Statistical issues emerging in training and evaluating classification models in presence of rare events

Giovanna Menardi[1]
Nicola Torelli[2]

[1]Department of Statistical Sciences,
University of Padova
menardi@stat.unipd.it
[2]Department of Economics, Business, Mathematics and Statistics,
University of Trieste
nicola.torelli@econ.units.it

June, 7th 2010, Ljubljana

Framework

- $y \in \mathcal{Y} = \{\mathcal{Y}_0, \mathcal{Y}_1\}$, dependent variable, $\mathbf{x} \in \mathcal{X}$, set of covariates.
- $T_n = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)$ i.i.d realizations from an unknown probability distribution F on $\mathcal{X} \times \mathcal{Y}$.
- Split data in:
  - training set $\longrightarrow$ used to estimate the model
  - test set $\longrightarrow$ used to evaluate the model
- $\mathbf{R}_{T_n} : \mathcal{X} \mapsto \mathcal{Y}$ allows for future prediction of the response variable y based on observation of $\mathbf{x}$ only.
- $\mathbf{R}_{T_n}$ produces a partition of $\mathcal{X}$ in subspaces, each of them associated with a label class $\mathcal{Y}_j$ of $\mathcal{Y}$ if:

$$\frac{\hat{P}(\mathcal{Y}_j|\mathbf{x})}{\hat{P}(\mathcal{Y}_k|\mathbf{x})} > t, \text{ for some } t, \quad k \neq j$$

- wide number of solutions!

Problem

What happens if one of the two classes is rare?

- Examples in many domains: finance (detection of defaulter credit applicants), epidemiology (diagnosis of rare diseases), social sciences (analysis of anomalous behaviors), computer sciences (identification of some features of interest in image data).
- The class imbalance is due to the nature of data or to limits in the data collection process, thus creating an artificial imbalance.

$\longrightarrow$ The class imbalance heavily affects the classification

Summary

- The class imbalance problem
- ROSE: Random OverSampling Examples for estimating and evaluating classifiers
- Some applications to real data

Effects of class imbalance

A skewed distribution of the classes highly affects the classification at the two levels:

- Model estimation
- Model assessment

1. Model assessment

- Choice of the measure of accuracy

1. Model assessment

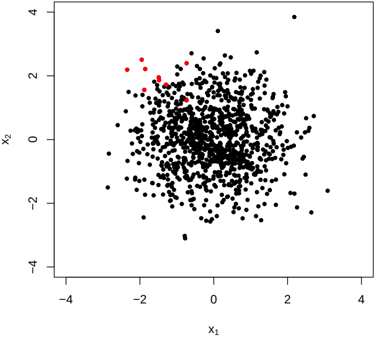
- Choice of the measure of accuracy

  Common error measures conduct to misleading results

  Example: the misclassification error measures the proportion of examples allocated to the wrong class.

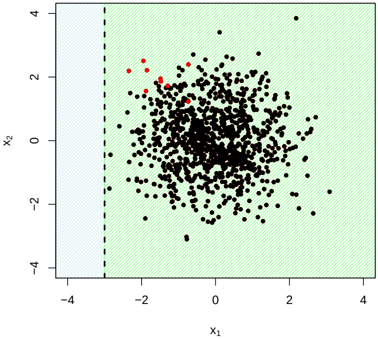
1. Model assessment

- Choice of the measure of accuracy
- Example: $n = 1000$, $P(\mathcal{Y}_1) = 0.01$



1. Model assessment

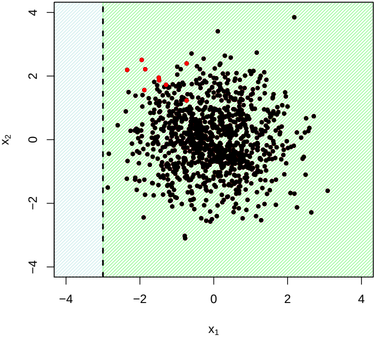
- Choice of the measure of accuracy
- Example: $n = 1000$, $P(\mathcal{Y}_1) = 0.01$



$$R_{T_n}(\mathbf{x}) = \begin{cases} \mathcal{Y}_0 \text{ if } \mathbf{x} = (x_1, x_2)' \text{ st } x_1 > -3 \\ \mathcal{Y}_1 \text{ if } \mathbf{x} = (x_1, x_2)' \text{ st } x_1 \leq -3 \end{cases}$$

1. Model assessment

- Choice of the measure of accuracy
- Example: $n = 1000$, $P(\mathcal{Y}_1) = 0.01$



$$R_{T_n}(\mathbf{x}) = \begin{cases} \mathcal{Y}_0 \text{ if } \mathbf{x} = (x_1, x_2)' \text{ st } x_1 > -3 \\ \mathcal{Y}_1 \text{ if } \mathbf{x} = (x_1, x_2)' \text{ st } x_1 \leq -3 \end{cases} \longrightarrow \text{Err} = 0.01.$$

1. Model assessment

- Choice of the measure of accuracy
  Common error measures conduct to misleading results
  $\longrightarrow$ use of class independent quantities based on different propensities toward false negative (FN) and false positive (FP) examples.

		predicted	
		$\mathcal{Y}_0$	$\mathcal{Y}_1$
true	$\mathcal{Y}_0$	TN	FP
	$\mathcal{Y}_1$	FN	TP

1. Model assessment

- Choice of the measure of accuracy
  Common error measures conduct to misleading results
  $\longrightarrow$ use of class independent quantities based on different propensity towards false negative (FN) and false positive (FP) examples.

		predicted	
		$\mathcal{Y}_0$	$\mathcal{Y}_1$
true	$\mathcal{Y}_0$	TN	FP
	$\mathcal{Y}_1$	FN	TP

  - precision, recall, F
  - ROC curve

1. Model assessment

- Choice of the measure of accuracy
- ROC curve
  classification is usually based on the following rule:
  - assign $\mathbf{x}$ to $\mathcal{Y}_1$ if

$$\frac{\hat{P}(\mathcal{Y}_1|\mathbf{x})}{\hat{P}(\mathcal{Y}_0|\mathbf{x})} > t, \text{ for some } t,$$

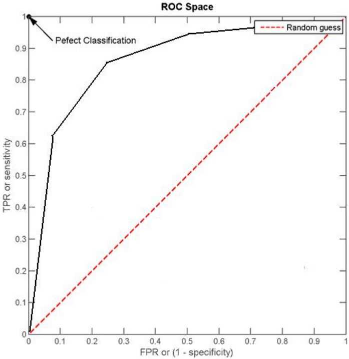
    that is

$$\hat{P}(\mathcal{Y}_1|\mathbf{x}) > t', \ t' \in (0,1).$$

  - $\forall t' \in (0,1)$:
  - assign the label class to every test point
  - compute the confusion matrix
  - plot the TP rate vs the FP rate
  - the area under the curve (AUC) summarizes the ROC behaviour.

1. Model assessment

- Choice of the measure of accuracy
- ROC curve



1. Model assessment

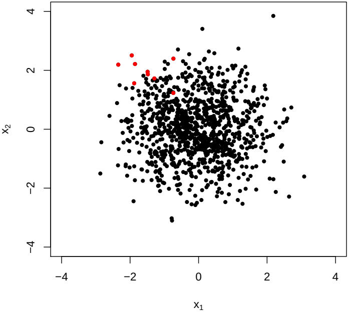
- Estimate of the accuracy

  Not enough examples from the rare class for both training and testing the classifier.

  The scarcity of data conducts to high variance estimates of the error rate, especially for the rare class.

