DATA MINING 2

Imbalanced Data and Performance Evaluation

Riccardo Guidotti

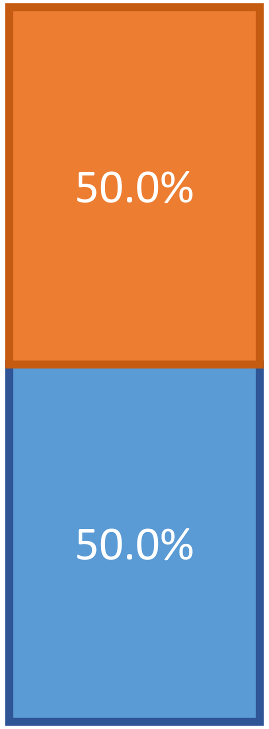
a.a. 2019/2020

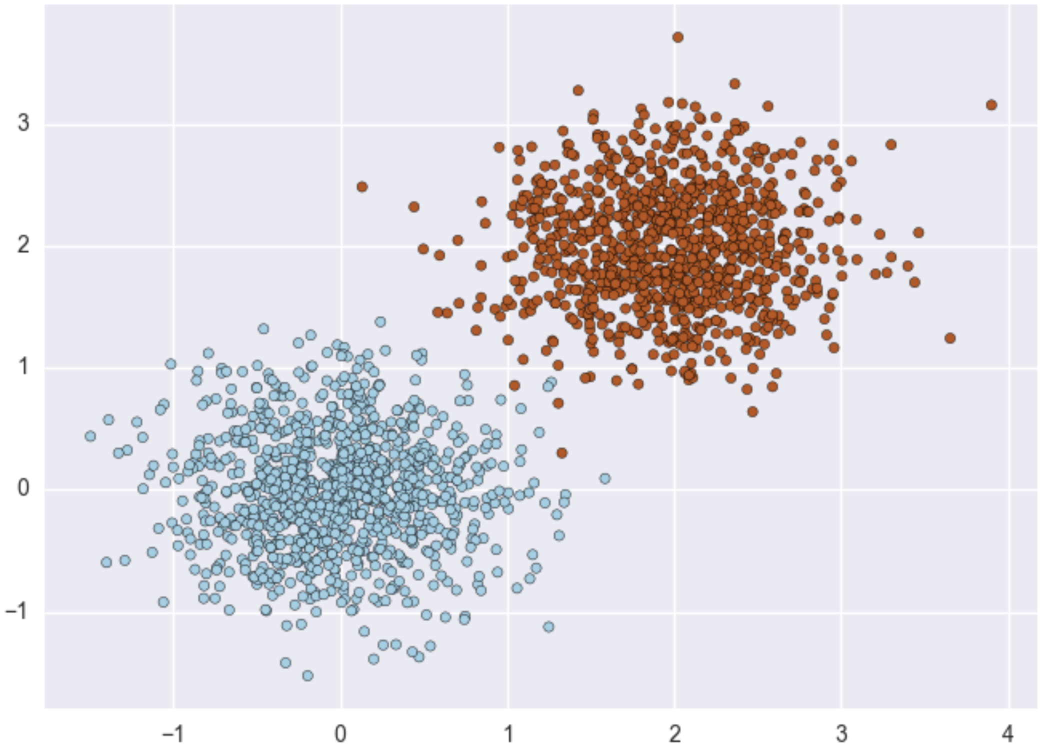
UNIVERSITÀ DI PISA

Imbalanced Data

Imbalanced Classes

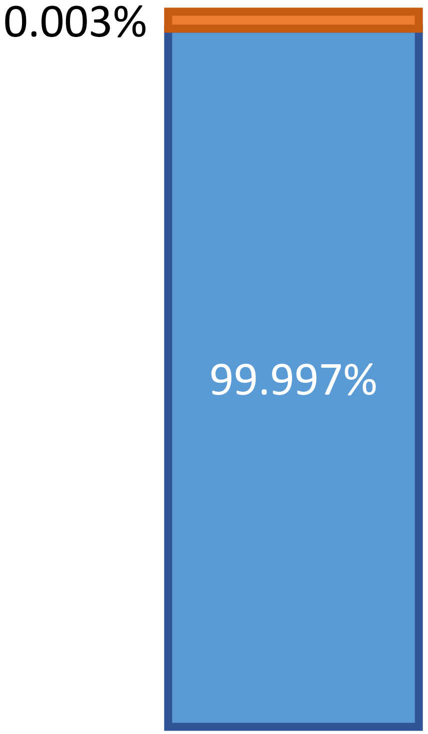
- Most classification methods assume classes are reasonably balanced.

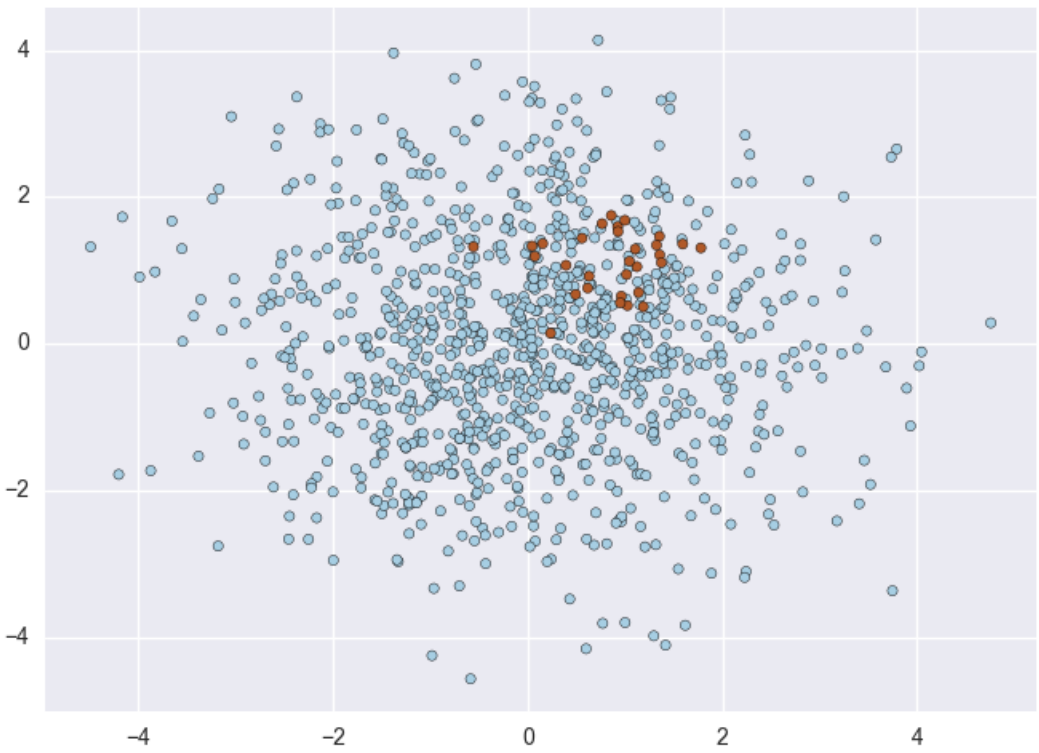




Imbalanced Classes

- In reality it is quite common to have a very popular class and a rare (yet interesting) class.





