ROBERT GORDON UNIVERSITY ABERDEEN

CMM500: Data Mining
Week 10 – Imbalanced Datasets

Pamela Johnston

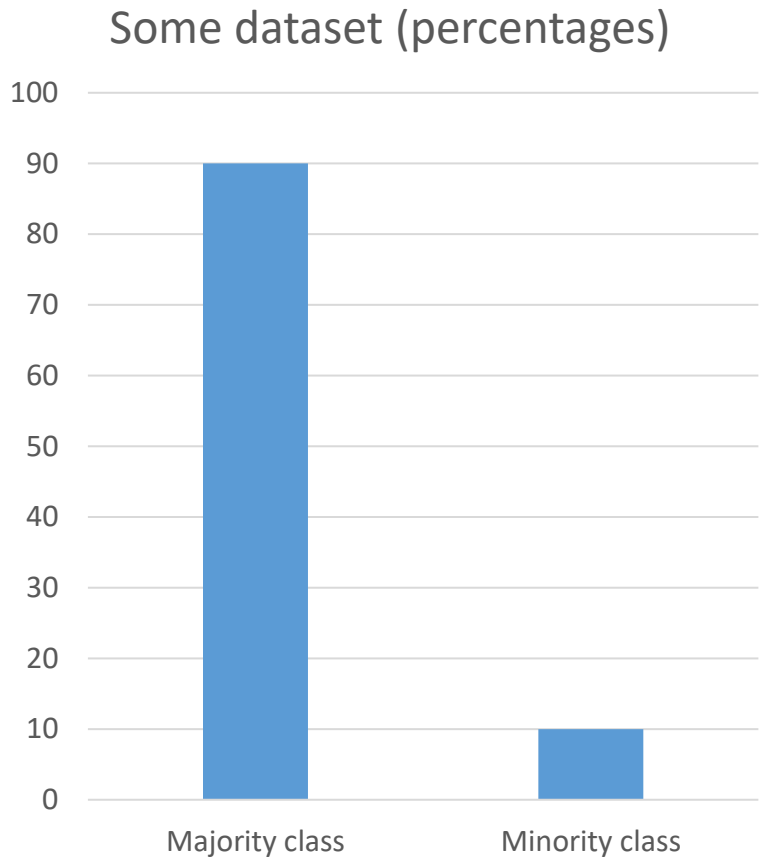
What is class imbalance?

Majority class

Minority class

There are *way more* instances in the majority class than in the minority class.

But the important class is the minority class.



Why is class imbalance a problem?

- Classifiers prioritise correct majority class.
- Poor performance on minority class.
- Insufficient data samples on minority class to compensate.
- “Accuracy” on the whole dataset becomes a useless metric.

Why is class imbalance a problem?

Pam's sweetie classifier

- Jar has 100 sweeties
- 1 is "poisoned"
- 99 are sweet
- Model is 99% accurate

The classifier

Is my dataset imbalanced?

50:50 split -> **not** imbalanced at all (two classes)

33:33:33 -> **not** imbalanced at all (three classes)

60:40 -> not much imbalance (don't worry about it)

20:30:50 -> not much imbalance (don't worry about it)