1. Model assessment

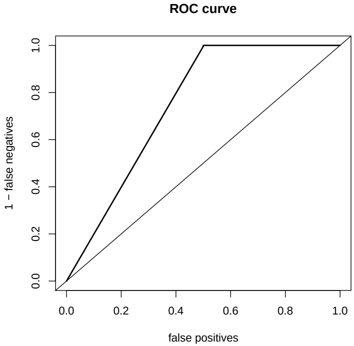
- Estimate of the accuracy
- Example



- split the data into a training set and a test set
- estimate a classification model based on the training set

1. Model assessment

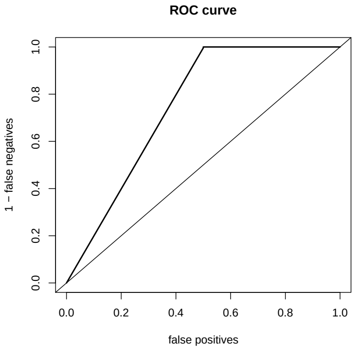
- Estimate of the accuracy
- Example



- obtain an estimate of the AUC based on the test set ($\hat{AUC} = 0.749$)

1. Model assessment

- Estimate of the accuracy
- Example

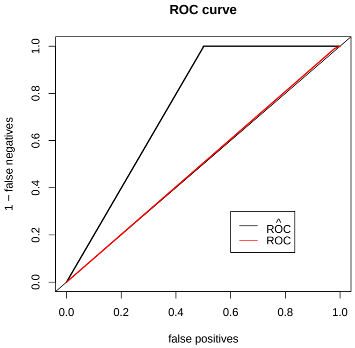


- obtain an estimate of the AUC based on the test set ($\hat{AUC} = 0.749$)
- obtain the distribution of the AUC estimator based on a large number of samples

MEAN($\hat{AUC}$)	0.673
SE($\hat{AUC}$)	0.227

1. Model assessment

- Estimate of the accuracy
- Example



- obtain an estimate of the AUC based on the test set ($\hat{AUC} = 0.749$)
- obtain the distribution of the AUC estimator based on a large number of samples

MEAN($\hat{AUC}$)	0.673
SE($\hat{AUC}$)	0.227

- approximate the actual AUC based on a huge test set $AUC = 0.505$

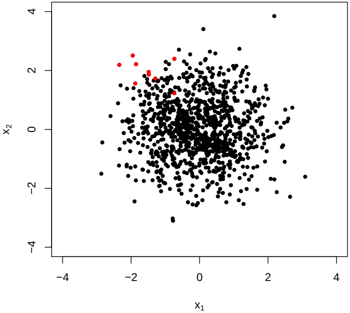
2. Model estimation

- Some methods estimate the classification rule that best fits the data according to some criterion of global accuracy

  $\longrightarrow$ when data are unbalanced the model tends to focus on the prevalent class and ignore the rare events.

2. Model estimation

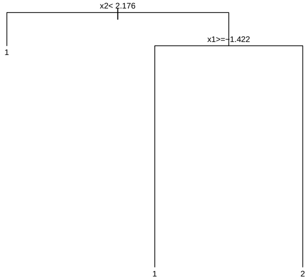
- Some methods estimate the classification rule that best fits the data according to some criterion of global accuracy
- Example

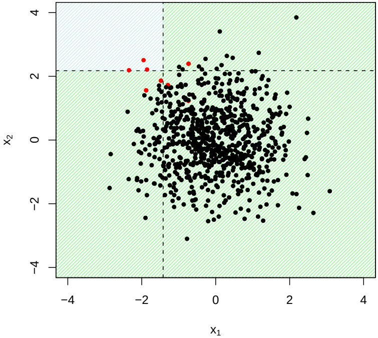


- split the data into a training set and a test set
- estimate a classification model based on the training set

2. Model estimation

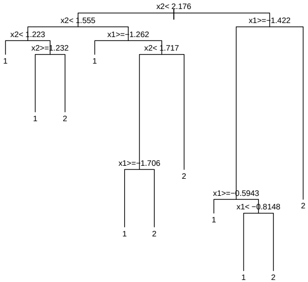
- Some methods estimate the classification rule that best fits the data according to some criterion of global accuracy
- Example

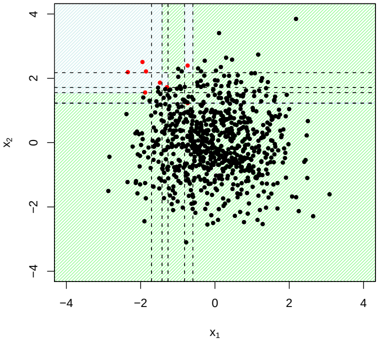




2. Model estimation

- Some methods estimate the classification rule that best fits the data according to some criterion of global accuracy
- Example





2. Model estimation

- Some methods estimate the classification rule that best fits the data according to some criterion of global accuracy (classification trees, $k-$nearest neighbors classifiers, SVM...)
- Other methods share different problems in estimating the classifier:
    - The conditional probabilities of the rare class are underestimated when using logistic regression (King and Zeng, 2001)
    - Problems arise in estimating the common covariance matrix of the two classes in linear discriminant analysis (Hand and Vinciotti, 2003)
    - ...

2. Model estimation

- Some solutions:
    - strengthening the process of learning with regards to the rare class.
        - optimization of a target function taking into account the skewed distribution of the classes
          Cieslak and Chawla (2008) build decision trees by using a skewed splitting criterion instead of the traditional Gini index.
        - use of a weighted-based distance criterion in $k-$nearest neighbors (Barandela et al., 2003)
        - minimization of the expected misclassification cost instead of the misclassification error (Lin et al. 2002).
        - ...

2. Model estimation

- Some solutions:
  - rebalancing the class distribution by some form of resampling (Japkovicz and Stephen, 2002).
    - random oversampling the rare class with replacement
    - random undersampling the prevalent class
    - combinations of over/undersampling techniques.
  - Increasing attention is given to the novel strategy of generating new artificial examples which are "similar" in some sense to the training data (Lee, 1999, 2000; Chawla *et al.*, 2002).

Summing-up

- when the distribution of the classes is skewed, the performance of classification models is comprehensively compromised but, even worst, poor-quality estimates of the chosen accuracy measure may preclude understanding the limits of the learning process.
- the problems of building an accurate classifier and assessing its performance should not be dealt with separately.

What to do?

- a simultaneous treatment of these two inseparable problems has not been considered yet.

↓

- a unified and systematic framework for dealing with the issue of class imbalance is proposed
- the proposed solution is based on the generation of new artificial data from the classes, according to a smoothed bootstrap approach

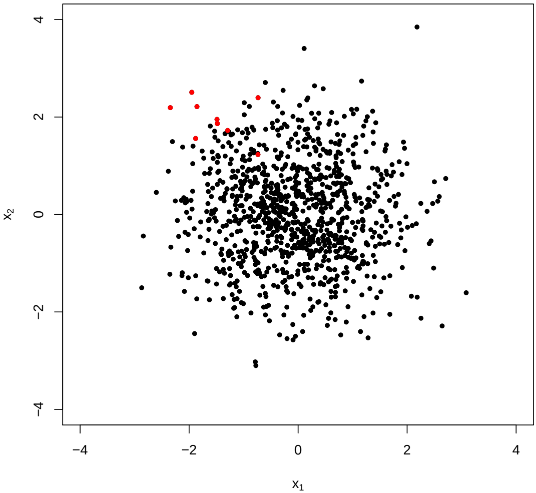
ROSE: Random Over Sampling Examples

1. select $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_j}\}, \mathbf{x}_i \in \mathcal{Y}_j$ with probability $P(\mathbf{x}_i) = \frac{1}{n_j}$, with n_j size of class $\mathcal{Y}_j$
2. sample $\mathbf{x}$ from $K_{H_j}(\mathbf{x}_i)$, with K_{H_j} local density of $\mathbf{x}_i$ ($e.g. K_{H_j}$ Gaussian distribution centered at $\mathbf{x}_i$ and H_j a scale parameter)

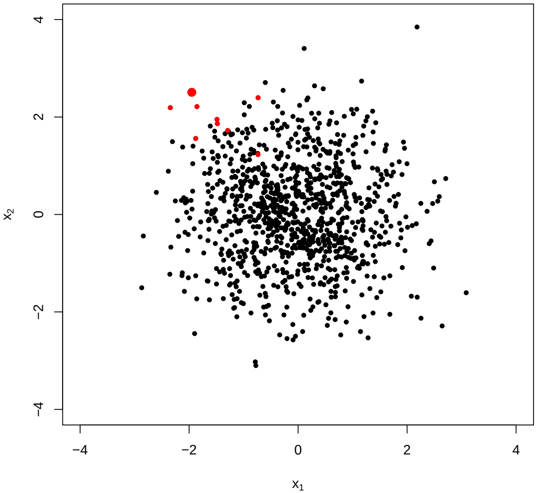
$$\begin{aligned}\hat{f}(\mathbf{x}|\mathcal{Y}_j) &= \sum_{i=1}^{n_j} P(\mathbf{x}_i) Pr(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} Pr(\mathbf{x}|\mathbf{x}_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{\mathbf{H}_j}(\mathbf{x} - \mathbf{x}_i)\end{aligned}$$

kernel density estimates of $f(\mathbf{x}|\mathcal{Y}_j)$.

