

Data Mining

S4

NOVA-IMS 2019/2020

Fernando Lucas Bação

bacao@isegi.unl.pt

<http://www.isegi.unl.pt/bacao>

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1



**Data Preparation and
Pre-processing**

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AACSB Accredited

UNIGIS

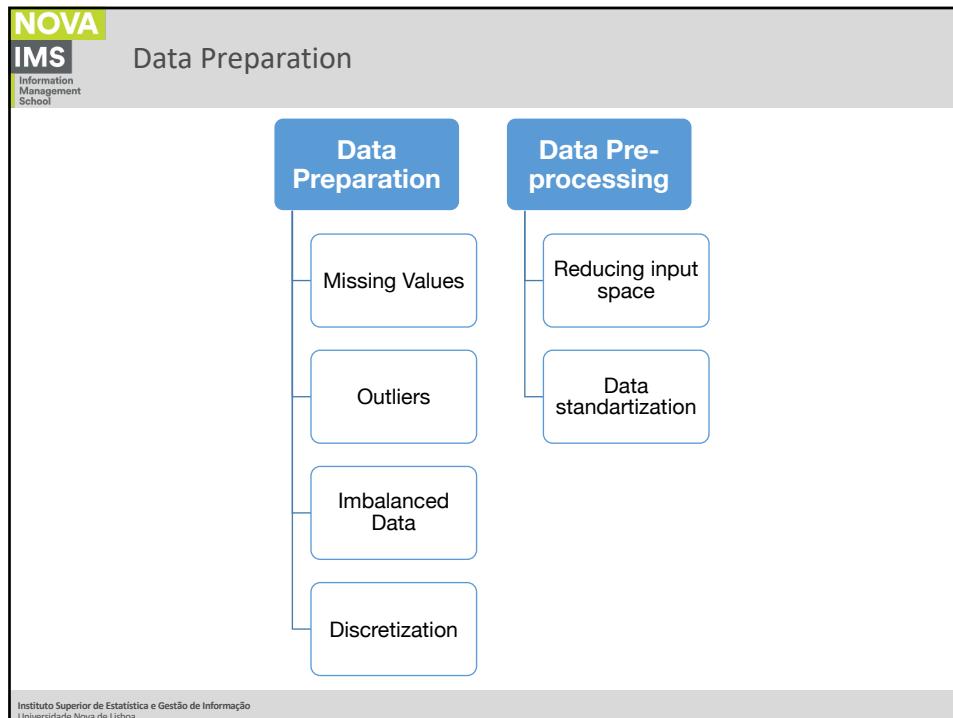
A3ES

EFMD

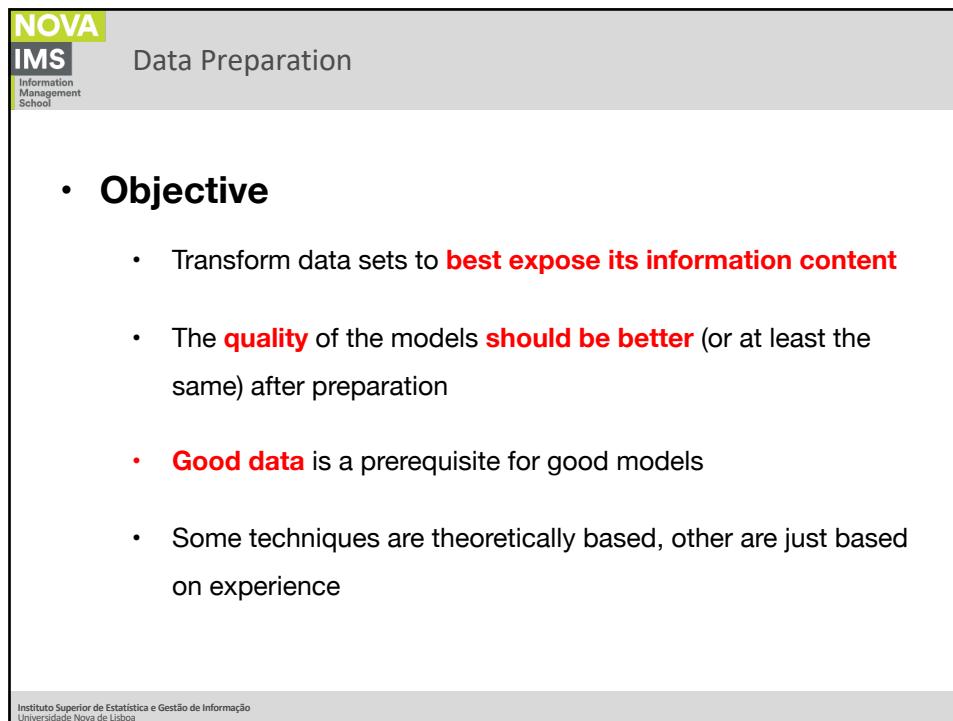
Network of Schools

eduniversal

2



3

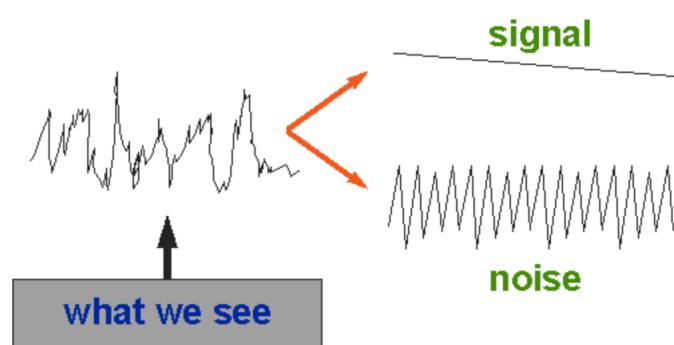