This occurs when there is a large discrepancy between the number of examples with each class label.

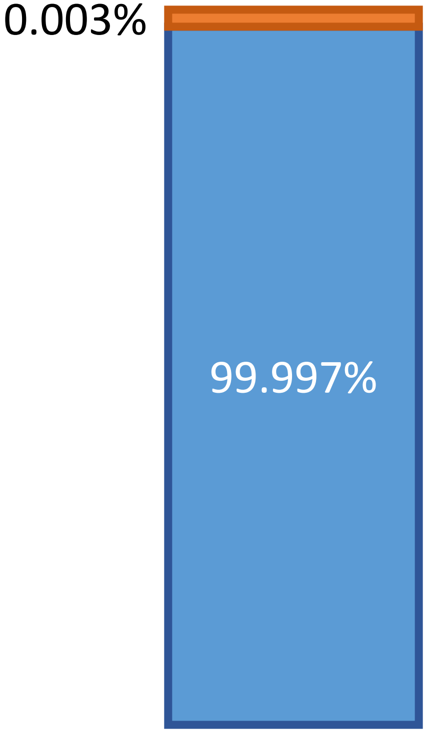
E.g. for 1M example dataset only about 30 represent an event.

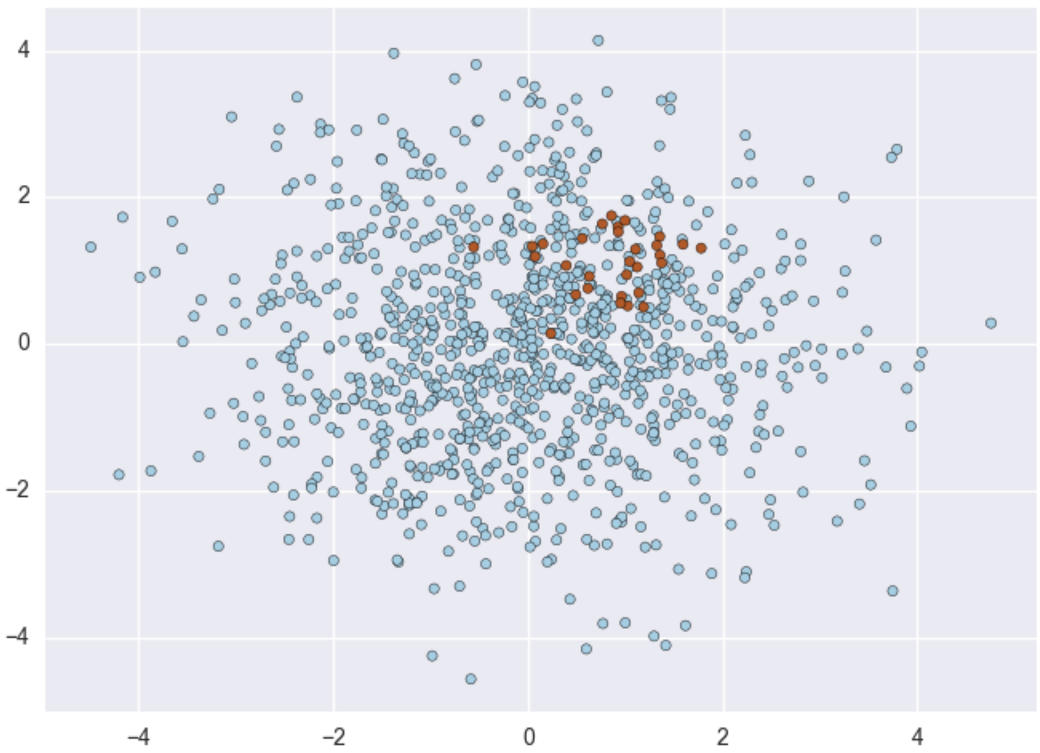
Examples

- About 2% of ***credit card*** accounts are defrauded per year1. (Most fraud detection domains are heavily imbalanced.)
- ***Medical screening*** for a condition is usually performed on a large population of people without the condition, to detect a small minority with it (e.g., HIV prevalence in the USA is ~0.4%).
- ***Disk drive failures*** are approximately ~1% per year.
- ***Factory production defect*** rates typically run about 0.1%.

What happens on classification?

- A classifier that always predict the most common class has an accuracy of 99.997%.





Evaluating Classifiers on Imbalanced Data

- When classes are slightly imbalanced, no balancing is need.
- Yet, take that into consideration when evaluating performances
- Assume the test set contains 100 records

  - Positive cases = 75, Negative cases = 25
  - Is a classifier with 70% accuracy good?
  - No, the trivial classifier (always positive) reaches 75%

  - Positive cases = 50, Negative cases = 50
  - Is a classifier with 70% accuracy good?
  - At least much better than the trivial classifier

Multiclass Problem

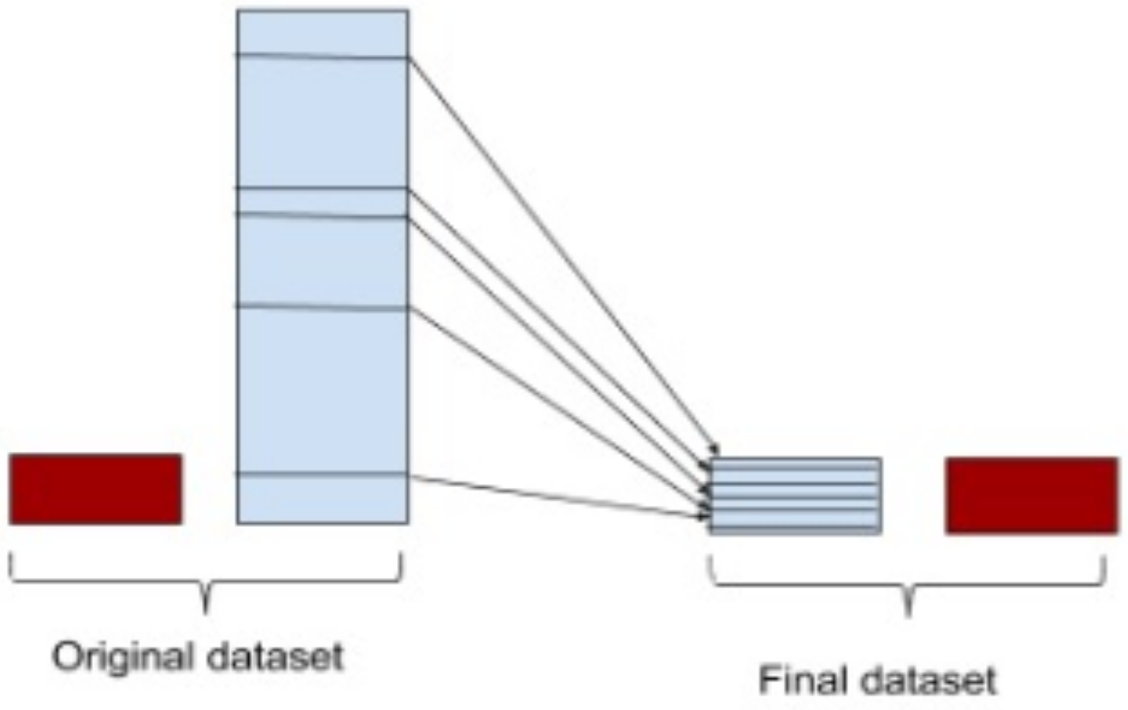
- Assume N classes
- If classes are perfectly balanced, a trivial classifier (e.g. majority) will yield $A_{trivial}$ ~100/N % accuracy
- N=2 → $A_{trivial}$ ~ 50%
- N=4 → $A_{trivial}$ ~ 25%
- Goodness of accuracy of a model should be compared against $A_{trivial}$
- E.g., If N=5, an accuracy of 40% would look large