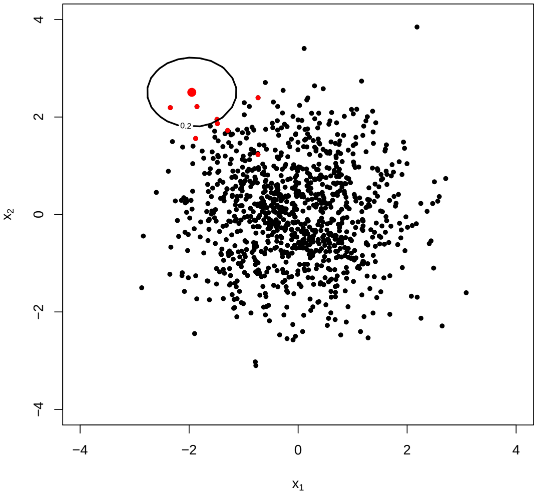
ROSE: a toy example



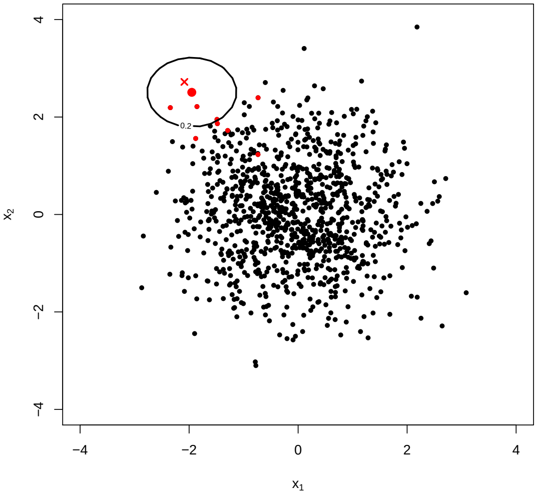
ROSE: a toy example



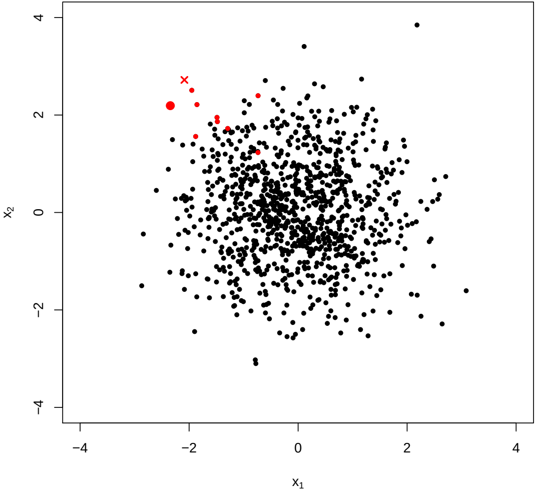
ROSE: a toy example



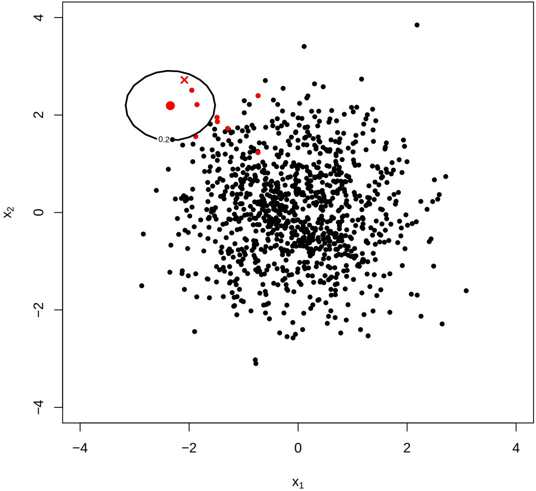
ROSE: a toy example



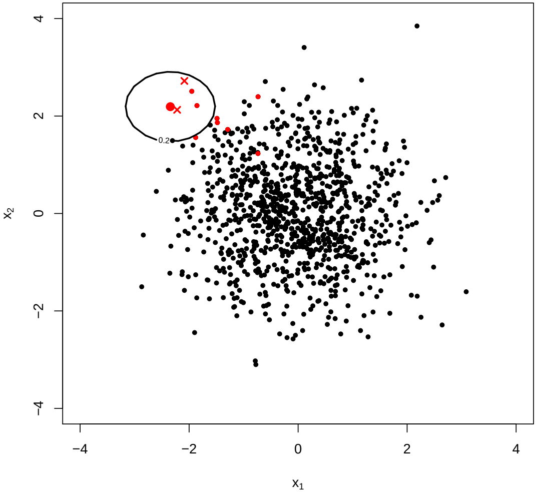
ROSE: a toy example



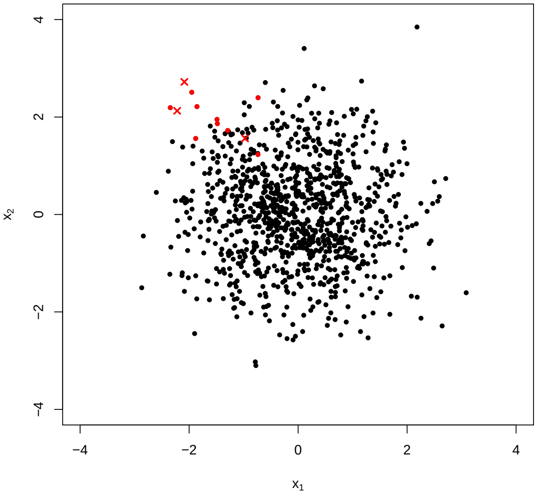
ROSE: a toy example



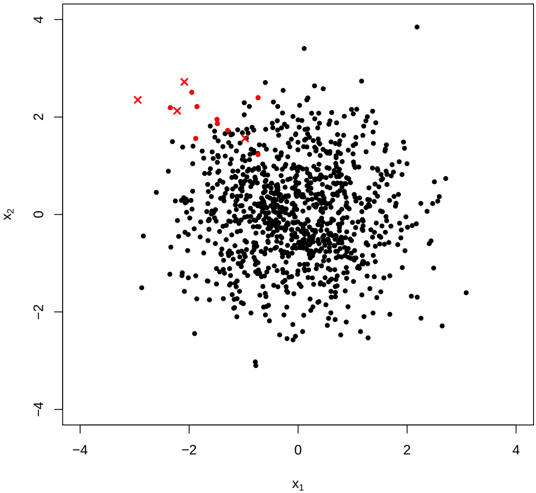
ROSE: a toy example



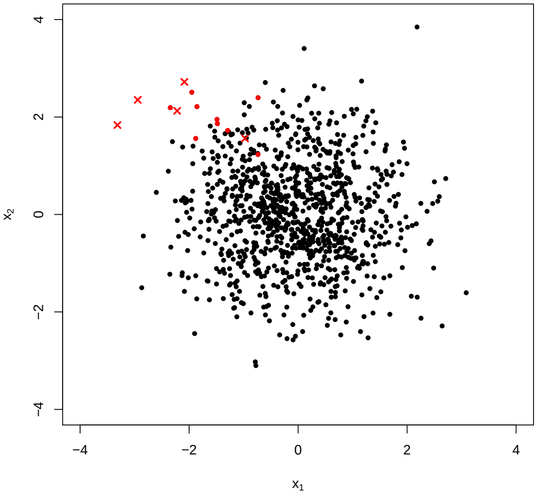
ROSE: a toy example



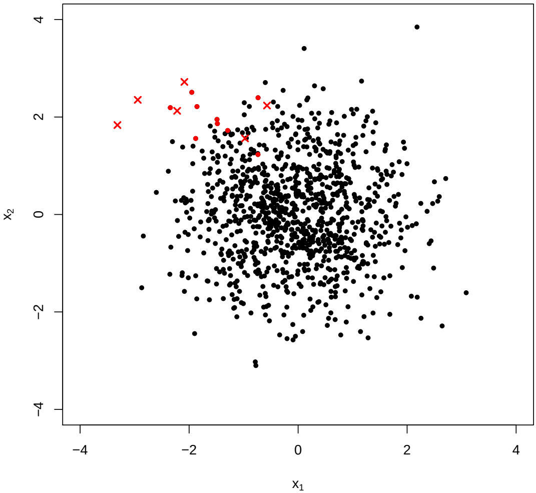
ROSE: a toy example



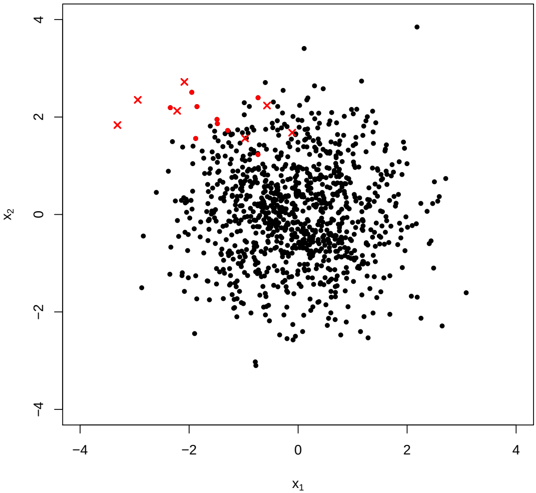
ROSE: a toy example



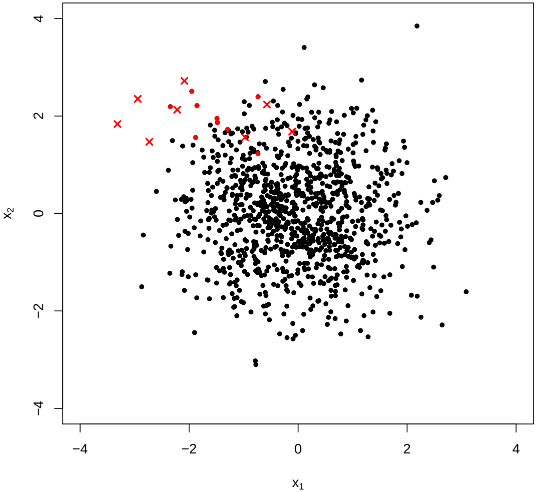
ROSE: a toy example



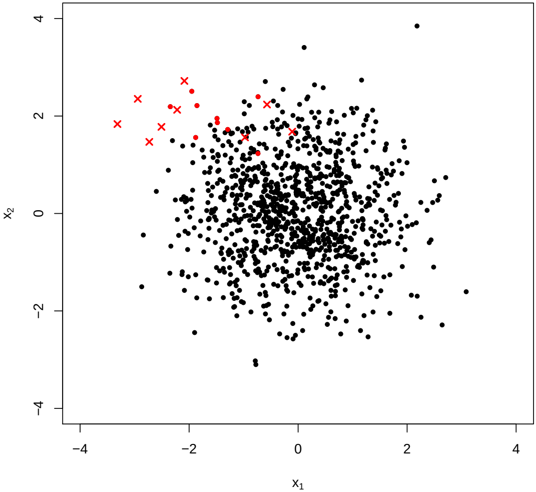
ROSE: a toy example



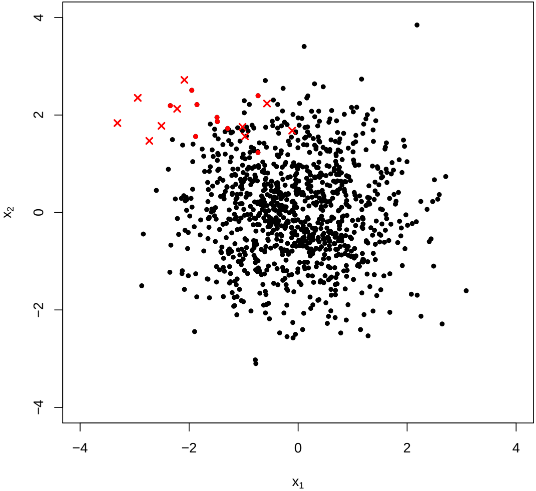
ROSE: a toy example



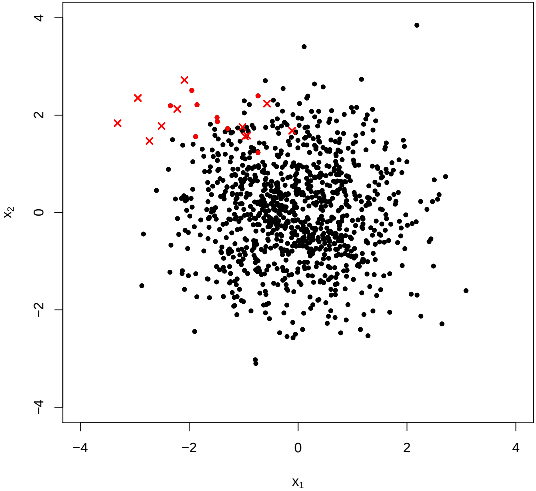
ROSE: a toy example



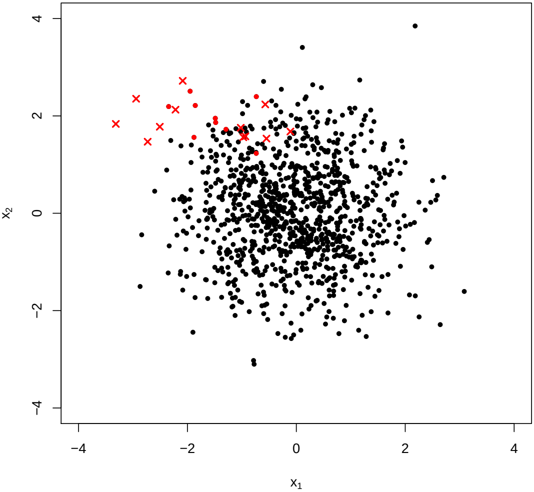
ROSE: a toy example



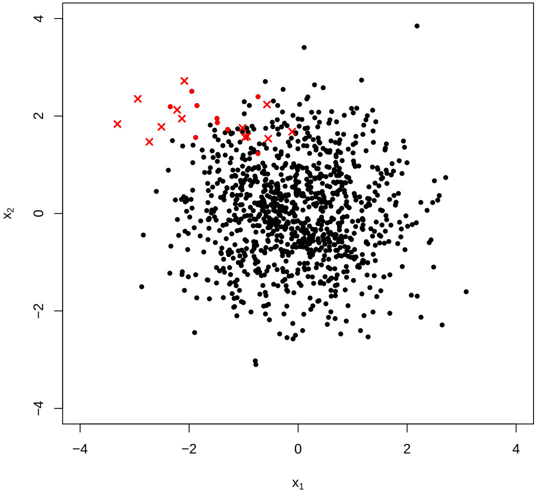
ROSE: a toy example



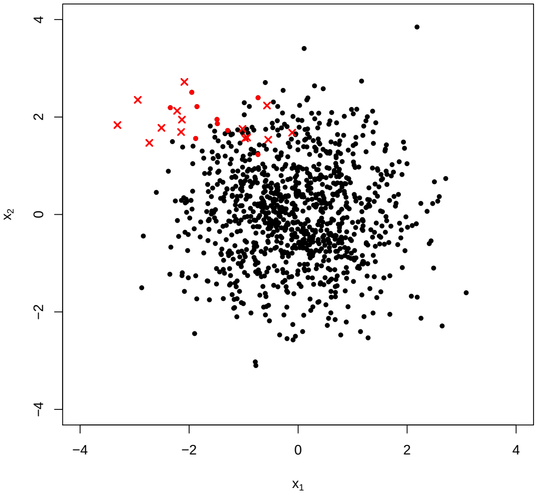
ROSE: a toy example



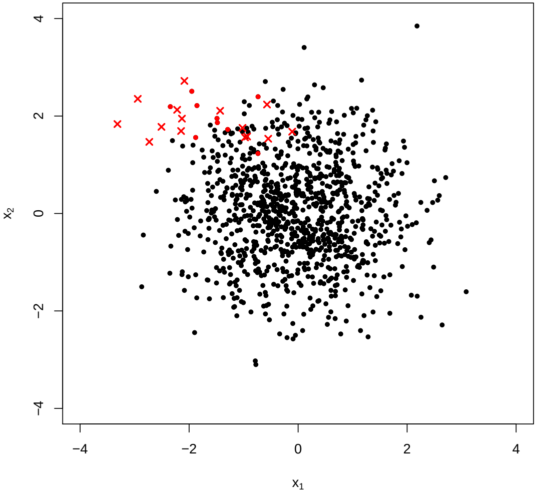
ROSE: a toy example



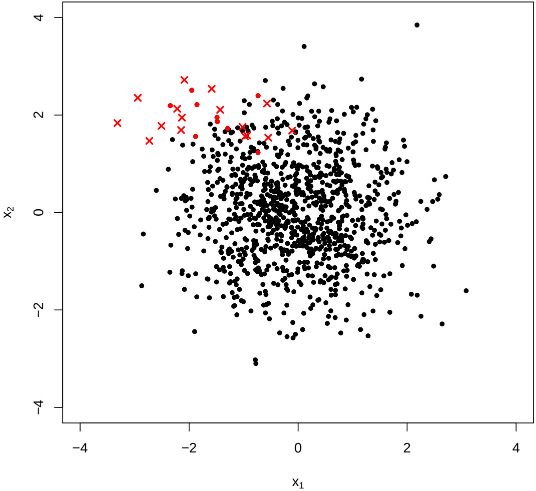
ROSE: a toy example