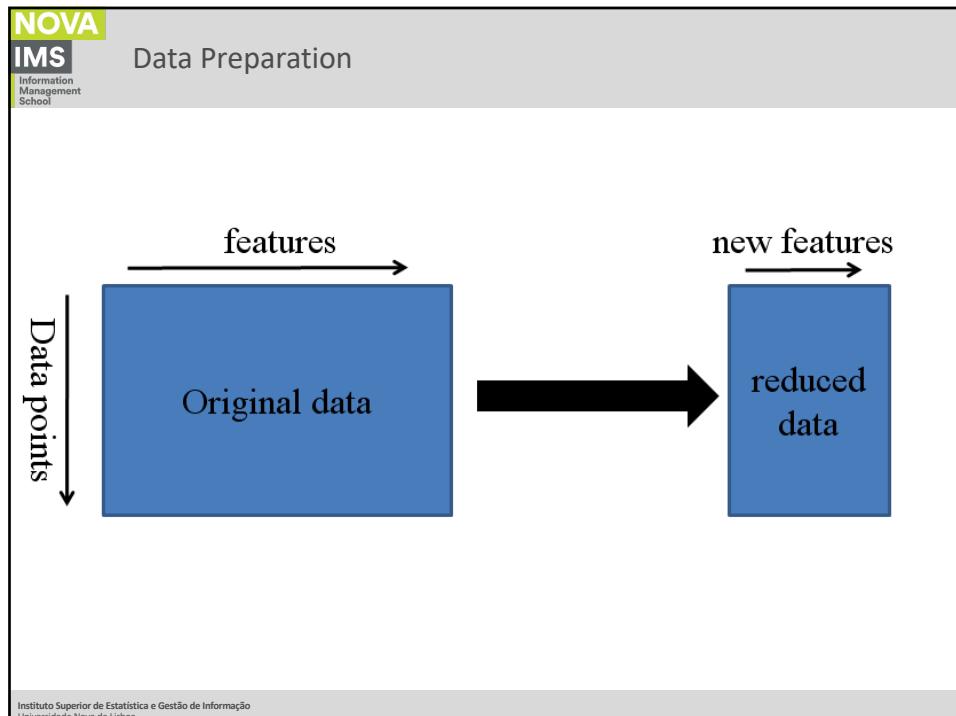
4

- **Signal vs. Noise**

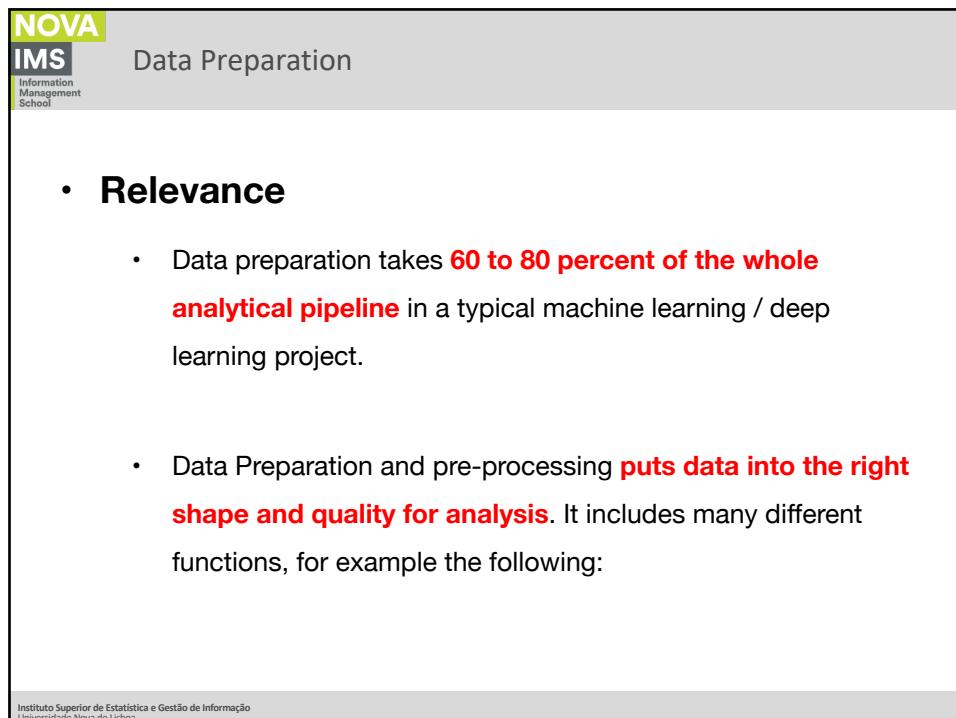
- In science (physics and telecommunications) noise is defined as **fluctuations and external disturbances** in the flow of information (signal) received;
- An **undesired disturbance** in relevant information;
- A disturbance that affects a signal and that may distort the information carried by the signal.

What we observe can be divided into:



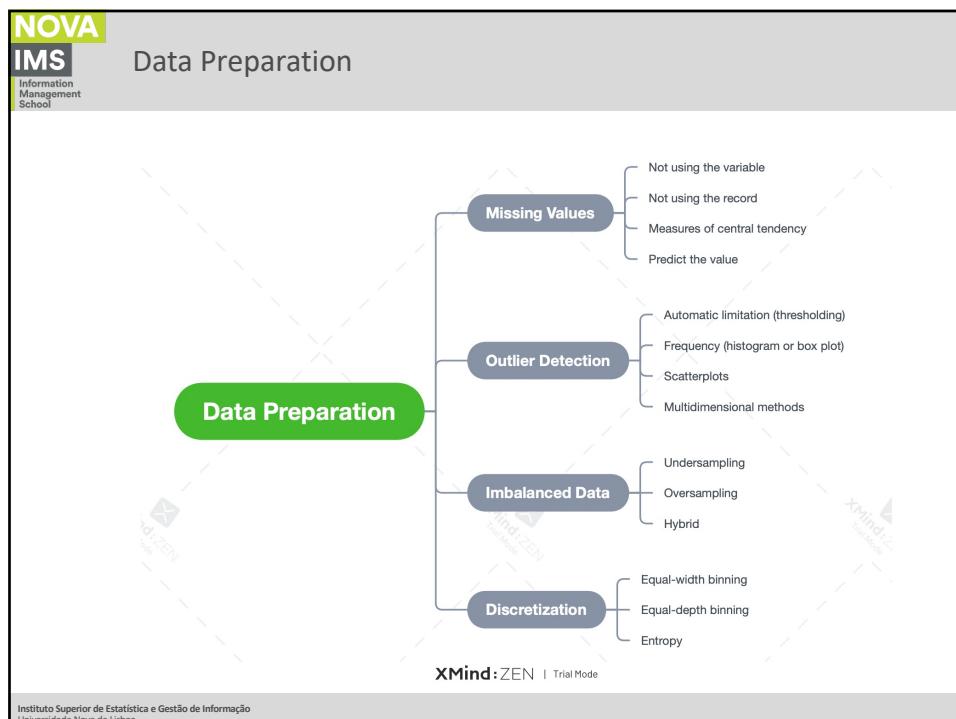


7

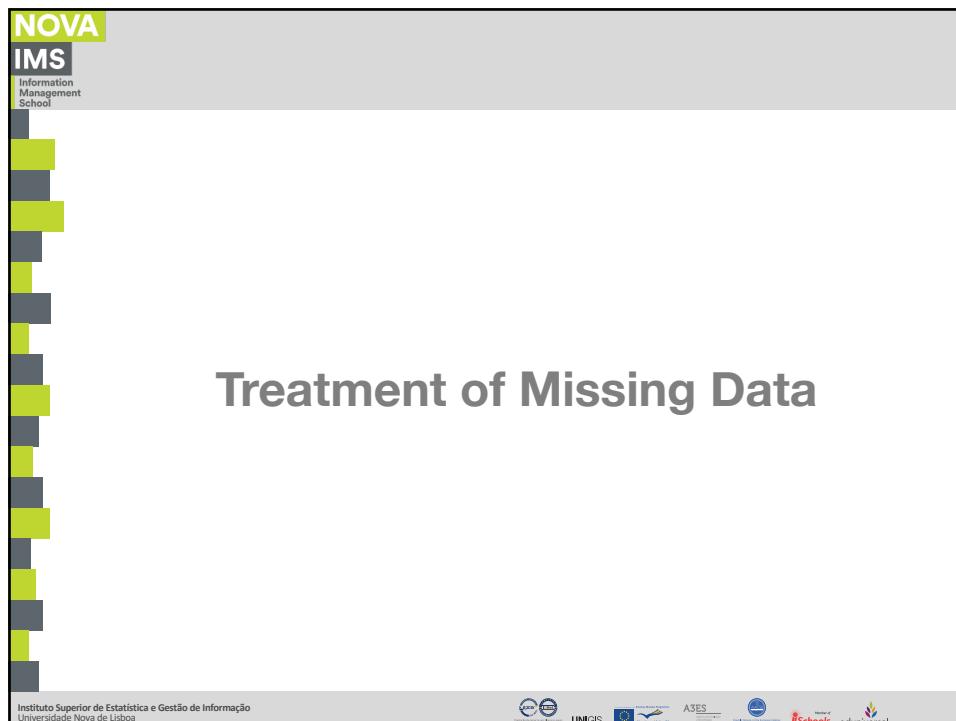


8

- **Real data suffers from several problems**
- Incomplete
 - Missing values, lacking attributes of interest, levels of aggregation
- Noisy
 - Errors and outliers
- Inconsistent
 - E.g. Age=42 Birthday=31/07/1997
 - Changes in scales
 - Duplicate records with different values



11



12

NOVA
IMS
Information Management School

Data Preparation

- Missing Data

Inputs

Inputs				
Records	1	2	3	4
1	?	?	?	?
2	?	?	?	?
3	?	?	?	?
4	?	?	?	?
5	?	?	?	?

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

13

NOVA
IMS
Information Management School

Data Preparation

- Missing Data

Inputs

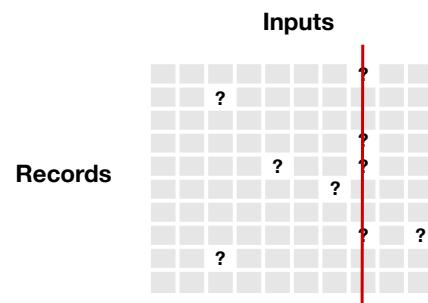
Inputs				
Records	1	2	3	4
1	?	?	?	?
2	?	?	?	?
3	?	?	?	?
4	?	?	?	?
5	?	?	?	?

10 to 3

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

14

- Missing Data



OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

NOVA**IMS**Information
Management
School**Data Preparation**

OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

OS	Global Mean
Android	1,804
iOS	3,027
iOS	8,788
Android	4,145
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	4,145
Android	1,146

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

17

NOVA**IMS**Information
Management
School**Data Preparation**

OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

OS	Global Mean	Group Mean
Android	1,804	1,804
iOS	3,027	3,027
iOS	8,788	8,788
Android	4,145	2,696
Android	3,735	3,735
Android	1,056	1,056
iOS	9,319	9,319
Android	6,199	6,199
Android	2,235	2,235
iOS	4,145	7,045
Android	1,146	1,146

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

18

NOVA
IMS
Information Management School

Data Preparation

Missing value record

Other dataset records

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

19

NOVA
IMS
Information Management School

Data Preparation

CUSTID	DAYSWU	AGE	EDU	INCOME
1138	951	29	17	55637
6313	1231	22	16	32795
9185	1186	31	18	40717
7735	1097	27	18	29184
7918	1196	19	13	16479
1585	849	34	19	55403
7897	1120	46	15	63603
7389	792	39	16	57402
6764	1060	missing	15	69803
8416	1217	31	18	56028
6541	1074	54	15	91934
8263	1185	46	18	78262
8052	1007	52	16	91292
2448	840	51	14	62699
10978	553	44	19	62675
1409	1191	missing	14	64449
6063	1222	32	18	55048
9767	961	34	20	53217
4489	1224	69	16	103191
10738	1190	missing	18	64412
10381	939	54	16	73709
8999	1054	45	15	73074
4482	1089	50	19	63983
4257	1074	missing	20	75006
7722	876	50	19	55332
7491	1165	49	14	63150
1037	1127	47	16	69713
10419	1112	48	15	74961
5124	1217	47	17	72890
8140	1229	54	16	74670
3218	1159	54	15	83956
4124	1247	53	15	81937
8990	1025	54	17	79364
7155	1235	54	18	81948
6346	1170	42	17	71182
9547	1125	missing	16	81442
8250	989	12	16	75566
2527	1234	51	15	77635
6663	1185	40	14	61974

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

20

- **Missing data**

- Delete variables (you loose information)
- Delete records (potential bias);
- To examine and manually enter a probable value (tedious + infeasible);
- Automatically fill in with a measure of central tendency (i.e. mean, median, mode);
- Automatically fill in with a measure of central tendency of a subset (e.g. men and women);
- To fill in with values from similar individuals (nearest neighbours);
- Predictive model (linear regression, multiple linear regression);
- Code the missing data explicitly.

- **Missing data**

- The most practical approach to the problem is to initially use the **quickest and simplest option**;
- After achieving some **preliminary results** we can comparatively analyze the performance of the model in the full sample patterns and in those where there was a need to estimate missing values;
- In the event that the **error is significantly higher** than in other data, then we will try to use another method in order to improve results.



Outlier treatment

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa



23



Data Preparation

- **Outliers**
 - In statistics, an outlier is an observation point that is distant from other observations.
 - An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.
 - Extreme cases in one or more variables and with great impact on the interpretation of results;
 - Outliers may come from:
 - Unusual but correct situations (the Bill Gates effect),
 - Incorrect measurements,
 - Errors in data collection;
 - Lack of code for missing data.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

24

**NOVA
IMS**
Information Management School

Data Preparation

- Outliers (leverage effect)**

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

25

**NOVA
IMS**
Information Management School

Data Preparation

- Remove Data Outliers**
 - Automatic limitation (thresholding)
 - The imposition of maximum and minimum values for the variables (age – 0 e 100)

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

26

**NOVA
IMS**
Information Management School

Data Preparation

- Remove Data Outliers

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

27

**NOVA
IMS**
Information Management School

Data Preparation

Q1: Quartile 1, or median of the *left* data subset
after dividing the original data set into 2 subsets via the median
(25% of the data points fall below this threshold)

Q3: Quartile 3, median of the *right* data subset
(75% of the data points fall below this threshold)

IQR: Interquartile-range, $Q_3 - Q_1$

Outliers: Data points are considered to be outliers if
value < $Q_1 - 1.5 \times IQR$ or
value > $Q_3 + 1.5 \times IQR$