Handling Imbalanced Data

- Balance the training set
  - Undersampling the majority class
  - Oversampling the minority class
- At the algorithm level
  - Adjust the class weight by making the algorithm more sensitive to rare classes
  - Adjust the decision threshold
  - Design new algorithm to perform well on imbalanced data
- Switch to anomaly detection
- Do nothing and hope to be lucky

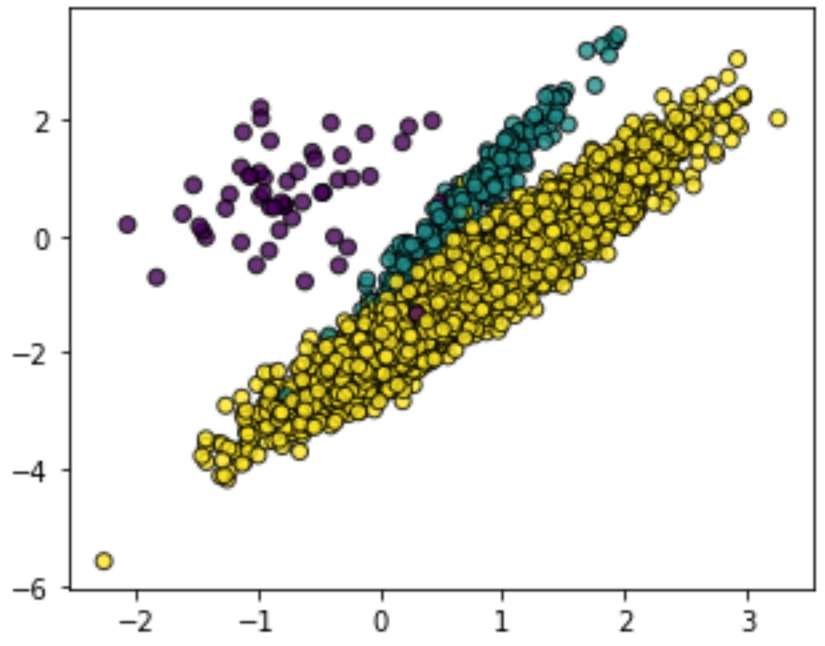
Undersampling the Majority Class

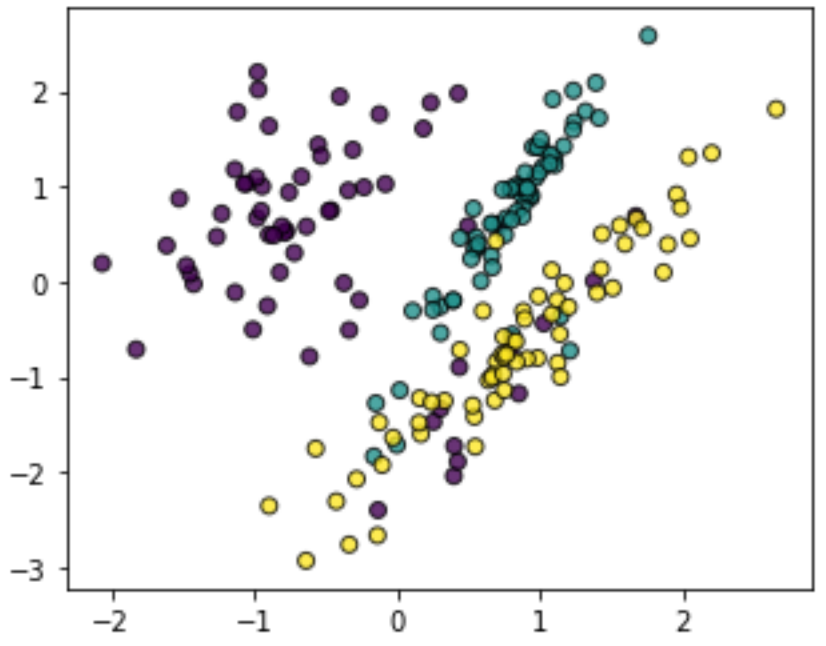
- Random Undersampling
- Neighbor-based approaches, e.g., Condensed Nearest Neighbor, Tomek Links, etc.



Random Undersampling

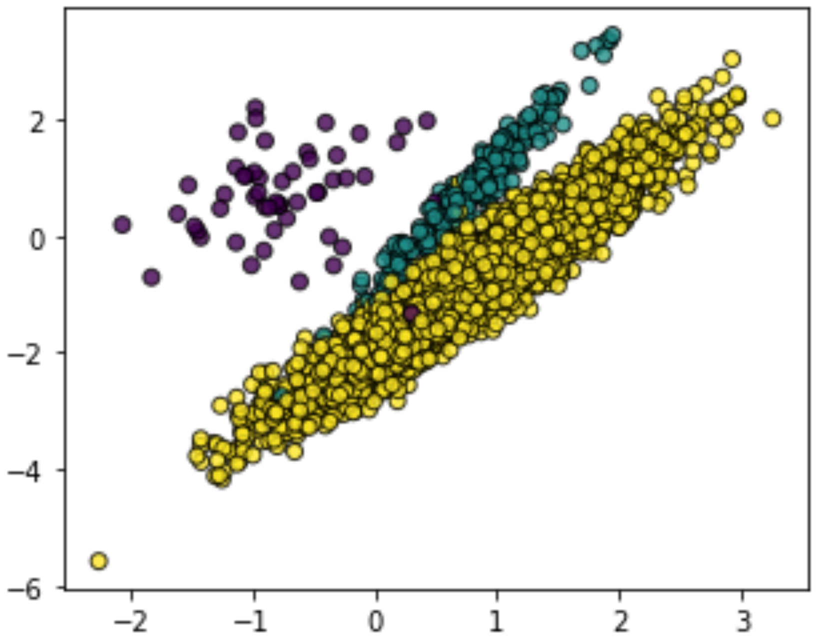
- Under-sample the majority class(es) by randomly picking samples with or without replacement.