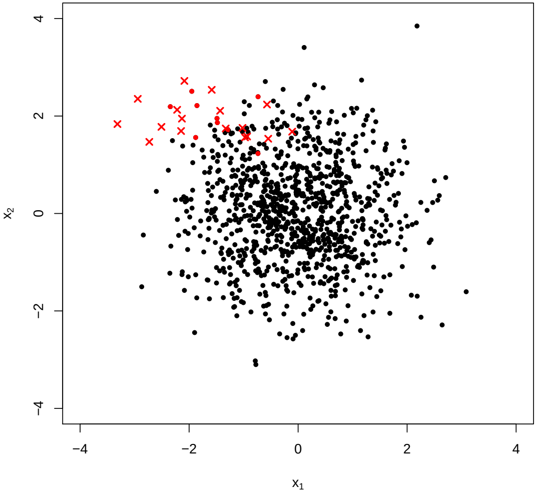
ROSE: a toy example



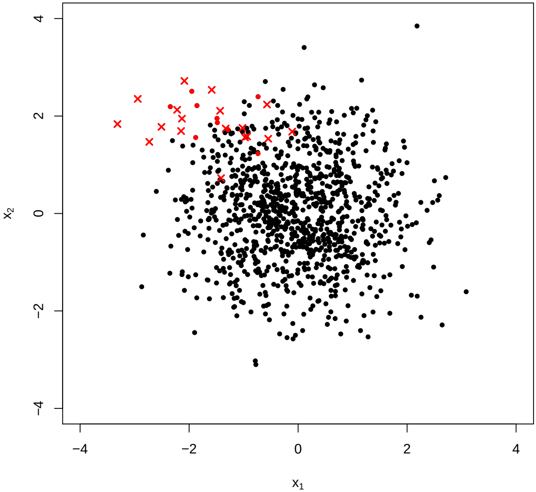
ROSE: a toy example



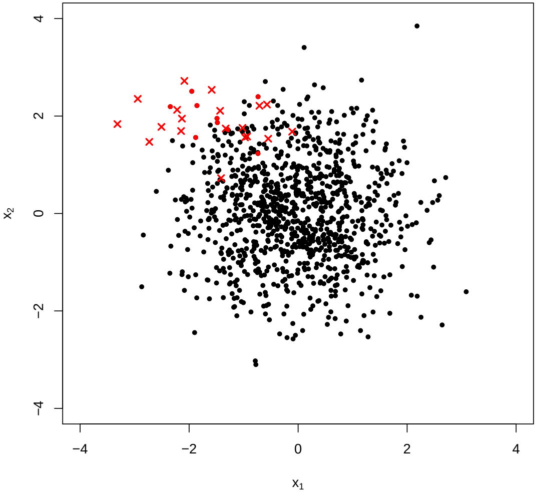
ROSE: a toy example



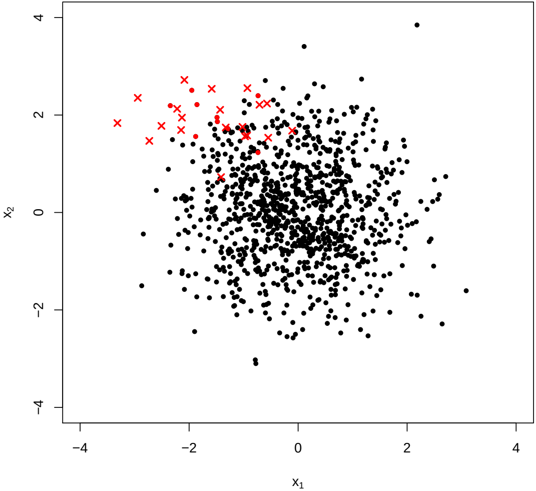
ROSE: a toy example



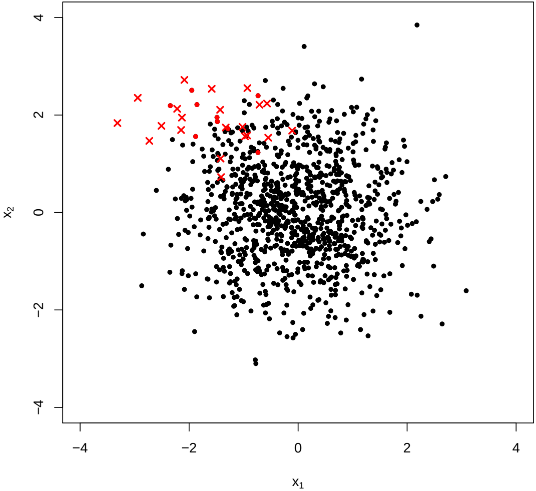
ROSE: a toy example



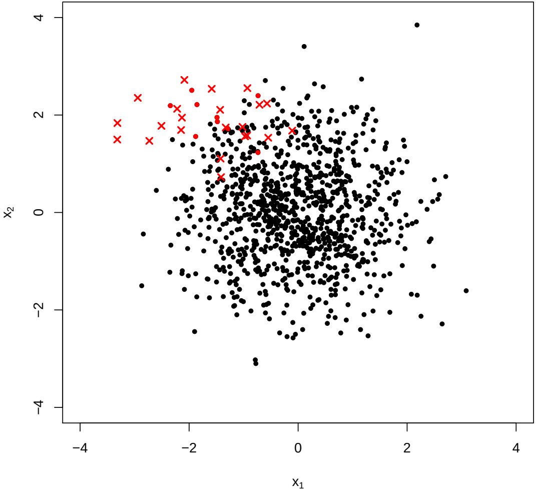
ROSE: a toy example



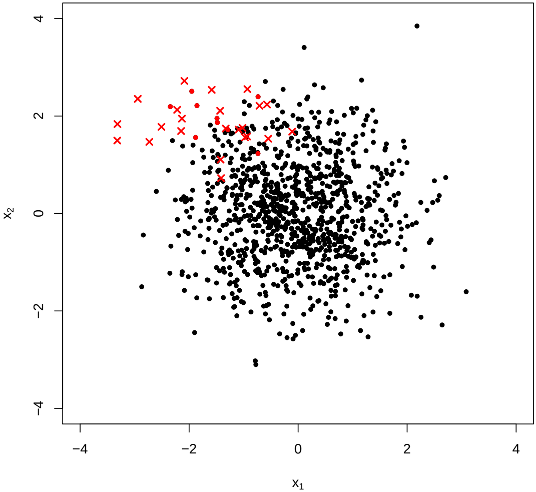
ROSE: a toy example



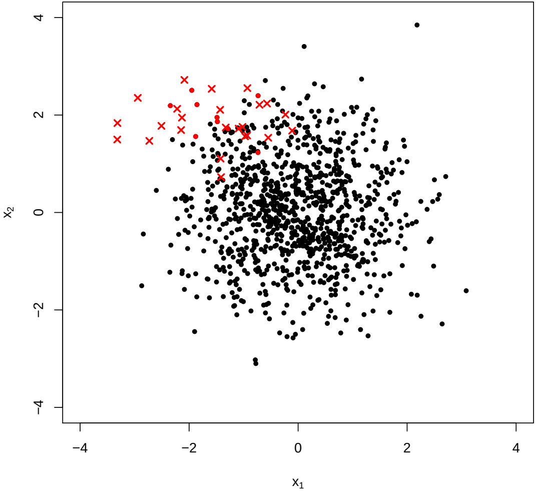
ROSE: a toy example



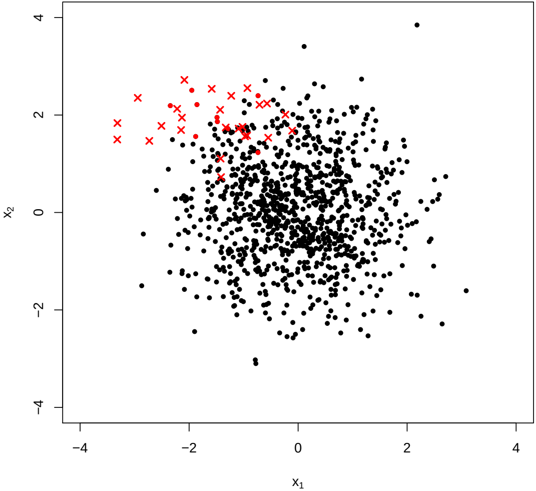
ROSE: a toy example



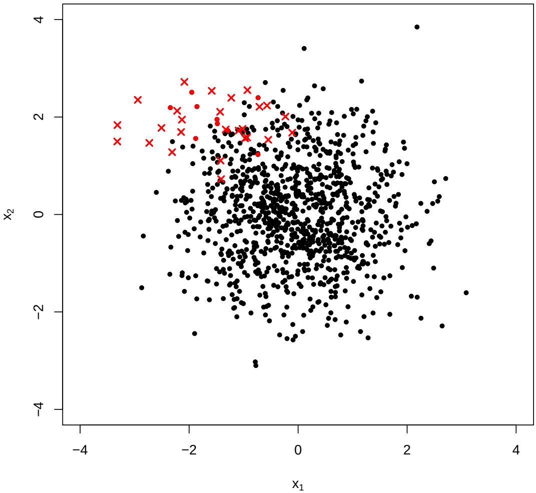
ROSE: a toy example



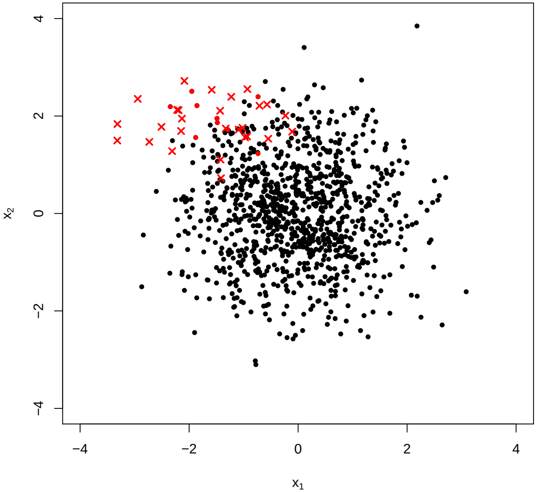
ROSE: a toy example



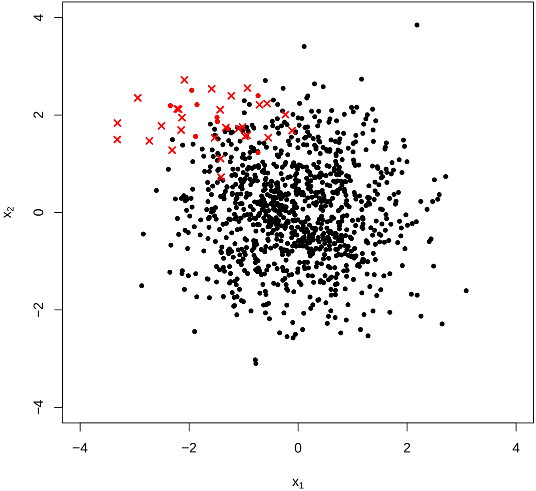
ROSE: a toy example



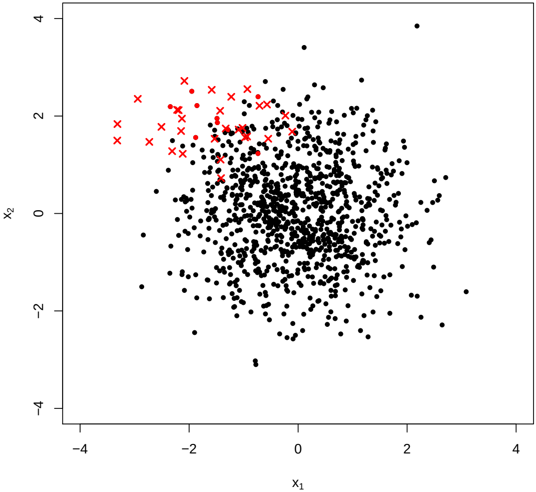
ROSE: a toy example



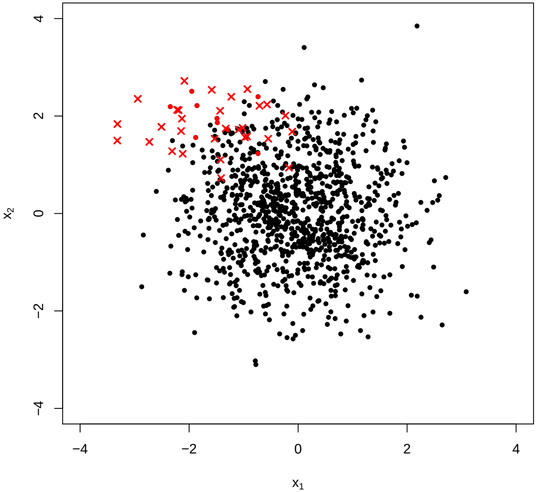
ROSE: a toy example



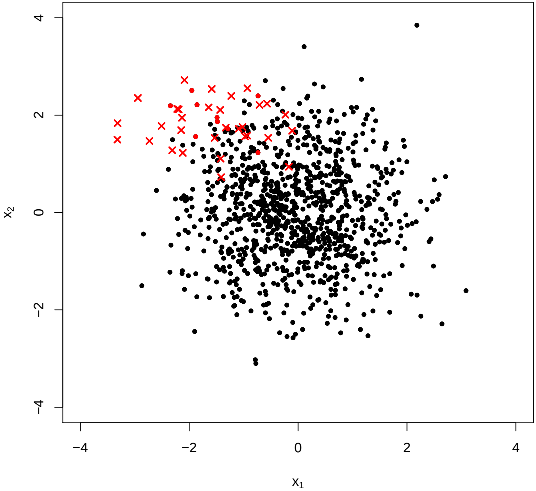
ROSE: a toy example



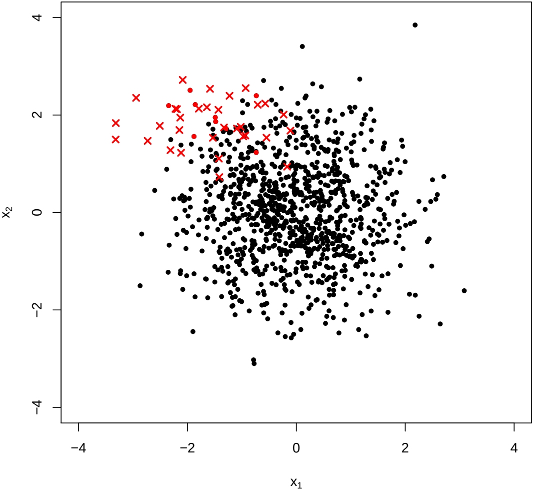
ROSE: a toy example



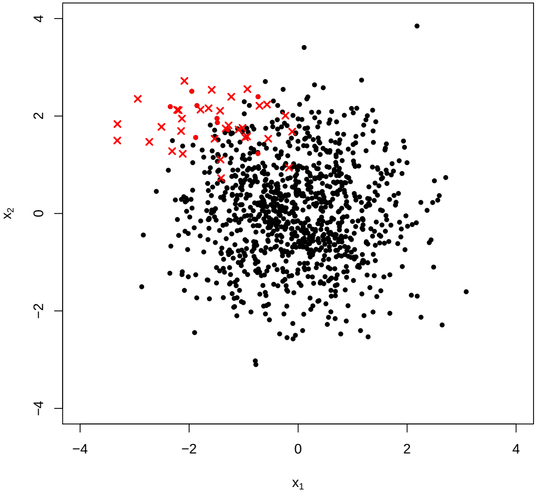
ROSE: a toy example



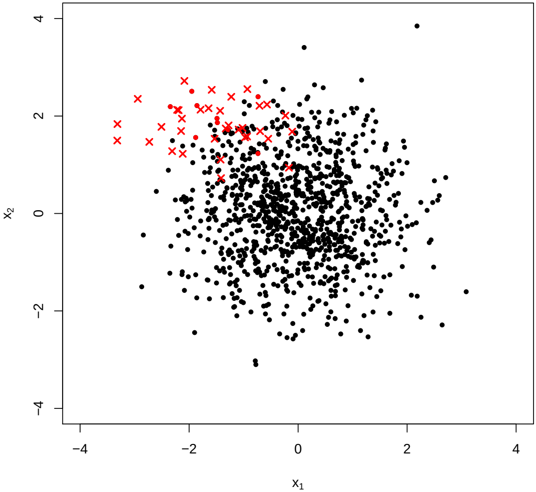
ROSE: a toy example



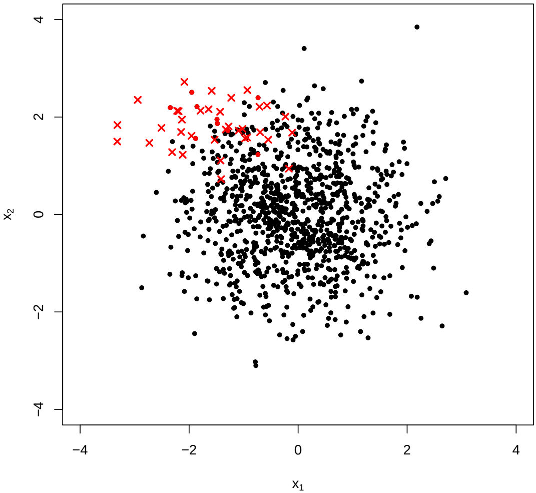
