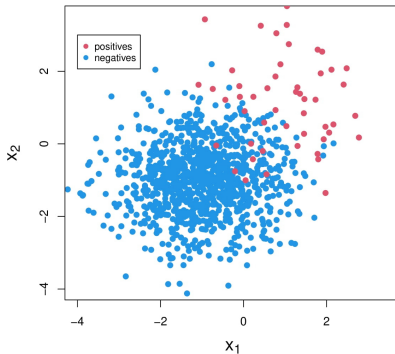
- Class imbalance: Ratio of classes is significantly different.
- Consequence: Undesirable predictive behavior for smaller class.
- Example: Sampling from two Gaussian distributions



Balanced Data Set



Imbalanced Data Set



IMBALANCED DATA SETS: EXAMPLES

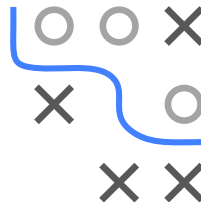
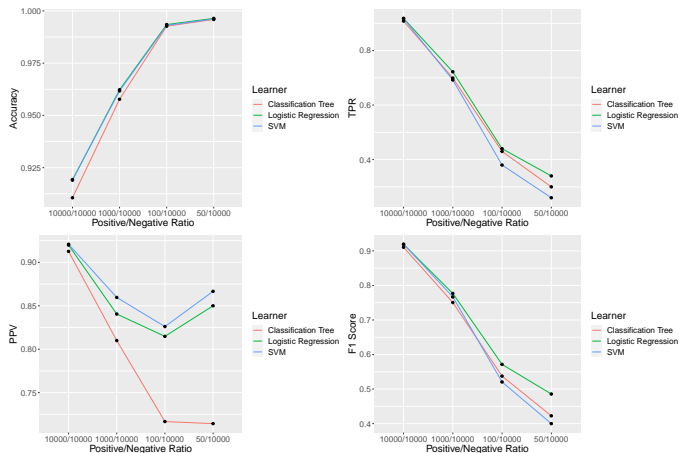
Domain	Task	Majority Class	Minor Class
Medicine	Predict tumor pathology	Benign	Malignant
Information retrieval	Find relevant items	Irrelevant items	Relevant items
Tracking criminals	Detect fraud emails	Non-fraud emails	Fraud emails
Weather prediction	Predict extreme weather	Normal weather	Tornado, hurricane



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

- 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

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	-