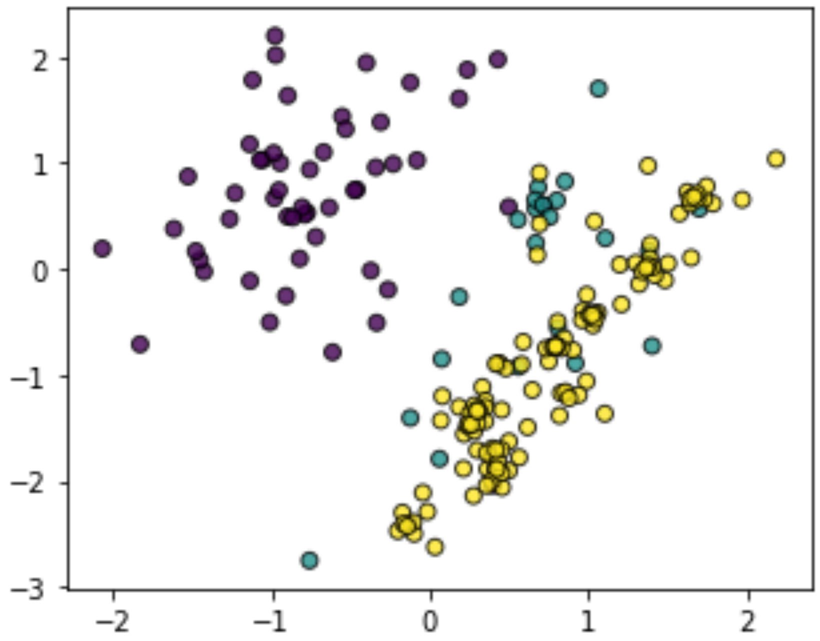




Condensed Nearest Neighbor

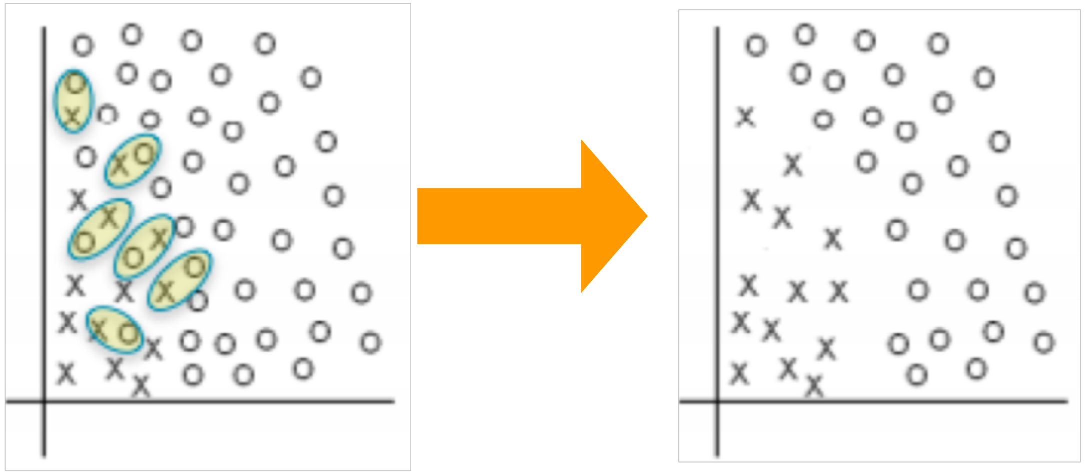
- Performs a smart undersampling by removing majority points having as k-NN a minority point.





Condensed Nearest Neighbor

P. Hart, “The condensed nearest neighbor rule,” In Information Theory, IEEE Transactions on, vol. 14(3), pp. 515-516, 1968



1) The first sample is placed in STORE.

2) The second sample is classified by the NN rule, using as a reference set the current contents of STORE. (Since STORE has only one point, the classification is trivial at this stage.) If the second sample is classified correctly it is placed in GRABBAG; otherwise it is placed in STORE.

3) Proceeding inductively, the ith sample is classified by the current contents of STORE. If classified correctly it is placed in GRABBAG; otherwise it is placed in STORE.

4) After one pass through the original sample set, the procedure continues to loop through GRABBAG until termination, which can occur in one of two ways:

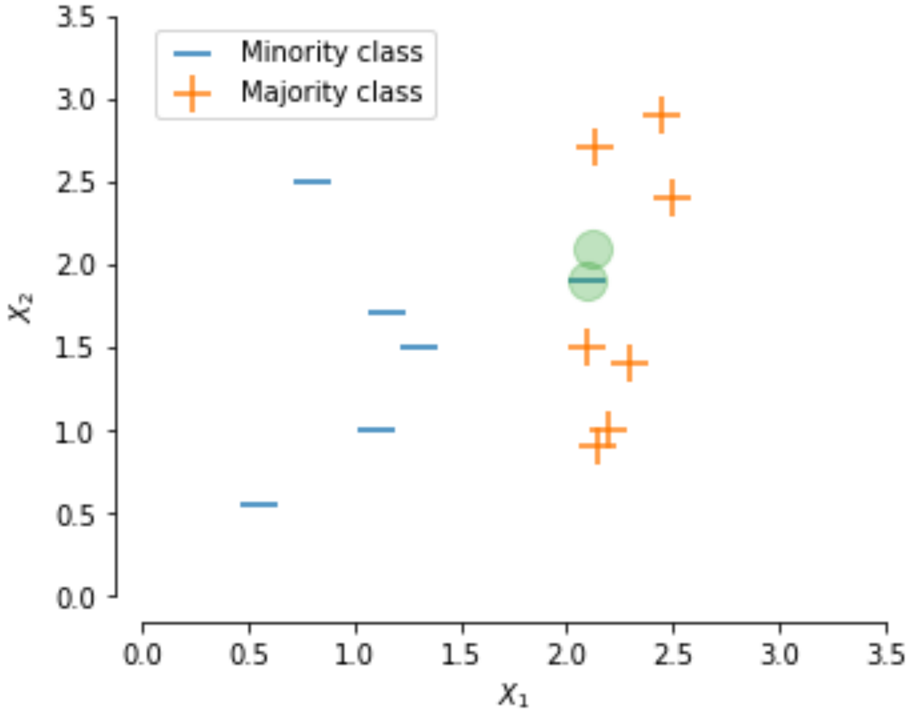
   a) The GRABBAG is exhausted, with all its members now transferred to STORE (in which case, the consistent subset found is the entire original set), or

   b) One complete pass is made through GRABBAG with no transfers to STORE. (If this happens, all subsequent passes through GRABBAG will result in no transfers, since the underlying decision surface has not been changed.)

5) The final contents of STORE are used as reference points for the NN rule; the contents of GRABBAG are discarded.