ROSE: a toy example



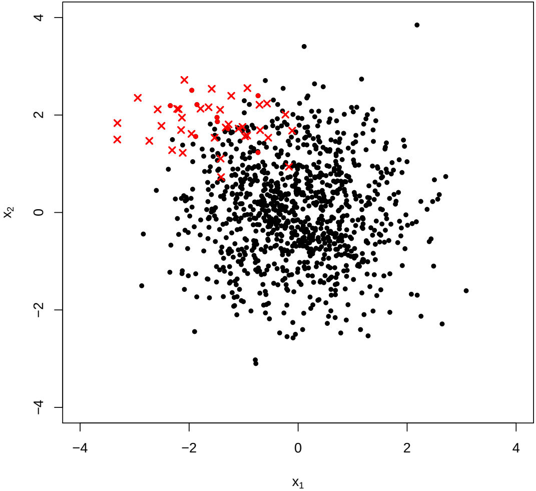
ROSE: a toy example



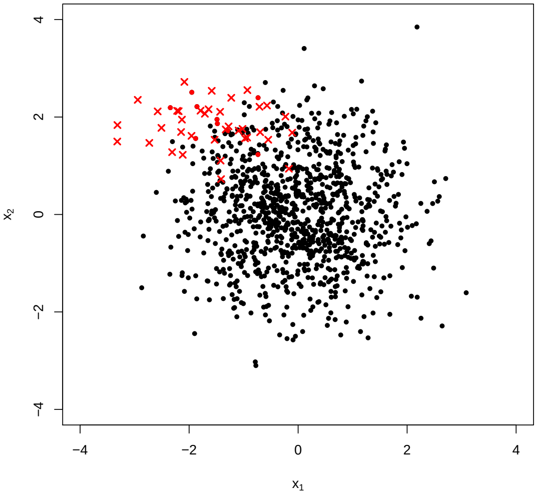
ROSE: a toy example



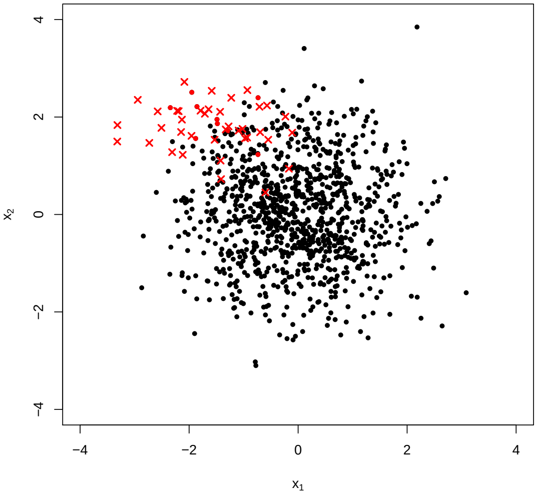
ROSE: a toy example



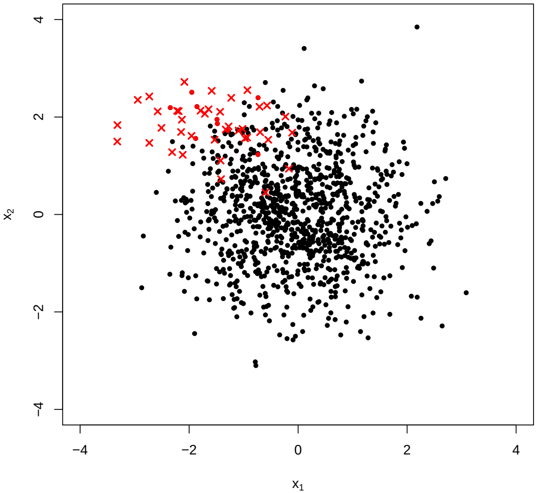
ROSE: a toy example



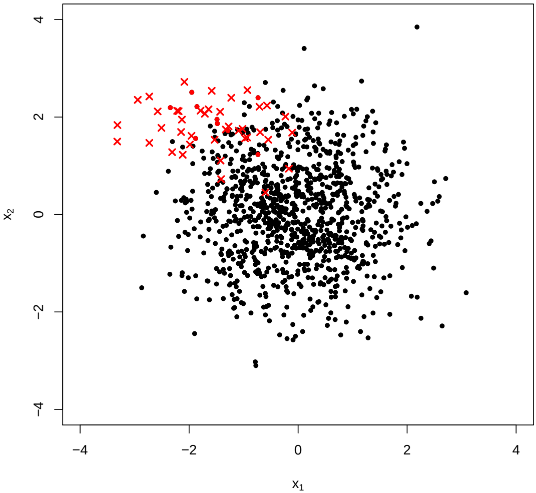
ROSE: a toy example



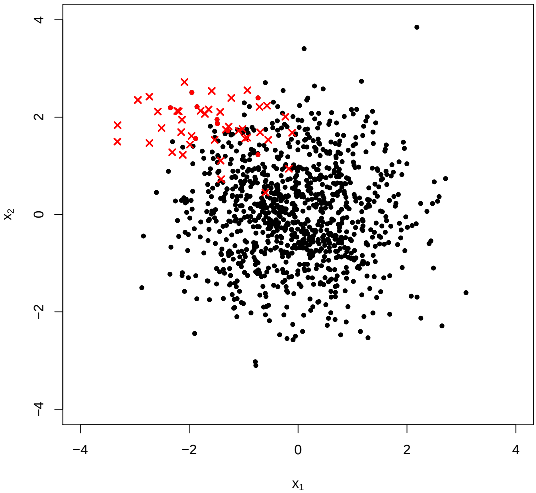
ROSE: a toy example



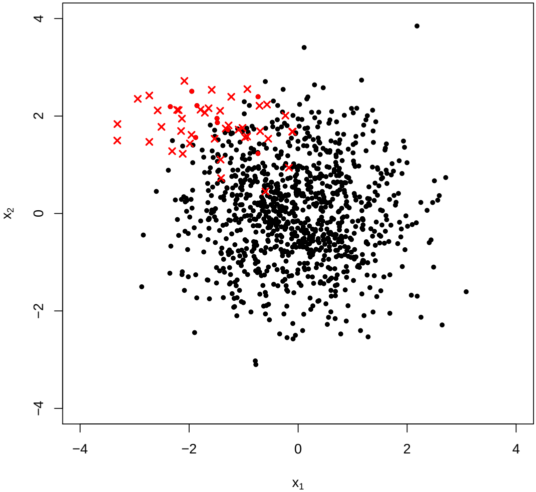
ROSE: a toy example



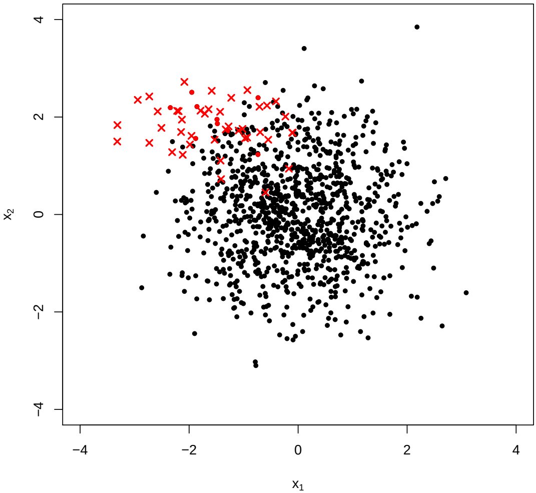
ROSE: a toy example



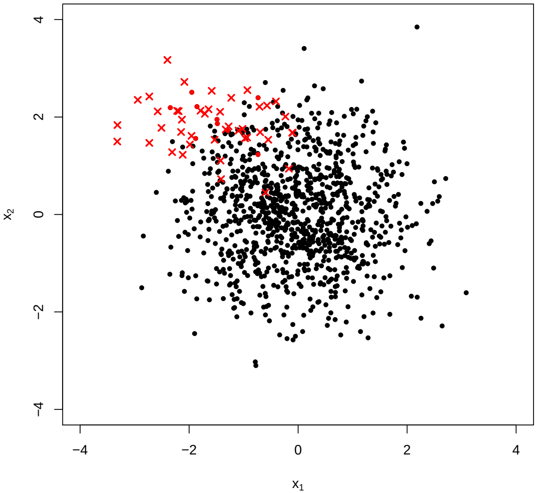
ROSE: a toy example



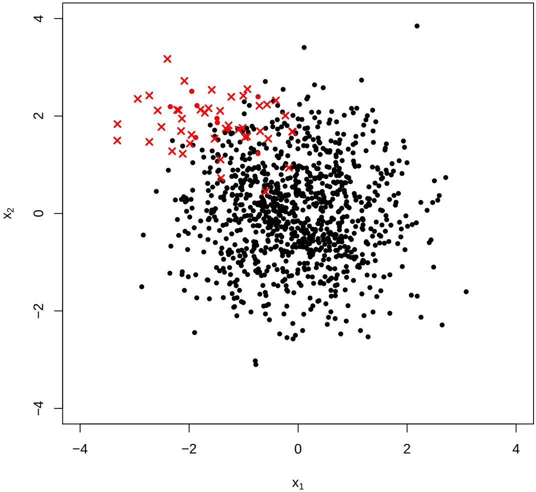
ROSE: a toy example



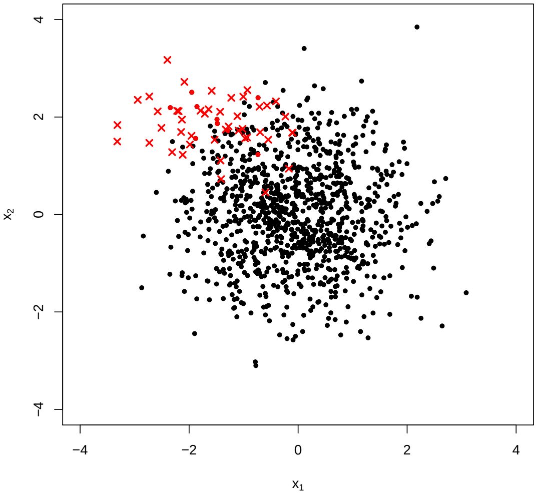
ROSE: a toy example



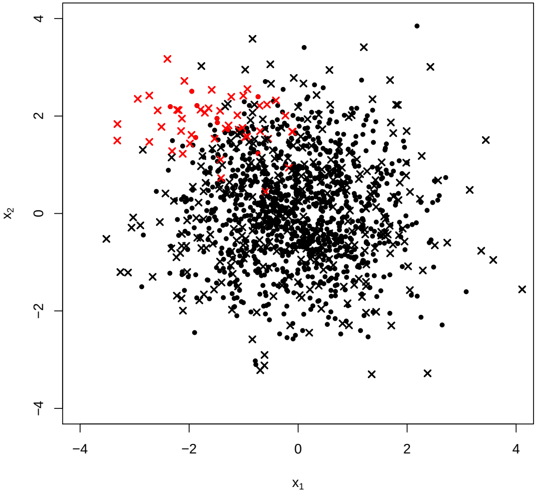
ROSE: a toy example



ROSE: a toy example



ROSE: a toy example



How to use ROSE?

- the synthetic generation of new examples allows for strengthening the process of learning as well as estimating the distribution of the chosen measure of accuracy.
- an augmented sample of data (especially belonging to the rare class) helps in estimating the model, because the classifier will address the same attention to both the classes.
