Outline: Unbalanced Supervised Problems (F. Chiaromonte)

Statistical Methods for Large, Complex Data

A reference article on subsampling in unbalanced classification problems:

The Annals of Statistics
2014, Vol. 42, No. 5, 1693–1724
DOI: 10.1214/14-AOS1220

LOCAL CASE-CONTROL SAMPLING: EFFICIENT SUBSAMPLING IN IMBALANCED DATA SETS

BY WILLIAM FITHIAN[1] AND TREVOR HASTIE[2]

Stanford University

For classification problems with significant class imbalance, subsampling can reduce computational costs at the price of inflated variance in estimating model parameters. We propose a method for subsampling efficiently for logistic regression by adjusting the class balance locally in feature space via an accept–reject scheme. Our method generalizes standard case-control sampling, using a pilot estimate to preferentially select examples whose responses are conditionally rare given their features. The biased subsampling is corrected by a post-hoc analytic adjustment to the parameters. The method is simple and requires one parallelizable scan over the full data set.

Standard case-control sampling is inconsistent under model misspecification for the population risk-minimizing coefficients θ^*. By contrast, our estimator is consistent for θ^* provided that the pilot estimate is. Moreover, under correct specification and with a consistent, independent pilot estimate, our estimator has exactly twice the asymptotic variance of the full-sample MLE—even if the selected subsample comprises a miniscule fraction of the full data set, as happens when the original data are severely imbalanced. The factor of two improves to $1 + \frac{1}{c}$ if we multiply the baseline acceptance probabilities by $c > 1$ (and weight points with acceptance probability greater than 1), taking roughly $\frac{1+c}{2}$ times as many data points into the subsample. Experiments on simulated and real data show that our method can substantially outperform standard case-control subsampling.

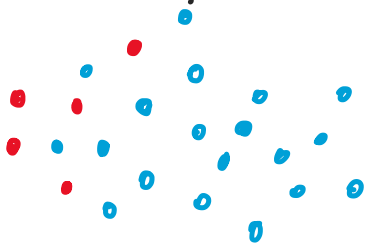
Some more references

Dubey, Rashmi et al. “Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study.” *NeuroImage* vol. 87 (2014): 220-41. doi:10.1016/j.neuroimage.2013.10.005

Menardi, Torrelli (2010) “Training and assessing classification rules with unbalanced data”. Working Paper 2-2010. Dipartimento B. De Finetti, Universita’ di Trieste.

Imbalanced Learning tools, MIT (includes SMOTE)
https://imbalanced-learn.org/stable/index.html

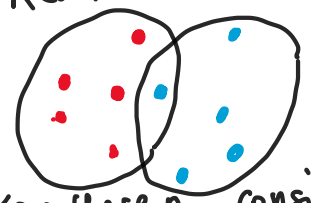
The data in feature space



"scarce" class n_R

"abundant" class n_B

(1) Reduce the abundant class



Keep these n_R fixed

consider only n_R of these at random

... can "repeat"

Form one, or several, datasets of size $2n_R < n_R + n_B$.

The red points are always the same.

(1B) A variant

Bootstrap the red points (resampling with replacement)

Sub-bootstrap the blue points (Select n_R blue points at random with replacement)

... can "repeat" both

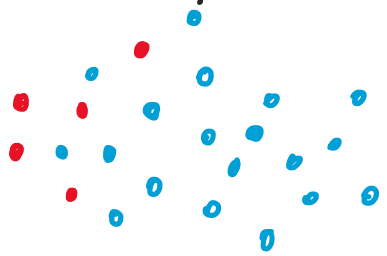
Form one, or several, datasets of size $2n_R < n_R + n_B$

The red points are bootstrapped too.

... MAKES MORE SENSE STATISTICALLY

we simulate sampling from the two populations.

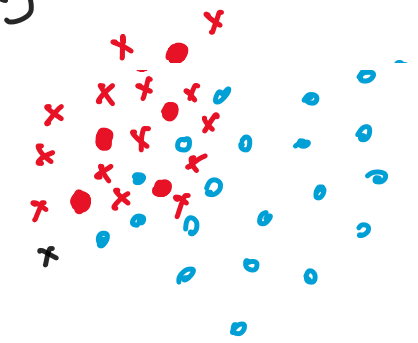
The data in feature space



"scarce" class n_R

"abundant" class n_B

② Augment the scarce class



• = actual red points

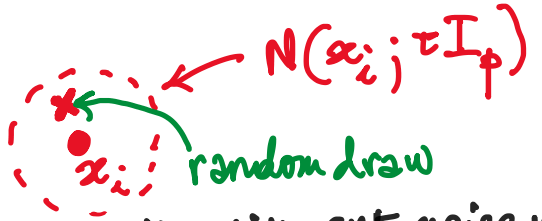
x = artificial red points

Form a dataset of size

$2n_B > n_R + n_B$

How do we create the artificial points? some options

(i) Over-bootstrap the red points (select n_B red points at random with replacement) ADDING NOISE to each draw



(can specify different noise models)

(ii) n_B times over: select at random two red points and a point between them

(can "localize" the selection of the red points, e.g., first at random, second at random among its closest red neighbors)

(2B) A variant

Focus the augmentation of the scarce class where it is harder to discriminate, i.e., close to the blue points.

... MAKES MORE SENSE STATISTICALLY