90:10 -> consider balancing techniques

45:45:10 -> consider balancing techniques

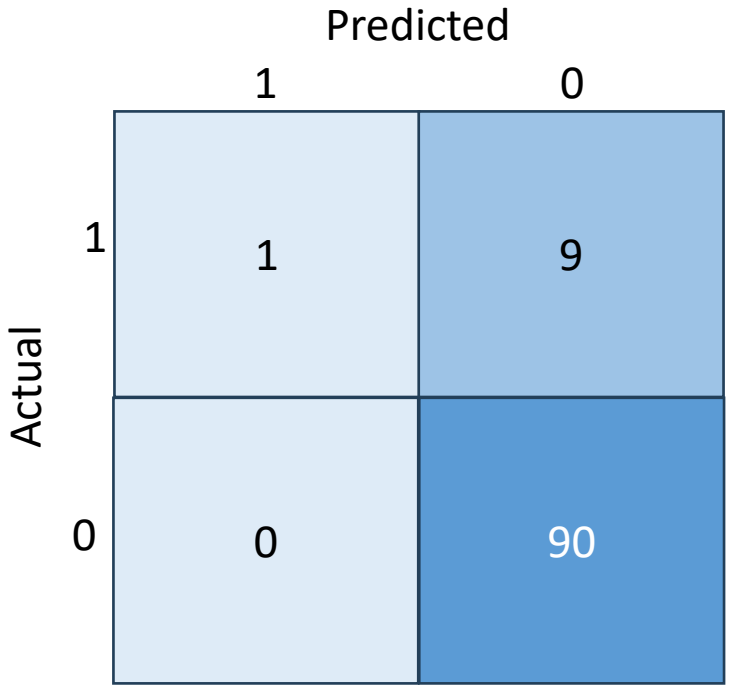
Ambiguity in dataset imbalance

If your dataset is more than ~60:40 imbalanced, then you might get a model improvement by balancing the data.

Is imbalance a problem? (the **real** check)

Once you've trained a classifier:

- Is the minority (important) class (almost) totally ignored?
- What does your confusion matrix look like?
- What percentage of your minority class is predicted correctly (i.e. what is the True Positive Rate)?



An alarming confusion matrix

Class imbalance and classification

- Check confusion matrix
- Which is the “important” class?
- Is the classifier accurate on the minority class or classes?
- Is overall accuracy a good measure?
- Which is the “better” classifier?

Class imbalance and cross validation

Sometimes a dataset is so imbalanced that splitting it into parts for stratified cross fold validation means that some folds do not have any instances of a minority class.

This will cause a warning in R. You will see this in today's lab.

"Class" imbalance and association rules

- What is the effect on "Support"?
- What might be the effect on "Lift"?
- What is the effect on "Confidence"?

- Interesting rules might have much less support than something that is very common.

5% of shopping baskets contain eggs!

100% of shopping baskets contain milk!

100% of shopping baskets with eggs also contain bacon!

Attribute imbalance and clustering

Minority class might form one very small cluster.

Clusters of the majority class might be bigger than the minority class cluster.

Analysis of the clusters might reveal this (but only if the minority class cluster is of sufficient size to appear in the analysis).

What can be done about class imbalance

Tackle it in the dataset:

- Undersample
- Oversample
- Generate synthetic samples

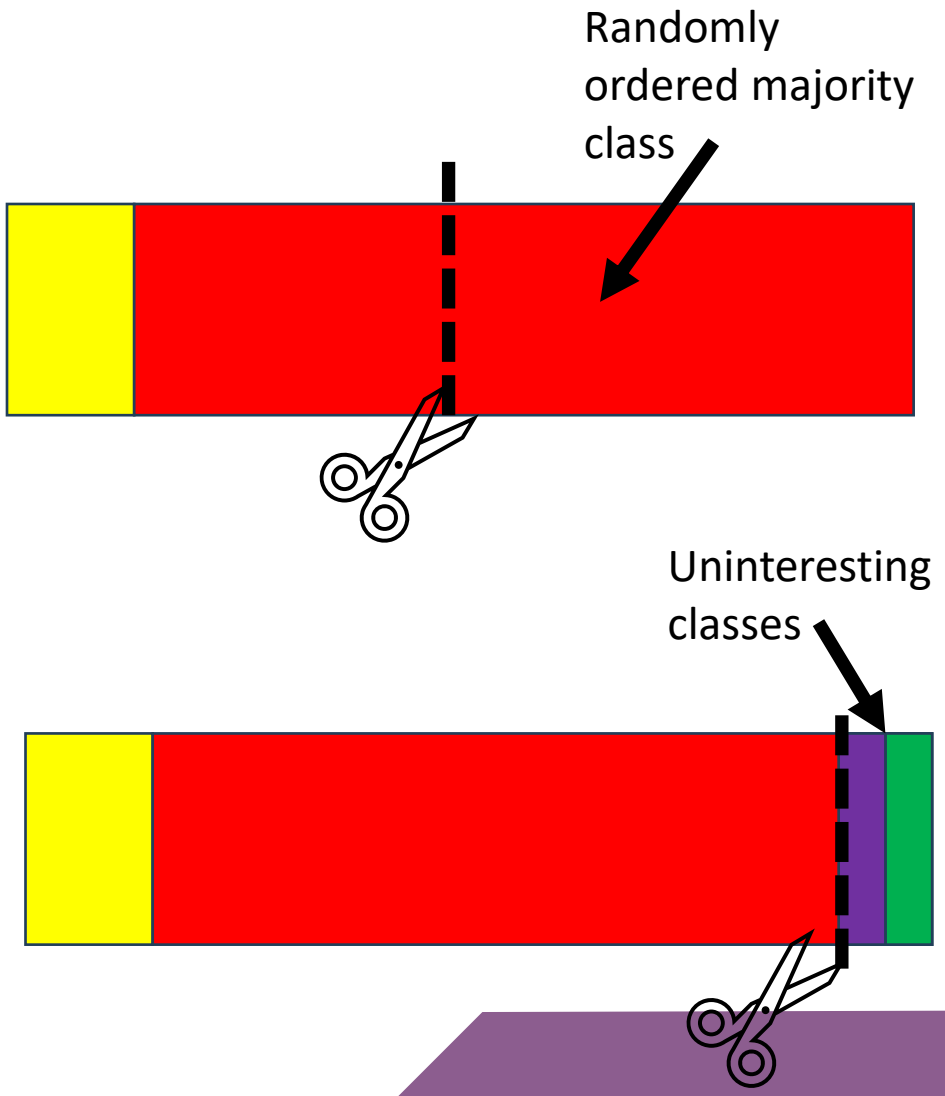
Tackle it in the model:

- Some models allow for class weights so that a particular class can be seen as more important.

Undersampling (easy)

Simply remove some of the majority class to reduce the class imbalance of the minority class.

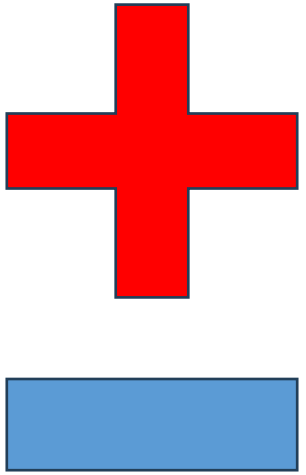
If you have multi-class classification but are not interested in the minority classes, you can remove whole classes.



Undersampling choices

In the majority class, you can:

- Decide to **keep** specific samples
- Decide to **throw away** specific samples
- Do a **combination** of both



Undersampling (more complex)

NearMiss Undersampling and **CondensedNearestNeighbours** : sample (**add**) the instances of the majority classes that are closest to the instances of the minority class.

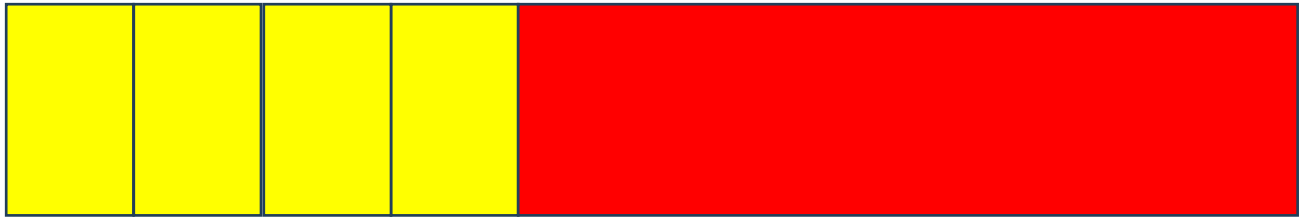
TomekLinks: decides which majority class instances to **remove** based on their distance from instances of the opposite class.

See the lab for examples.

Oversampling (easy)

Just double up on a selection of minority class samples.