- the originally observed data may be used to assess the model's accuracy.
- smoothed bootstrap methods aid estimating the error distribution.

An application to credit scoring

- GOAL: building an as accurate as possible rule to separate defaulter enterprises from non defaulter ones.
- DATA: vital statistics, balance sheet records and financial ratios of all the commercial business enrolled to the Business Register and located in a province of the North East Italy.
- the bankruptcy condition as the default event (occurring the 0.7% of the cases)

An application to clinical data

- GOAL: distinguish patients suffering from diabetes from healthy patients.
- DATA: physical and clinical measurements of 768 females of Pima Indians, a population in which a high incidence of diabetes has been historically reported (UCI repository of machine learning).
- The response variable is the positive or negative result from a diabetes test.
- the distribution of the classes has been made unbalanced by considering only a proportion of 1% rare cases, randomly selected from the class of diabetic patients.

How to proceed

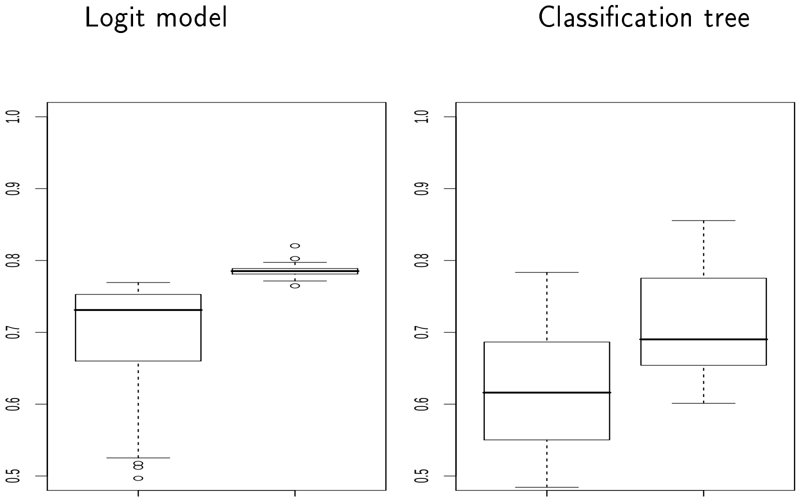
- select B ROSE samples T_n^R from the data