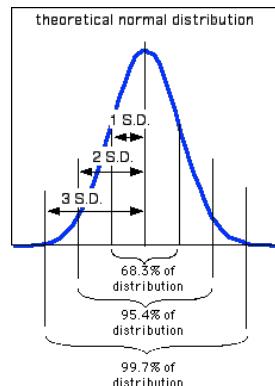
fmi:
http://sebastianraschka.com/Articles/2014_dixon_test.html
http://www.itl.nist.gov/div898/handbook/prc/section1/prc1_6.htm

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

28

- Remove Data Outliers

- Normal data distribution
 - $3\sigma +/\text{-} \text{ Average}$

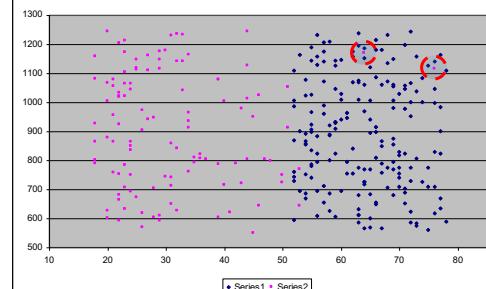
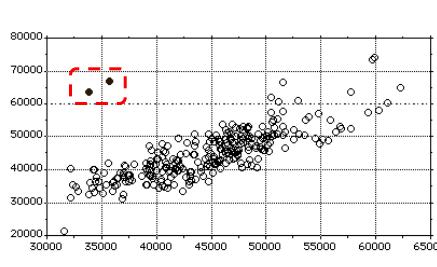


Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

29

- Remove Data Outliers

- In two dimensions...



Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

30

- Remove Data Outliers
 - Cluster Analysis (K-means)
 - Self-Organizing Maps

Imbalanced Learning

Quick quizz:

	Yes	No
Is it possible to have a useless classifier with 99% accuracy?		
Is it possible to achieve a 99.9% accuracy with a trivial classifier?		

Quick quizz:

	Yes	No
Is it possible to have a useless classifier with 99% accuracy?	x	
Is it possible to achieve a 99.9% accuracy with a trivial classifier?	x	

Quick quizz:

	Yes	No
Is it possible to have a useless classifier with 99% accuracy? If the minority class is 1%	X	
Is it possible to achieve a 99.9% accuracy with a trivial classifier? If the minority class is 0,1%	X	

- Class Imbalance
 - Is this frequent in real-world applications?
 - **Credit Card** frauds - ~2% per year.
 - **HIV prevalence** in the USA - ~0.4%.
 - **Disk drive** failures - ~1% per year.
 - **Factory production** defects - ~ 0.1%.
 - **Business churn** - ~3%

- Imbalanced Learning

- An **imbalanced learning** problem is defined as a classification task for binary or multi-class datasets where a **significant asymmetry** exists between the **number of instances for the various classes**.
- The dominant class is called the **majority class (negative cases)** while the rest of the classes are called the **minority classes (positive cases)**
- The **Imbalance Ratio** (IR), is the ratio between the majority class and the minority class, (depends on the type of application and for binary problems values between 100 and 100.000 have been observed)

- Imbalanced Learning

- **Standard learning methods induce a bias** in favor of the majority class during training.
- This happens because the **minority classes contribute less to the maximization** of the objective function, which is usually accuracy.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TotalPopulation}}$$

Imbalanced Learning

- Imbalanced Learning
 - Standard learning methods induce a bias in favor of the majority class during training.
 - This happens because the minority classes contribute less to the maximization of the objective function, which is usually accuracy.

		Actual Value	
		Positives	Negatives
Predicted Value	Positives	TP True Positives	FP False Positives
	Negatives	FN False Negatives	TN True Negatives

Imbalanced Learning

- Imbalanced Learning
 - Standard learning methods induce a bias in favor of the majority class during training.
 - This happens because the minority classes contribute less to the maximization of the objective function, which is usually accuracy.

		Actual Value	
		Positives	Negatives
Predicted Value	Positives	TP True Positives	FP False Positives
	Negatives	FN False Negatives	TN True Negatives

NOVA
IMS
Information Management School

Imbalanced Learning

- Imbalanced Learning
 - By optimizing classification accuracy, **most algorithms assume a balanced class distribution**

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Accredited by EQUIS, AMBA, AACSB
UNIGIS
A3ES
BOM
ISchools
eduniversal

41

NOVA
IMS
Information Management School

Approaches to the Imbalanced Learning Problem

Solutions to Imbalanced Learning

Undersampling

Oversampling

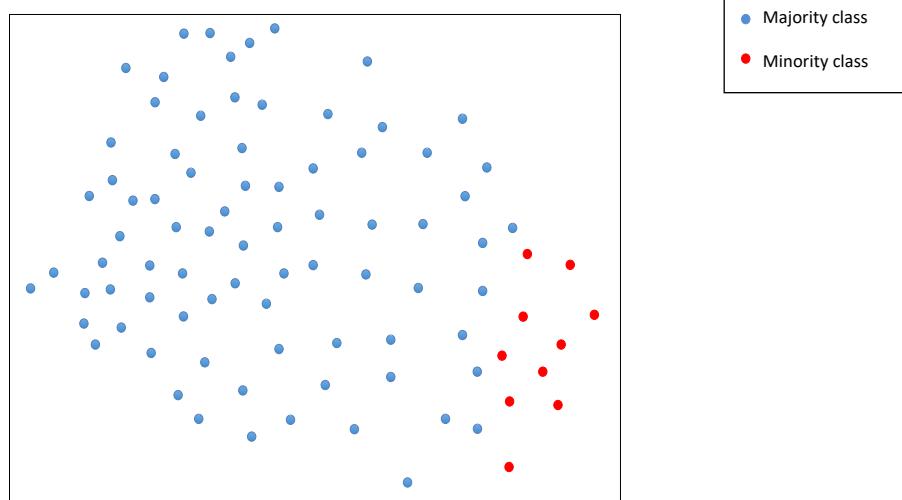
Hybrid approaches

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Accredited by EQUIS, AMBA, AACSB
UNIGIS
A3ES
BOM
ISchools
eduniversal

42

Approaches to the Imbalanced Learning Problem



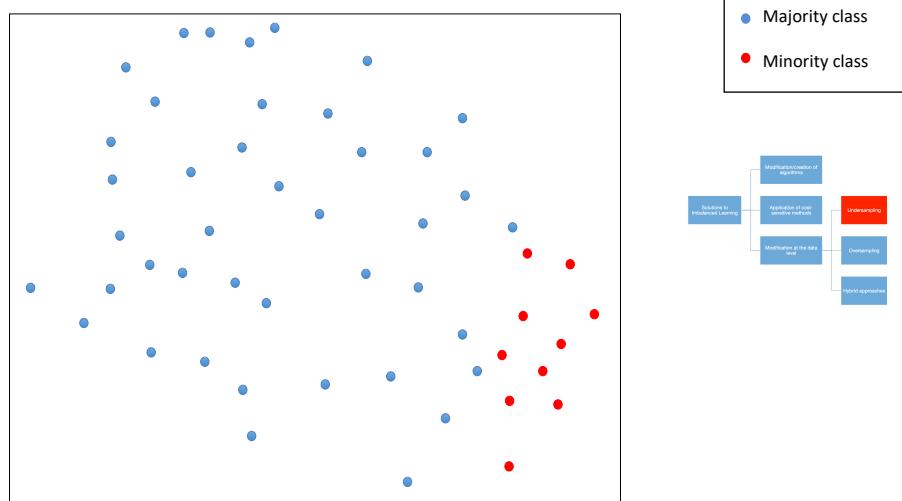
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