Condensed Nearest Neighbor

a) pass $\leftarrow$ 1,
b) choose $x \in D$ randomly, $D(1) = D - \{x\}$, $E = \{x\}$,
c) D (pass + 1) = $\varnothing$, count $\leftarrow$ 0,
d) choose $x \in D$ (pass) randomly, classify x by NN using E,
e) if classification found in d) agrees with actual membership of x
  then D(pass + 1) = D(pass + 1) $\cup \{x\}$
  else $E = E \cup \{x\}$, count $\leftarrow$ count + 1,
f) D(pass) = D(pass) $- \{x\}$,
g) if D(pass) $\neq \varnothing$ go to d),
h) if count = 0
  then end of algorithm
  else pass $\leftarrow$ pass + 1, go to b).

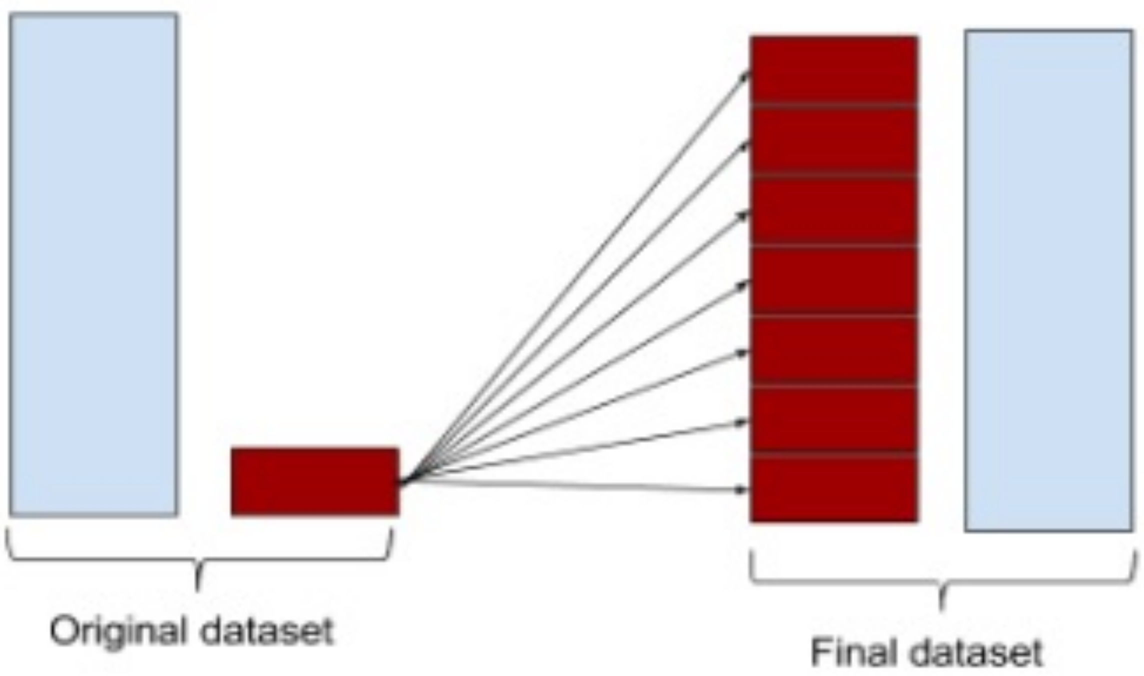


CNN alternatives

- Tomek’s links
- One Sided Selection

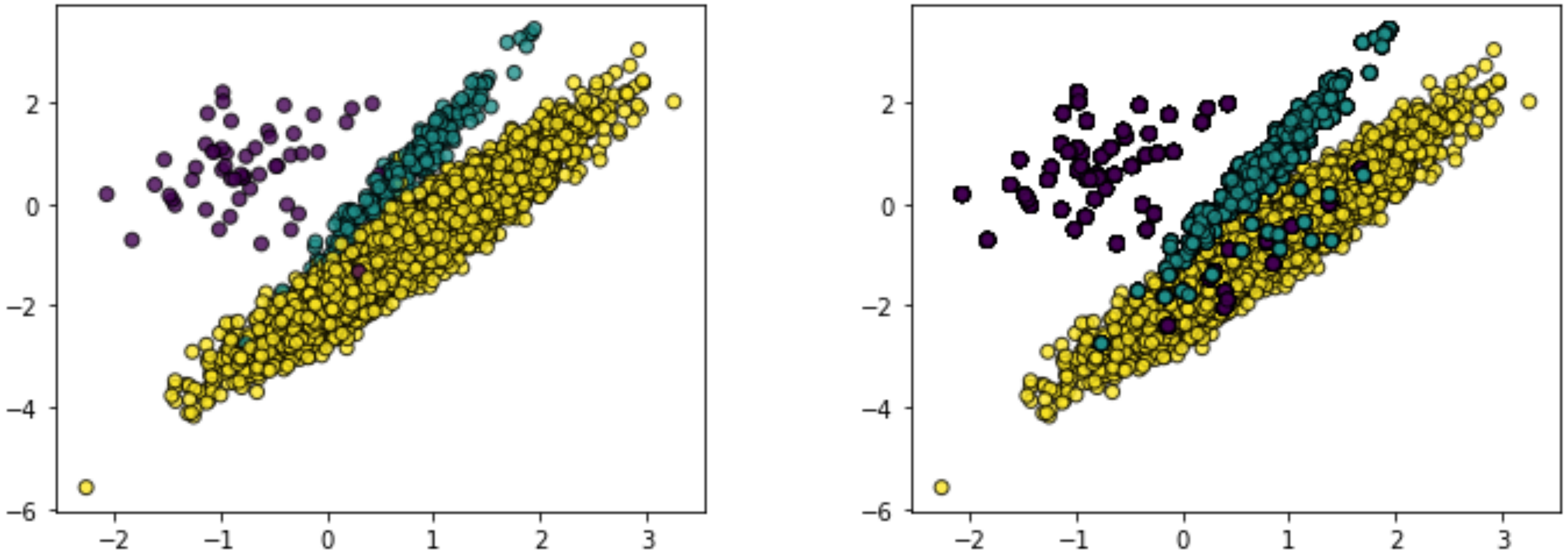
Oversampling the Majority Class

- Random Oversampling
- Synthetic Minority Oversampling Technique (SMOTE)
- Adaptive Synthetic (ADASYN) sampling approach



