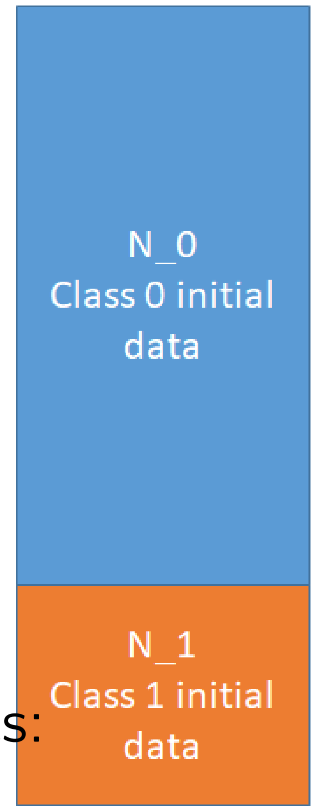
Statistical Learning for Data Mining
*

Class Imbalance

*George Runger 2019

Class Imbalance *



- Data sets can differ dramatically in the counts of instances in classes

- For example:
  Among 1000 patients, 900 *normal*, 100 *ill*
  Among 10000 manufactured parts: 9900 *conform*, 70 *marginal*, 30 *fail to conform*

- Accuracy scores can mislead us when class imbalance is present

* *George Runger 2019*

Class Imbalance *

- For example, in the table below, if all parts are classified as Class 0, regardless of predictor attribute values, accuracy is 90%

	Predicted	
Actual	Class 1	Class 0
Class 1	0	100
Class 0	0	900

- For unbalanced data, another measure is **balanced error rate (BER)**
  $BER = (FPR + FNR)/2$
  For the table above
  $BER = (0 + 1)/2 = 0.5$

- Two common approaches to handle class imbalance are: **weight instances** or **adjust training data**

*George Runger 2018

Class Imbalance *

- Common to weight instances from a class inversely proportional to the class proportion
  - Weight instances so that the weight totals for each class are equal

- For example:
  Class 0 800 instances, Class 1 200 instances
  Class 0 weights = 1, Class 1 weights = 4

- Could also use Class 0 weights = 2, Class 1 weights = 8

* *George Runger 2018*

Class Imbalance *

- Easiest to assign weight 1 to majority class

- Among 10000 manufactured parts: 9900 *conform*, 70 *marginal*, 30 *fail to conform*
  Class *conform* weight = 1
  Class *marginal* weight = 9900/70
  Class *fail to conform* weight = 9900/30

* *George Runger 2018*

Class Imbalance *

- Algorithms can often be adapted for weights $w_i, i = 1, 2, \ldots, N$ applied to instances

- Consider squared error loss $\sum_{i=1}^{N}(y_i - \widehat{y}_i)$
  Modify to $\sum_{i=1}^{N} w_i(y_i - \widehat{y}_i)$

- Consider log-likelihood $\sum_{i=1}^{N}$ likelihood$(\vec{x}_i, y_i)$
  Modify to $\sum_{i=1}^{N} w_i$likelihood$(\vec{x}_i, y_i)$

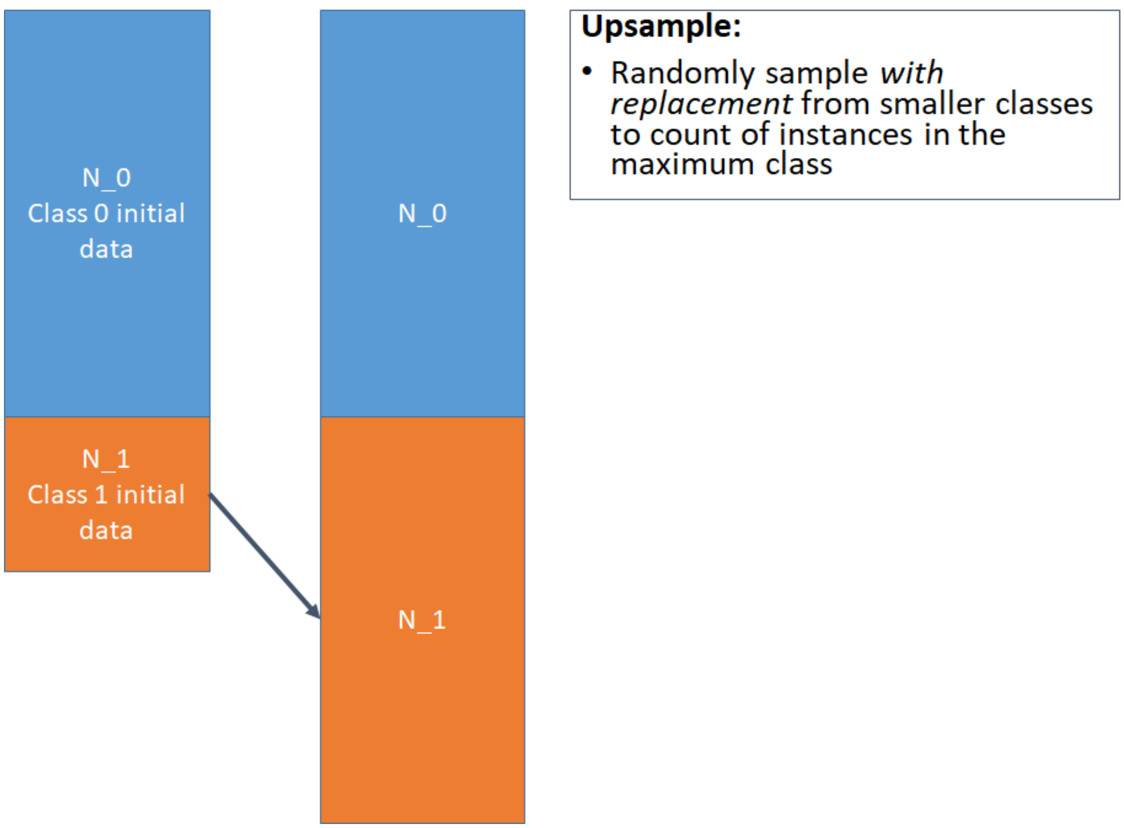
- Consider proportions for impurity indices
  For example, weights on 7 instances:
  Class 0: 1, 1, 1, 1, 1; Class 1: 2.5, 2.5
  Unweighted estimate of Class 0 proportion
  $p_0 = 5/7$
  Weighted estimate of Class 0 proportion
  $p_0 = \frac{1+1+1+1+1}{1+1+1+1+1+2.5+2.5} = 5/10$