UNIGIS A3ES

43

Approaches to the Imbalanced Learning Problem

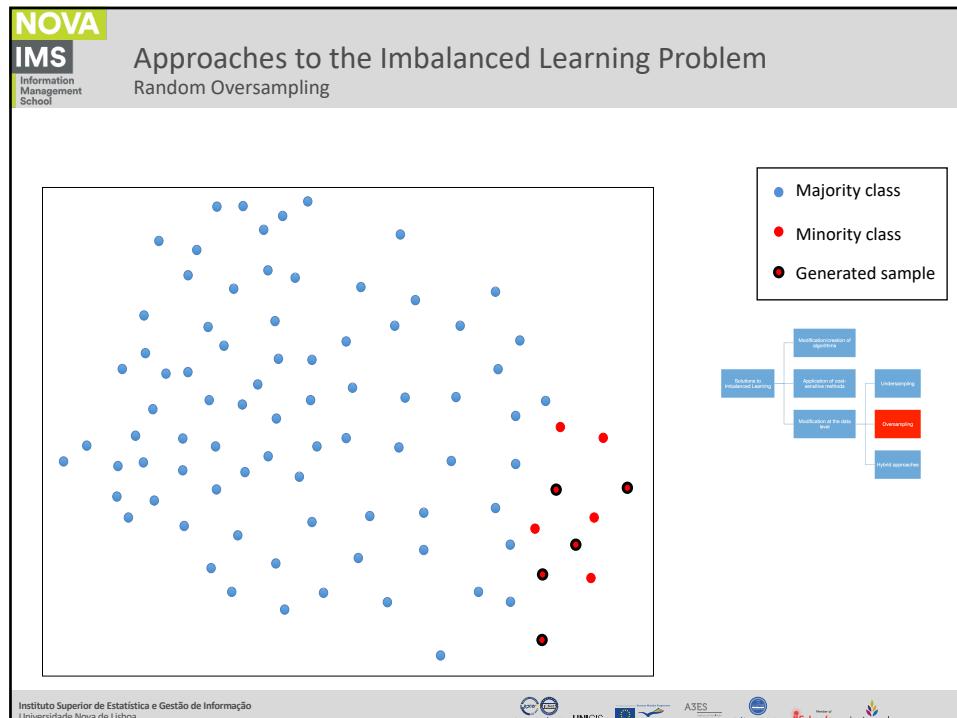
Random Undersampling



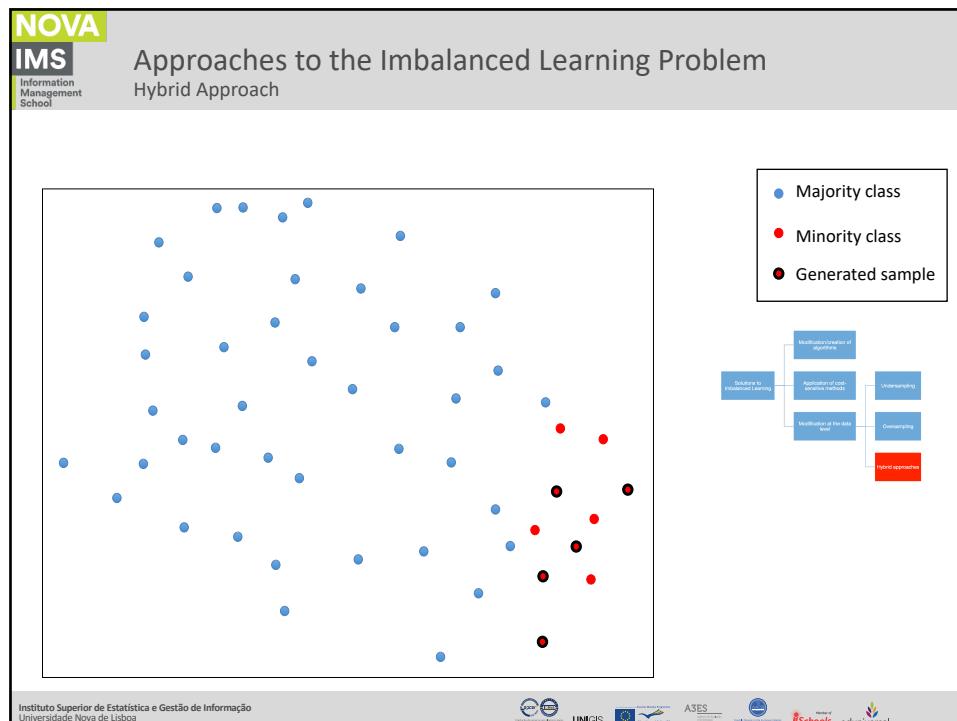
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

UNIGIS A3ES

44



45



46

NOVA
IMS
Information Management School



SMOTE: Synthetic Minority Over-sampling TEchnique

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

47

Accreditation Logos: EQUIS, UNIGIS, A3ES, AACSB, iSchools, eduniversal

47

NOVA
IMS
Information Management School

Imbalanced Learning

- SMOTE
 - The idea underlying SMOTE is **as simple as it is clever**.
 - The **basic steps** are:
 - randomly selecting a minority class instance x ;
 - then it defines the set of k-nearest neighbors (x_{knn});
 - randomly selects another minority class sample x' from the x_{knn} set.
 - x_{gen} is generated by using a linear interpolation of x and x' , which can be expressed as:

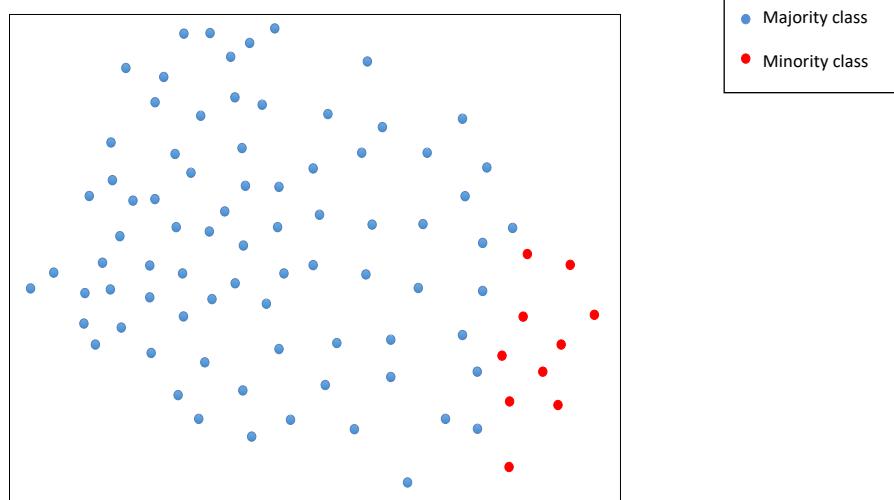
$$x_{gen} = x + a \cdot (x' - x)$$

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Accreditation Logos: EQUIS, UNIGIS, A3ES, AACSB, iSchools, eduniversal

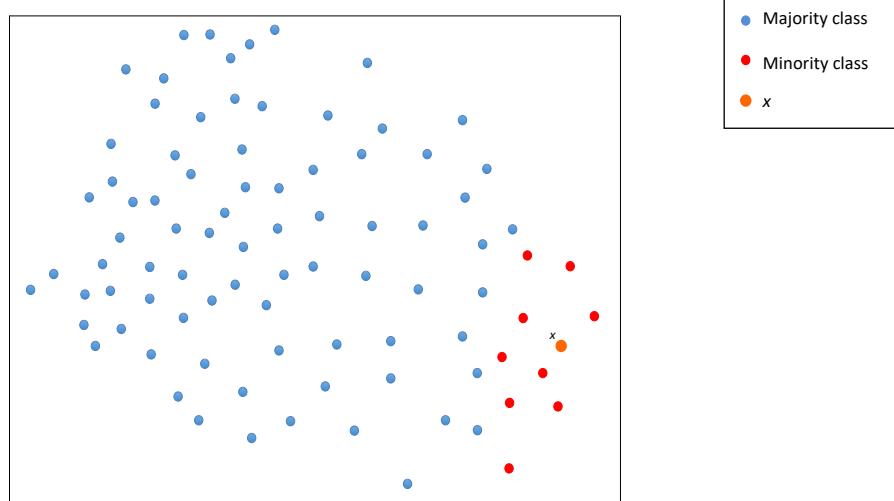
48

General aspects of data collection – acquiring knowledge

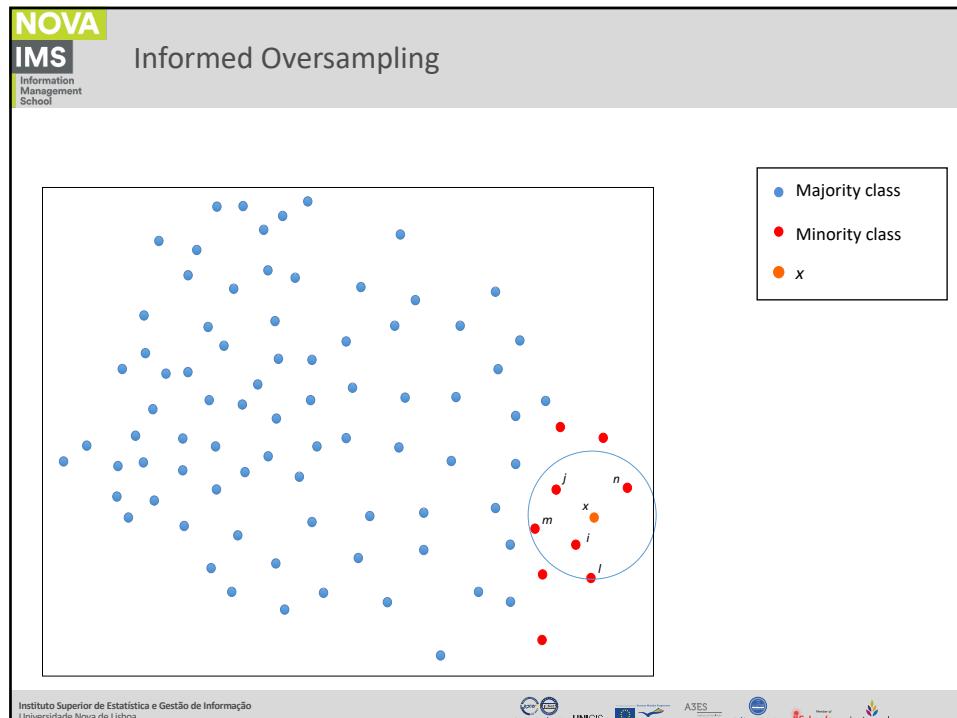
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