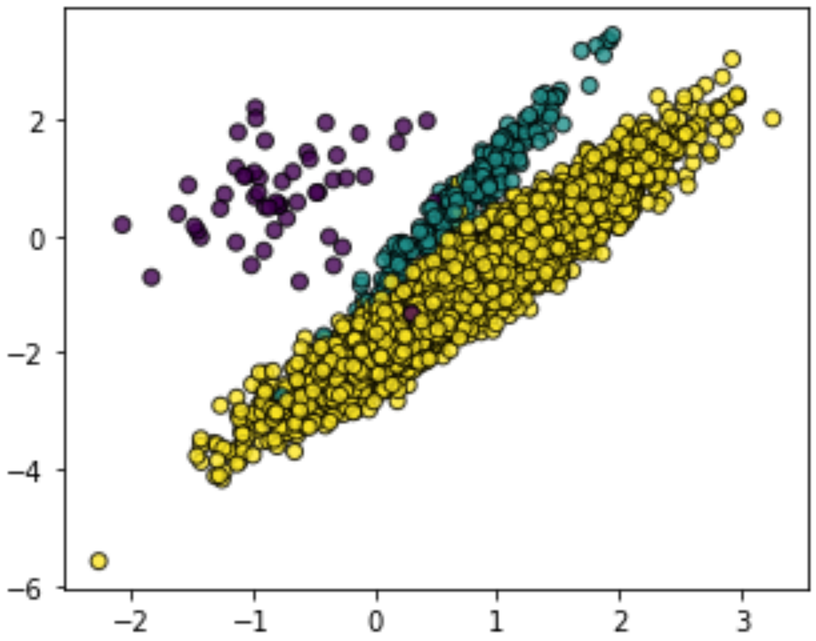
Random Oversampling

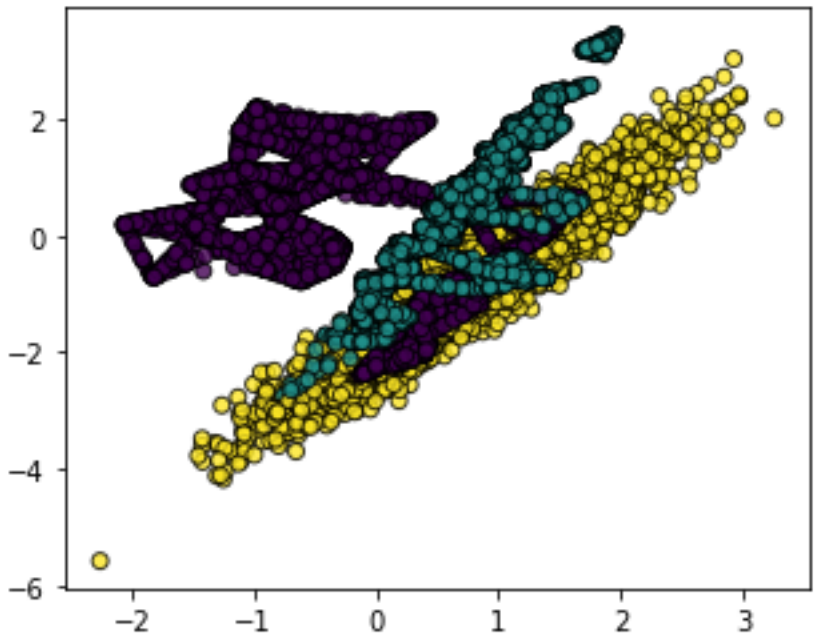
- Over-sample the minority class(es) by picking samples at random with replacement.



SMOTE Oversampling

- Over-sample the minority class(es) by adding points through interporlation.



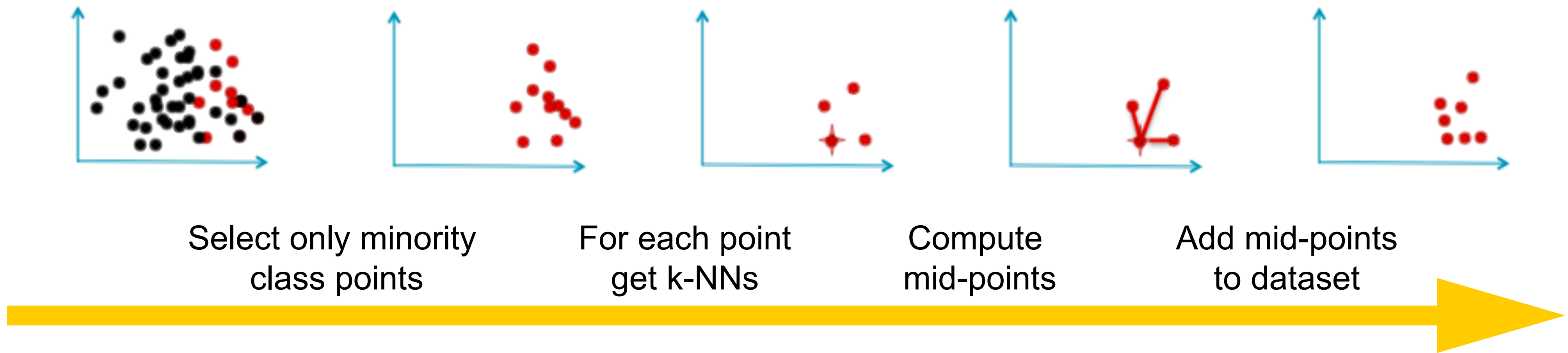


SMOTE

- It operates in the ***"feature space"*** rather than in the "data space", and effectively forces the decision region of the minority class to become more general.
- The minority class is over-sampled by taking each minority class sample and ***introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors***.
- Depending upon the amount of over-sampling required, ***neighbors from the k nearest neighbors are randomly chosen*** (by default k=5).
- E.g., if the amount of over-sampling needed is 200%, only two neighbors from the five are chosen and one sample is generated in the direction of each.

SMOTE – Samples Generation

- Take the difference between the feature vector (sample) under consideration and its nearest neighbor.
- Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.
- This causes the selection of a random point along the line segment between two specific features.