- build the classification rule on each T_n^R
- evaluate the distribution of the AUC (area under the ROC curve) on the original data

- use of logit models and classification trees

- benchmark: AUC distribution deriving from (unbalanced) smoothed bootstrap samples

Distribution of the AUCs

Credit scoring data

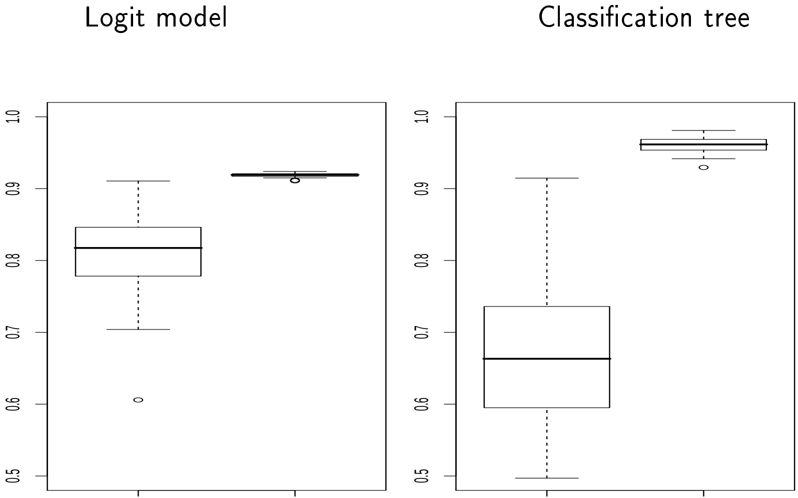


Logit model

Classification tree

Distribution of the AUCs

Diabetes data



Logit model

Classification tree

Final remarks

- ROSE provides an unified and systematic approach for dealing with rare classes, supported by the good properties of kernel methods.
- It aims at producing new balanced data to be used both for model building and evaluating its accuracy.
- ROSE includes most of existing resampling methods as a special case.
- It has the advantage of both reducing the risk of overfitting and increasing the classifier ability of generalization.
- Significant improvements of accuracy due to the data generation have resulted from the application to real and simulated data.

References

[1] Barandela, R., Sánchez, J. S., García, V., Rangel, E. Strategies for learning in class imbalance problems. *Pattern Recognition*, vol. 36, pp. 849-851, 2003.

[2] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, P. (2002), Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence 16, 321-357.

[3] Cieslak, D., Chawla, N. (2008), Learning decision trees for unbalanced data. *Lecture Notes in Computer Science*, 5211, 241-256.

[4] Hand, D.J. and Vinciotti, V. (2003). Choosing K for Two-Class Nearest Neighbour Classifiers with Unbalanced Classes, *Pattern Recognition Letters*, vol. 24, pp. 1555-1562,.

[5] Japkowicz, N., Stephen, S. (2002), The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis Journal*, 6.

[6] King, G. and Zeng, L., Logistic regression in rare events data. *Political Analysis*, vol. 9, pp. 137-163, 2001.

[7] Lee, S. Noisy replication in skewed binary classification. *Computational Statistics and Data Analysis*, vol. 34, pp. 165-191, 2000.

[8] Lee, S. Regularization in skewed binary classification. *Computational Statistics*, vol. 14, pp. 277-292, 1999.

[9] Lin, Y., Lee, Y., Wahba, G.. Support Vector Machines for Classification in Nonstandard Situations, *Machine Learning*, vol. 46, pp. 191-202, 2002.

[10] Silverman, B.W. (1986), *Density estimation for statistics and data analysis*, Chapman & Hall, London.