*George Runger 2019

Class Imbalance *

- Another approach to class imbalance is to adjust the training data through sampling

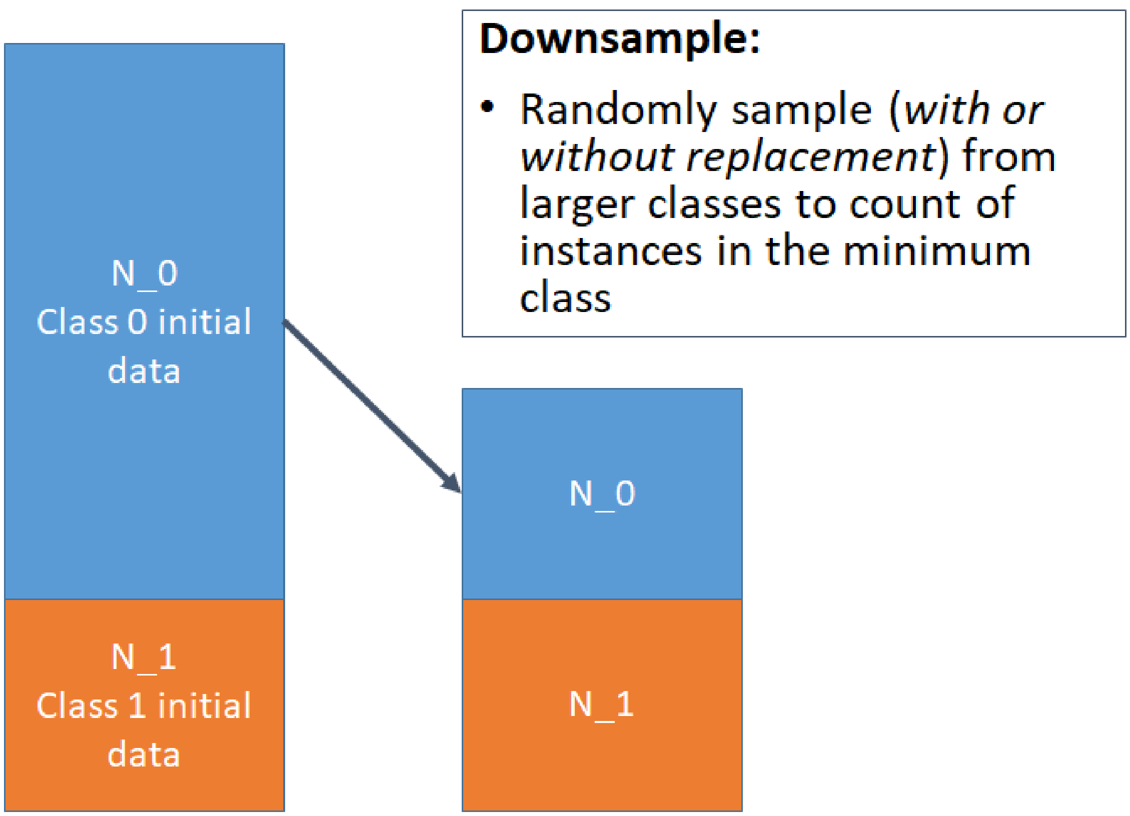
- Upsample to the count of the largest class



* *George Runger 2019*

Class Imbalance *

- Downsample to the count of the smallest class



**George Runger 2019*

Class Imbalance *

- Up and down sampling are simple approaches Can be applied to the data before models are generated

- Because of the random sampling, replicates of the samples and models are useful to evaluate the results
  - E.g., maybe 10 replicates
  - For an ensemble model, different samples can be selected for each base learner (considered a way to regularize the model)

- Models are trained with up or down sampling, but accuracy, balanced error rate, and other measures are usually evaluated from un-adjusted data

* *George Runger 2019*

Class Imbalance *

- For example, after a method to adjust for class imbalance is applied, the previous table might be changed to one such as

	Predicted	
Actual	Class 1	Class 0
Class 1	70	30
Class 0	180	720

- Accuracy is reduced to 790/1000 = 79%, but BER improved to $BER = (180/900 + 30/100)/2 = 0.25$

George Runger 2018