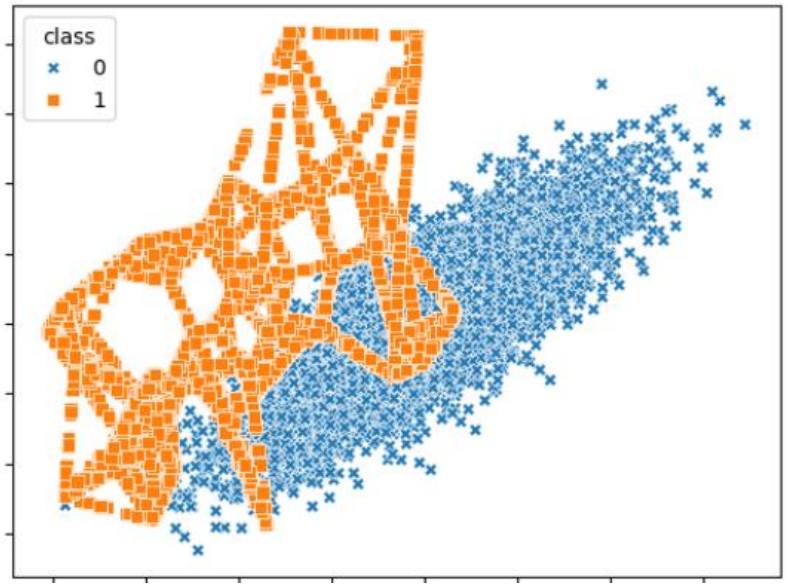
This effectively re-weights the sample so that the model sees minority class instances as more important than the appear in the distribution.



Oversampling (SMOTE)

SMOTE combines both undersampling the majority class and oversampling the minority class.

“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.”



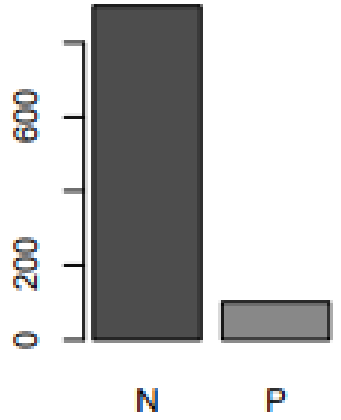
Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002)

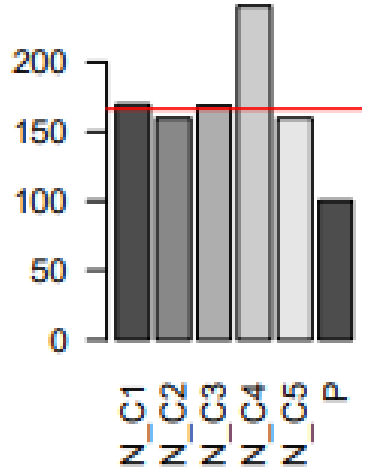
Increasing the number of classes available in the majority class.

Find within-class similarity in the dominant class

Oversample minority-class instances

Classify and compare





Elyan, Eyad, Carlos Francisco Moreno-Garcia, and Chrisina Jayne. "CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification." *Neural computing and applications* 33 (2021)

CDSMOTE motivation

Binary classification is dominant in many medical datasets, with the positive class denoting the existence of a particular disease in medical diagnosis applications

Such labelling does not depict the reality of having **different categories** of the same disease

What if datasests were **decomposed using clustering** of each class to reveal hidden categories?

Such class decomposition has three (potential) advantages:

1. Diversification of the input that enhances the ensemble classification;
2. Improving class separability, easing the follow-up classification process;
3. Finer-grained training.

Class decomposition – example Iris dataset

Sepal len	Sepal wid	Petal len	Petal wid	class
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
5.0	3.6	1.4	0.4	Setosa
5.4	3.9	1.7	0.3	Setosa
4.4	3.4	1.4	0.2	Setosa

Three classes:

- Setosa
- Virginica
- Versicolor

Class decomposition: Iris

Cluster instances within each class into 2 groups

Table: Original Dataset

No	Class-Label	Frequency
1	setosa	50
2	versicolor	50
3	virginica	50

Table: Clustered Dataset (*k=c(2,2,2)*)

No	Class-Label	Frequency
1	setosa_**c1**	28
2	setosa_**c2**	22
3	versicolor_**c1**	24
4	versicolor_**c2**	26
5	virginica_**c1**	22
6	virginica_**c2**	28

Oversampling (synthetic samples)

Synthetic data samples are a growing research trend.