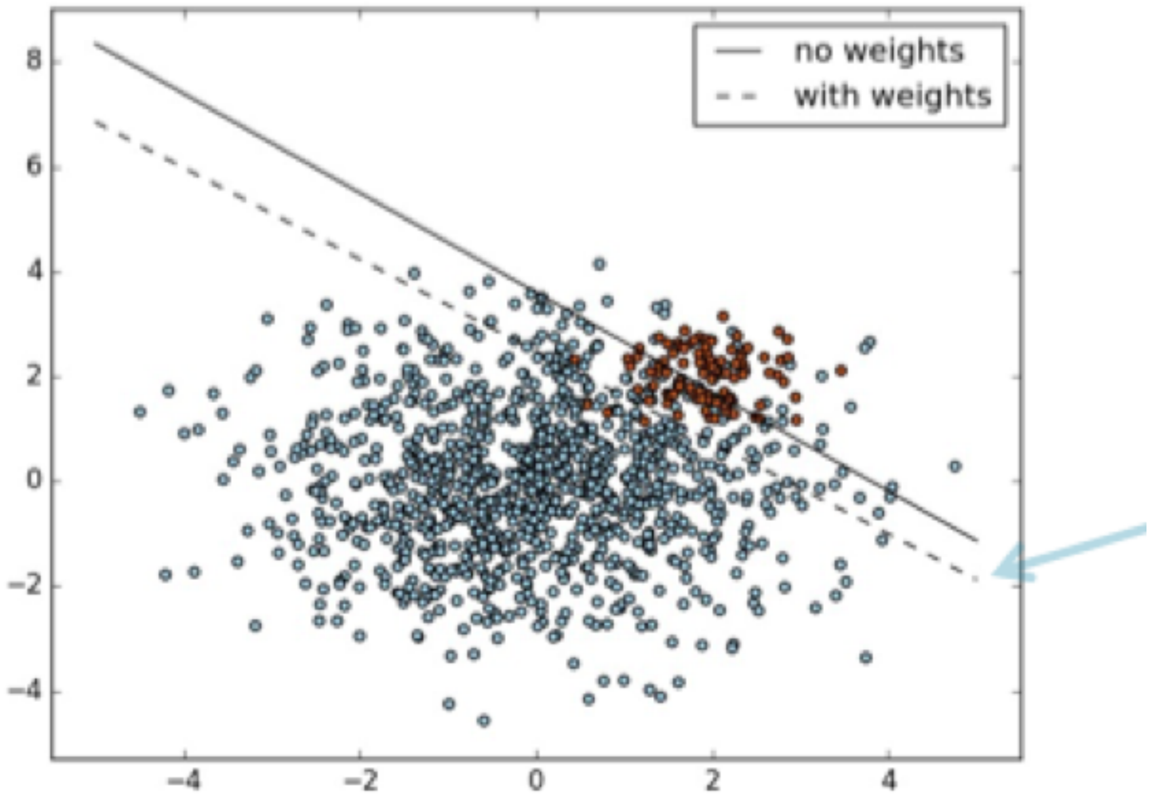


SMOTE alternatives

- SMOTENC: Over-sample for continuous and categorical features.
- BorderlineSMOTE: Over-sample using the borderline variant.
- SVMSMOTE: Over-sample using the SVM variant.
- ADASYN: Over-sample using ADASYN.

Adjust the Class Weight

- The classifier can be trained considering **different costs** to be paid for misclassification errors on minority classes.
- This is generally done using a "class weight''.



Adjust the Class Weight

- Each outcome with respect to a confusion matrix can be associated to a weight in a corresponding weight (or cost) matrix.
- Thus, the objective of the classification algorithm is to find the model that minimizes the total cost.
  - $\sum_X weight(x) freq(x)$

Confusion Matrix

		Actual	
		Y	N
Predicted	Y	50	7
	N	3	40

Weight Matrix

		Actual	
		Y	N
Predicted	Y	0	95
	N	5	0

Cost = 0.03*5 + 0.07*95

Meta-Cost Sensitive Classifier

- Apply a classifier getting probability of a class label P(j|x)
- Compute expected risk of classifying x with class i:

$$R(i|x) = \sum_j P(j|x)C(i,j)$$

- Re-label the train data with the class i having lower risk
- Learn a model on the cost-sensitive train data

Adjust the Decision Threshold

- Several classification methods compute scores in terms of probability of belonging to a class, and then assign class.
- Generally we have:
  - Score $p > 50\%$ → class = Y
  - Otherwise → class = N

- E.g.: decision trees have p = #positive/#negative cases over each leaf

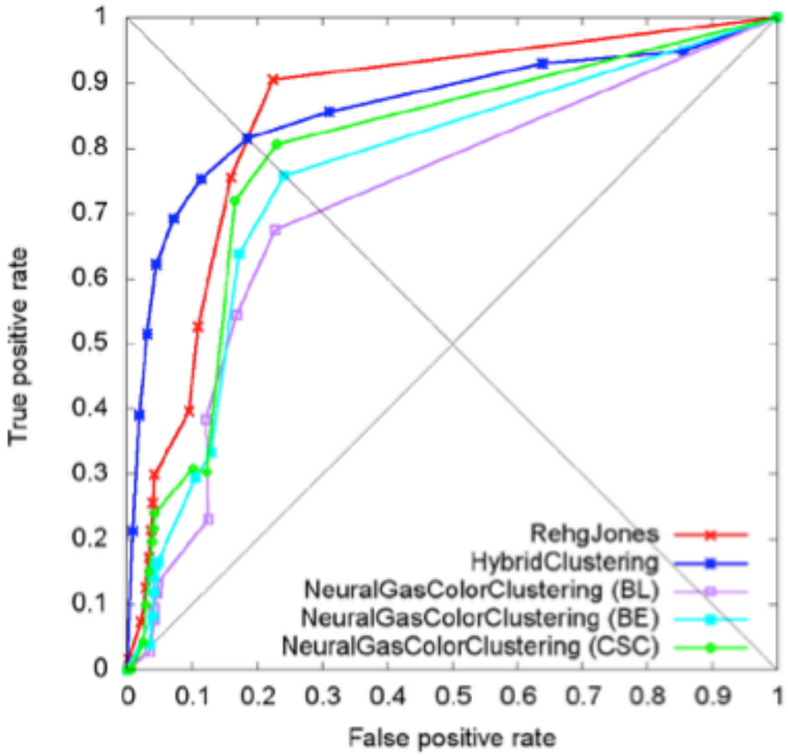
Adjust the Decision Threshold

- What if we generalize the schema into:
  - Score p > THR% → class = Y
  - Otherwise → class = N
- For each THR (in [0-100]) we get a different set of predictions
- The confusion matrix changes and all indicators derived from it change
  - Accuracy
  - True Positive Rate (TPR)
  - False Positive Rate (FPR)
  - ….

Performance Evaluation

Receiver Operating Characteristic Curve

- It illustrates the ability of a binary classifier as its discrimination threshold THR is varied.
- The ***ROC*** curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various THR.
- The TPR = TP / (TP + FN) is also known as ***sensitivity***, ***recall*** or probability of detection.
- The FPR = FP / (TN + FP) is also known as probability of ***false alarm*** and can be calculated as (1 – specificity).



https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Lift Chart

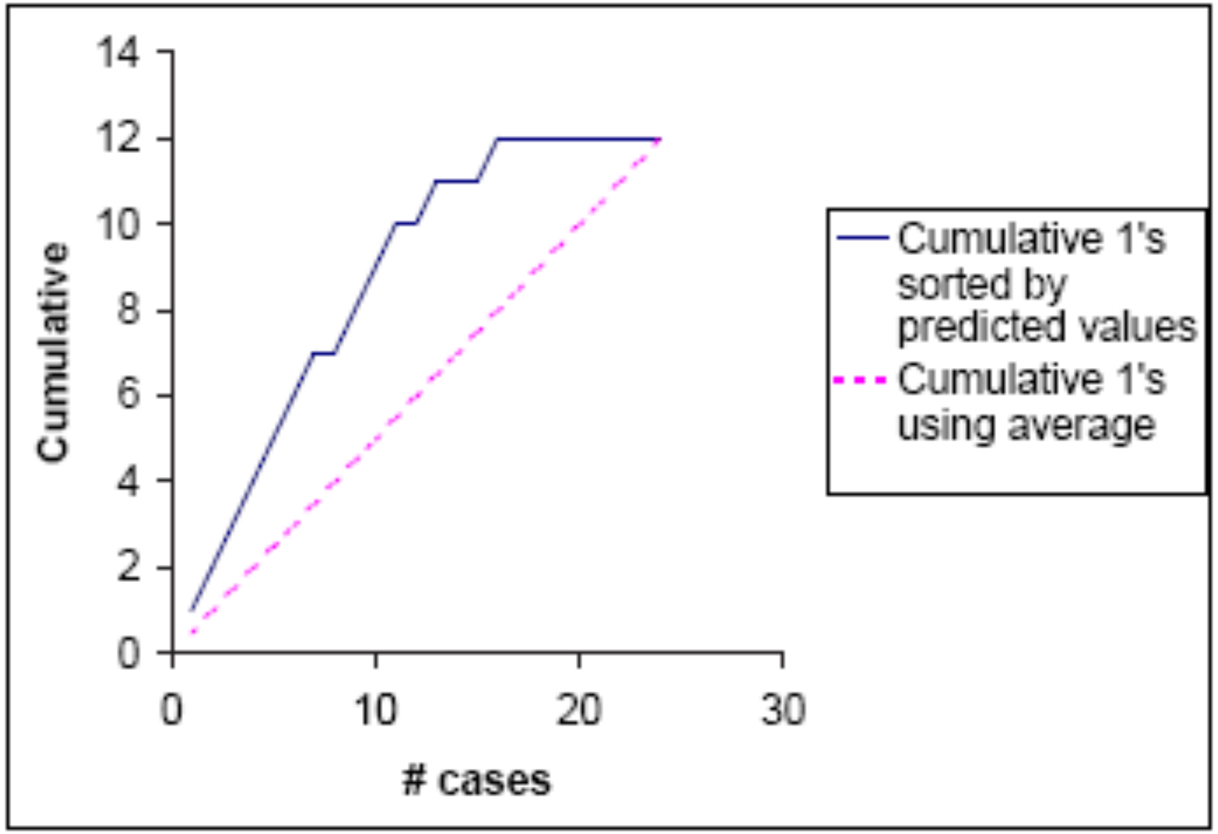
http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html
http://mlwiki.org/index.php/Cumulative_Gain_Chart