49

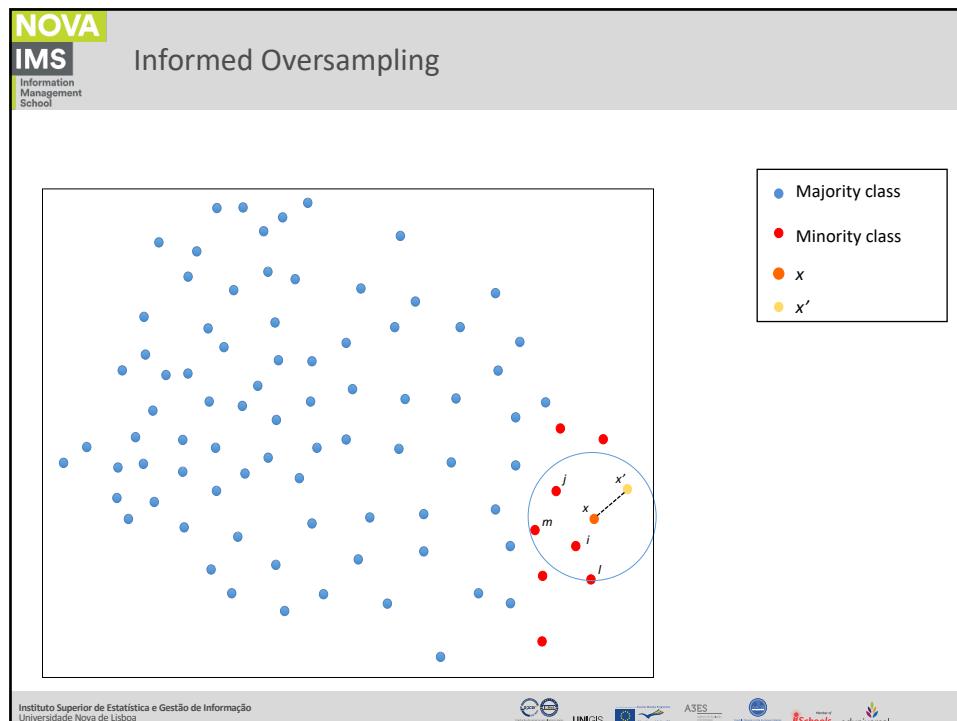
Informed Oversampling

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

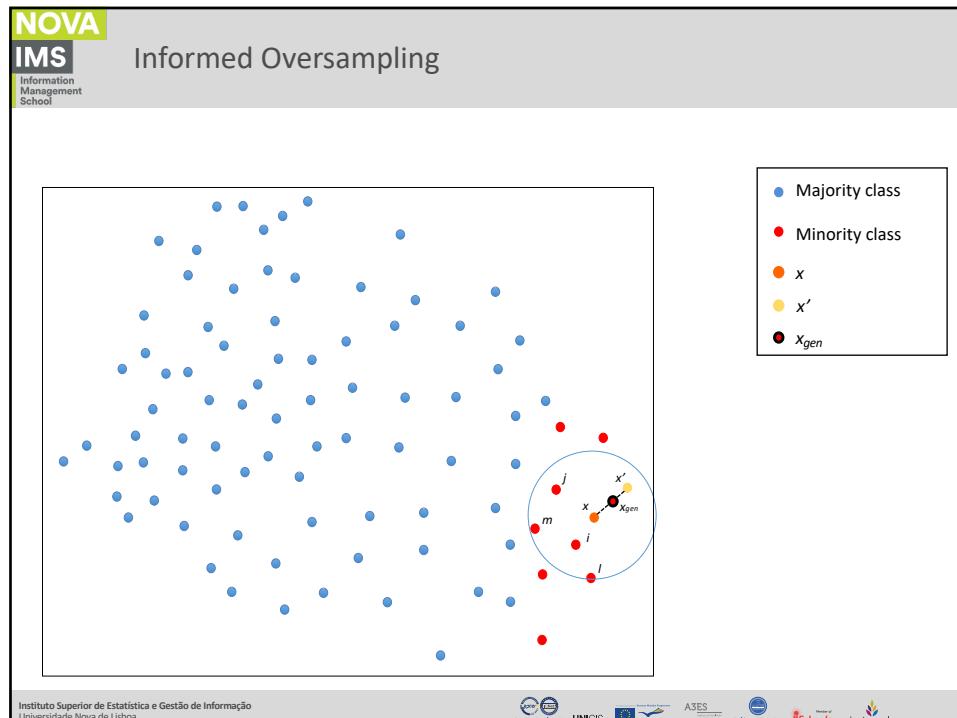
50



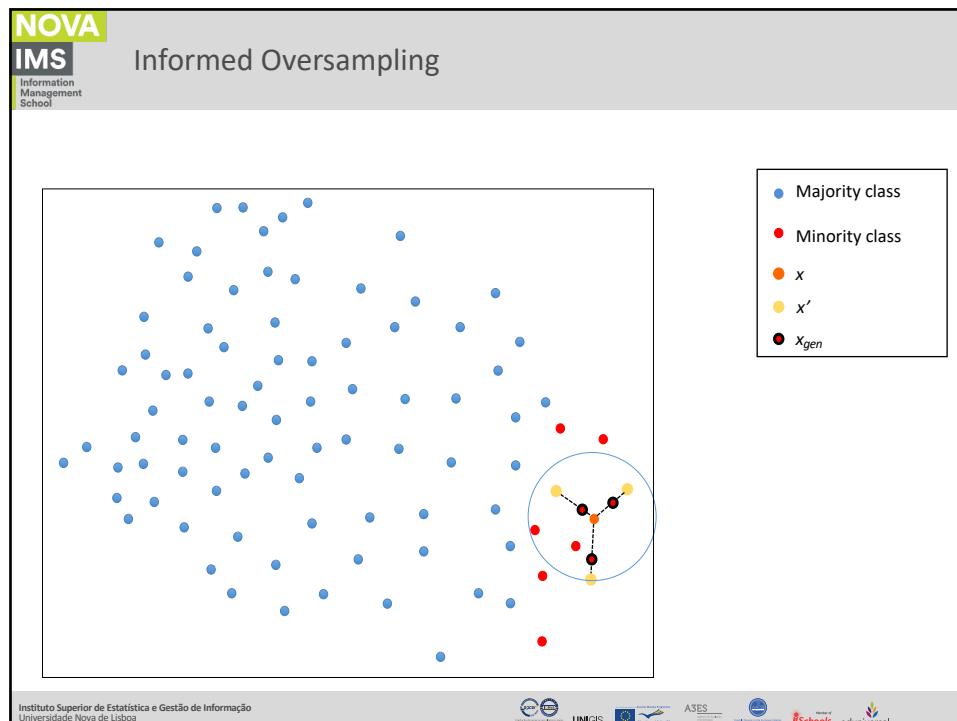
51



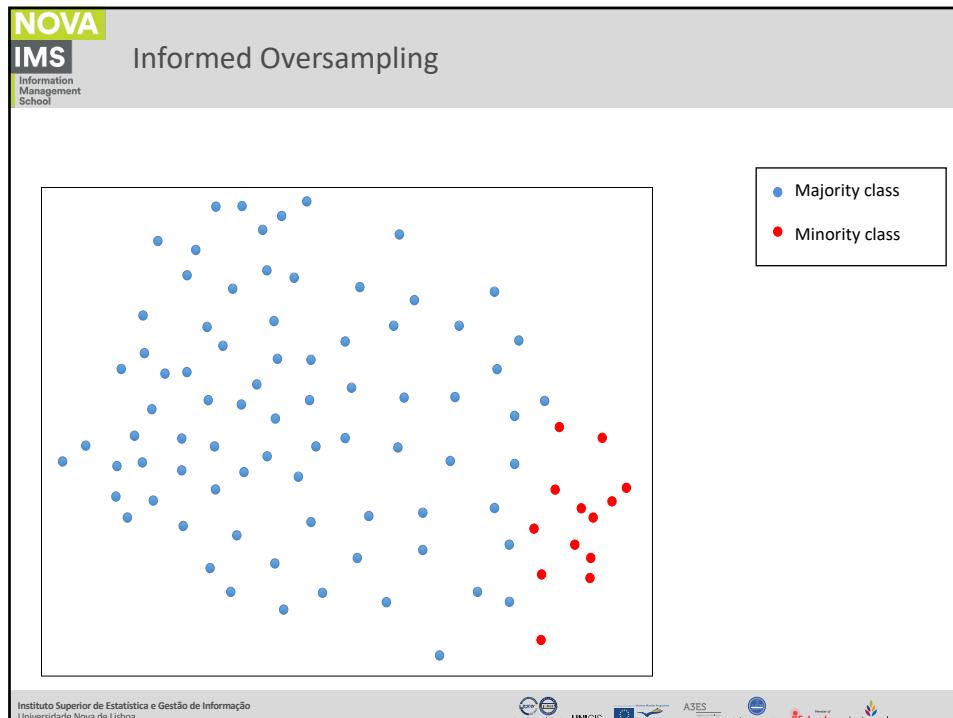
52



53



54



55

NOVA
IMS
Information Management School

General aspects of data collection

- Use of artificial data:
 - It is always preferable to use real data;
 - Create data as realistic as possible;
 - Make artificial data as representative as possible.
 - The quality of the model is constrained by the quality of the data;
 - Creating artificial data translates into the introduction of some noise.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