Useful for classification

Data samples can be used to supplement the minority class.

Should also be present in the majority class or classifier could learn that *synthetic == minority class*

Generative AI is aiding the generation of synthetic samples.

ROBERT GORDON UNIVERSITY ABERDEEN

Synthetic image examples from GANs

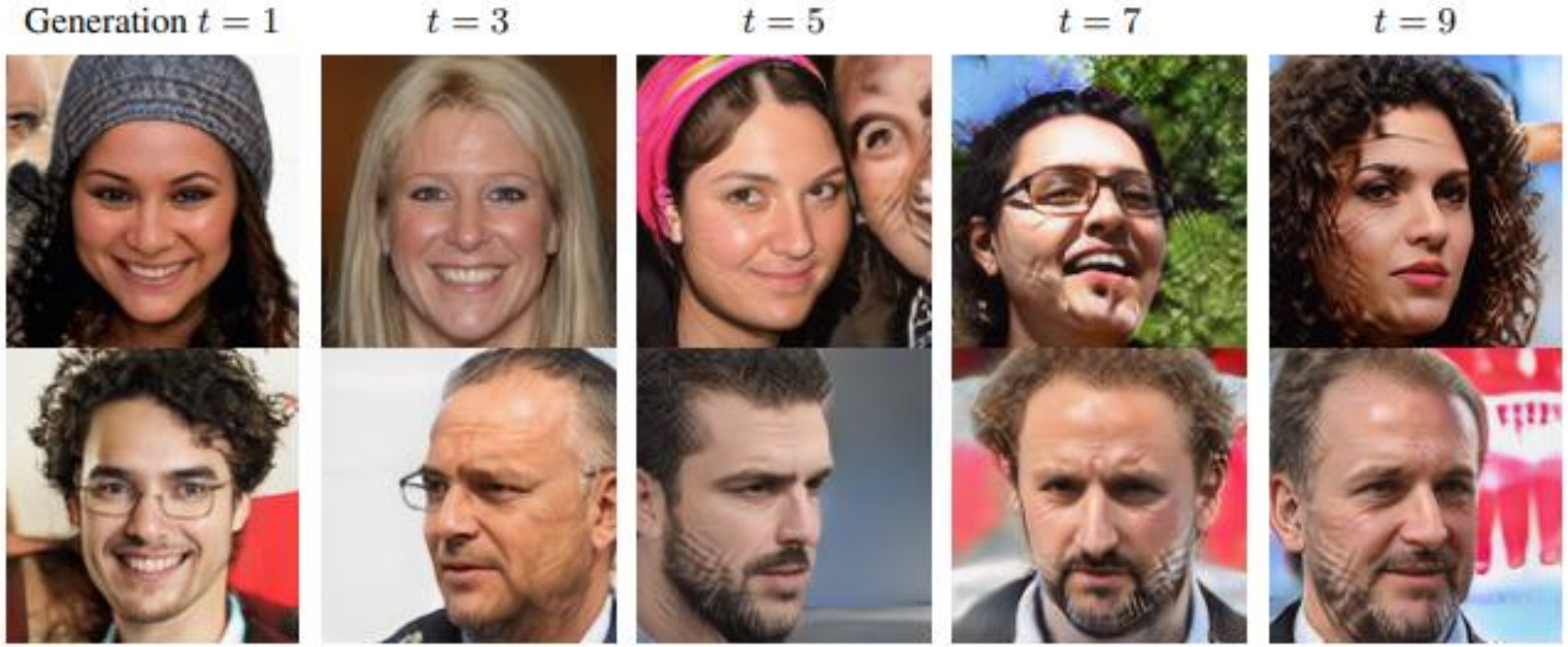
Images do not have to look perfect to supplement the dataset.

Useful for imbalanced image datasets.

Ali-Gombe, Adamu, and Eyad Elyan. "MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial network." *Neurocomputing* 361 (2019)

ROBERT GORDON UNIVERSITY ABERDEEN

Synthetic data: use caution if training generative AI.



Alemohammad, Sina, et al. "Self-consuming generative models go mad." *arXiv preprint arXiv:2307.01850* (2023).