- The lift curve is a popular technique in direct marketing.
- The input is a dataset that has been "scored'' by appending to each case the estimated probability that it will belong to a given class.
- The cumulative ***lift chart*** (also called ***gains chart***) is constructed with the cumulative number of cases (descending order of probability) on the x-axis and the cumulative number of true positives on the y-axis.
- The dashed line is a reference line. For any given number of cases (the x-axis value), it represents the expected number of positives we would predict if we did not have a model but simply selected cases at random. It provides a benchmark against which we can see performance of the model.

Notice: "Lift chart" is a rather general term, often used to identify also other kinds of plots. Don't get confused!

Lift Chart – Example

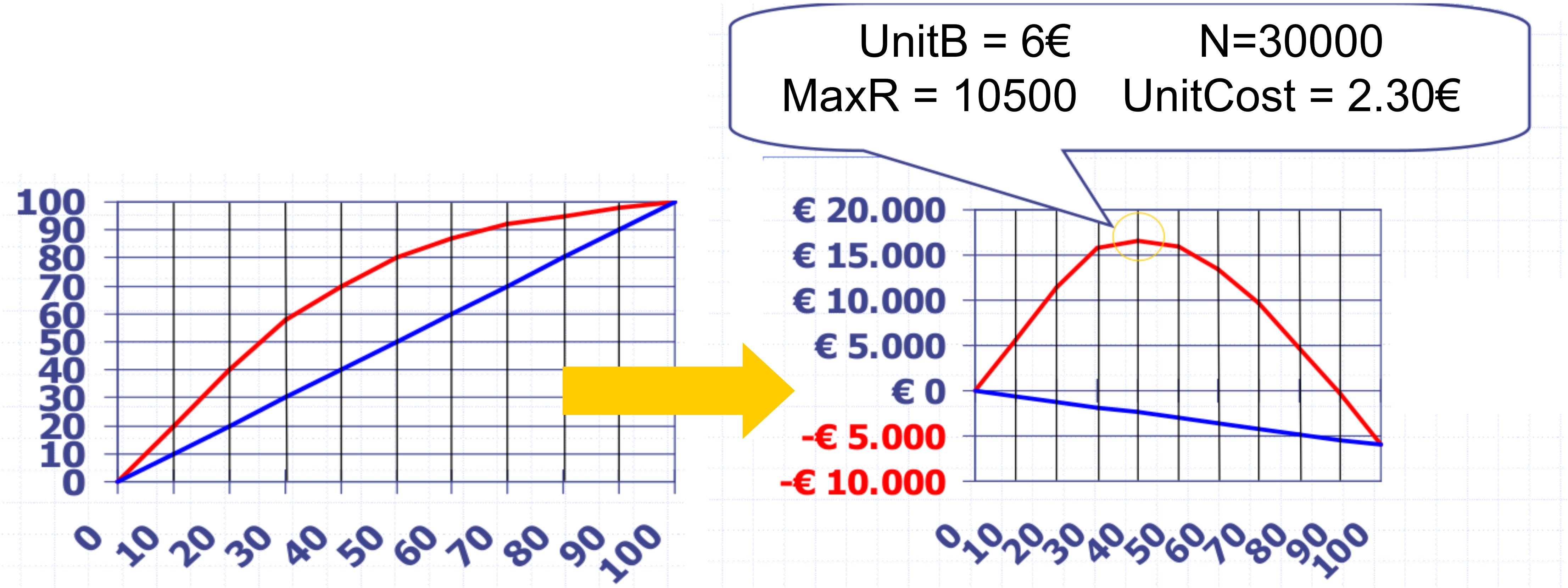
Serial no.	Predicted prob of 1	Actual Class	Cumulative Actual class
1	0.995976726	1	1
2	0.987533139	1	2
3	0.984456382	1	3
4	0.980439587	1	4
5	0.948110638	1	5
6	0.889297203	1	6
7	0.847631864	1	7
8	0.762806287	0	7
9	0.706991915	1	8
10	0.680754087	1	9
11	0.656343749	1	10
12	0.622419543	0	10
13	0.505506928	1	11
14	0.47134045	0	11
15	0.337117362	0	11
16	0.21796781	1	12
17	0.199240432	0	12
18	0.149482655	0	12
19	0.047962588	0	12
20	0.038341401	0	12
21	0.024850999	0	12
22	0.021806029	0	12
23	0.016129906	0	12
24	0.003559986	0	12



Lift Chart – Application Example

- From Lift chart we can easily derive an “economical value” plot, e.g. in target marketing.
- Given our predictive model, how many customers should we target to maximize income?
- Profit = UnitB*MaxR*Lift(X) - UnitCost*N*X/100
- UnitB = unit benefit, UnitCost = unit postal cost
- N = total customers
- MaxR = expected potential respondents in all population (N)
- Lift(X) = lift chart value for X, in [0,..,1]

Lift Chart – Application Example



References

- I. Tomek, “Two modifications of CNN,” In Systems, Man, and Cybernetics, IEEE Transactions on, vol. 6, pp 769-772, 2010.
- N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, 321-357, 2002.
- Domingos, Pedro. "Metacost: A general method for making classifiers cost-sensitive." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999.
- P. Hart, “The condensed nearest neighbor rule,” In Information Theory, IEEE Transactions on, vol. 14(3), pp. 515-516, 1968.
- Python *imblearn* library: https://imbalanced-learn.readthedocs.io/en/stable/index.html