56

NOVA
IMS
Information Management School



Discretization

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa



57

NOVA
IMS
Information Management School

Data Preprocessing

- **Discretization**
 - Divide the range of a continuous variable into intervals
 - Some classification algorithms only accept discrete attributes
 - Reduce data size
 - Prepare for further analysis
 - Frequently called binning

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

58



Discretization

Unsupervised

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

59



Data Preprocessing

- **Discretization**
 - Unsupervised binning methods transform numerical variables into categorical counterparts but do not use the target (class) information.
 - Equal Width
 - Equal Depth (or frequency)
 - Other methods such as quantiles

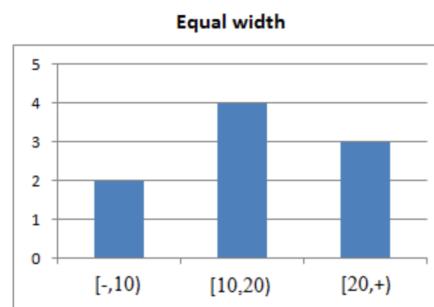
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

60

- **Discretization**

- Equal-width binning

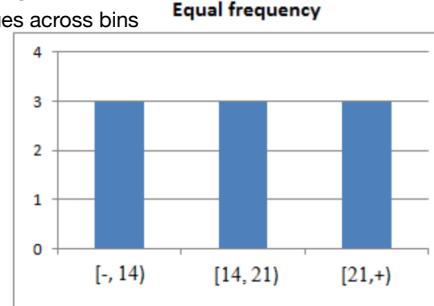
- Divides the range into n intervals of equal size
- If A and B are the minimum and the maximum values of the attribute, the width of the intervals will be: $w=(B-A)/N$
- Most simple method
- Outliers may dominate



- **Discretization**

- Equal-depth binning

- Divides the range into n intervals, each containing approximately the same number of samples
- Generally preferred avoids clumps
- Gives more intuitive breakpoints
- Shouldn't break frequent values across bins



- **Discretization**

- Class-independent methods (unsupervised)
 - Equal Width is simpler, good for many classes
 - can fail miserably for unequal distributions
 - Equal Height gives better results
- Class-dependent methods can be better for classification
 - Decision tree methods build discretization on the fly
 - Naïve Bayes requires initial discretization
- Many other methods exist ...

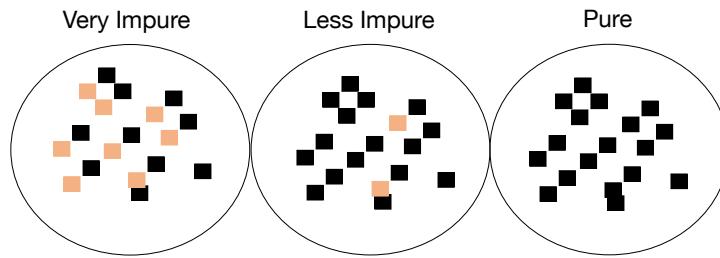


Discretization

Supervised

- **Discretization**

- Entropy (also called Expected Information) based discretization

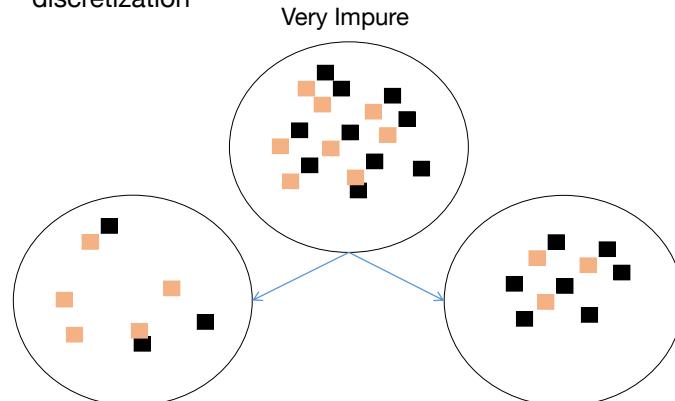


- **Discretization**

- Entropy (also called Expected Information) based discretization
 - Sort examples in increased order
 - Each value forms an interval (m intervals)
 - Calculate the entropy measure of each discretization
 - Find the binary split boundary that minimizes the entropy function over all possible partitions. The split is selected as a binary discretization
 - Apply the process recursively until some stopping criteria is met

- **Discretization**

- Entropy (also called Expected Information) based discretization



Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

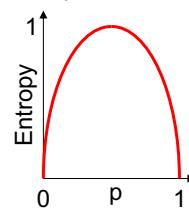
67

- **Discretization**

- Entropy based discretization
 - Entropy
 - Idea: maximize info
 - It measures the purity of a partition:

$$E = -p \log_2(p)$$

- Where p is the probability of the examples belong to a specific class



Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

68

- **Discretization**

- Entropy based discretization

	Income <= 50K	Income > 50K
Age < 25	4	6
	Income <= 50K	Income > 50K
Age < 25	9	1

- **Discretization**

- Entropy based discretization

- **Partition entropy:**

$$Ent(S) = - \sum_{i=1}^{\#C} p_i \log_2(p_i)$$

- **Gain in choosing A attribute:**

$$Gain(Ent_{new}) = Ent_{initial} - Ent_{new}$$

$$Gain(S, A) = Ent(S) - \sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v)$$

- **Discretization**

- Entropy based discretization

	Income <= 50K	Income > 50K
	13	7

$$Ent(S) = -(13/20 \log_2(13/20) + 7/20 \log_2(7/20)) = 0.403 + 0.530 = 0.934$$

- **Discretization**

- Entropy based discretization

	Income <= 50K	Income > 50K
Age < 25	4	6

$$Ent(Age < 25) = -(4/10 \log_2(4/10) + 6/10 \log_2(6/10)) = 0.529 + 0.442 = 0.971$$

	Income <= 50K	Income > 50K
Age \geq 25	9	1

$$Ent(Age \geq 25) = -(9/10 \log_2(9/10) + 1/10 \log_2(1/10)) = 0.137 + 0.332 = 0.469$$

- **Discretization**

- Entropy based discretization

	Income <= 50K	Income > 50K
Age < 25	9	1
Age ≥ 25	4	6

$$\sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v) = \frac{1}{2}(0.469) + \frac{1}{2}(0.971) = 0.72$$

$$Gain(S, A) = Ent(S) - \sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v)$$

$$Ent(S) = -(13/20 \log_2(13/20) + 7/20 \log_2(7/20)) = 0.403 + 0.530 = 0.934$$

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa