NOVA IMS Information Management School

Data Mining

S6

Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

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

1

Data Preparation

Data Preparation RoadMap

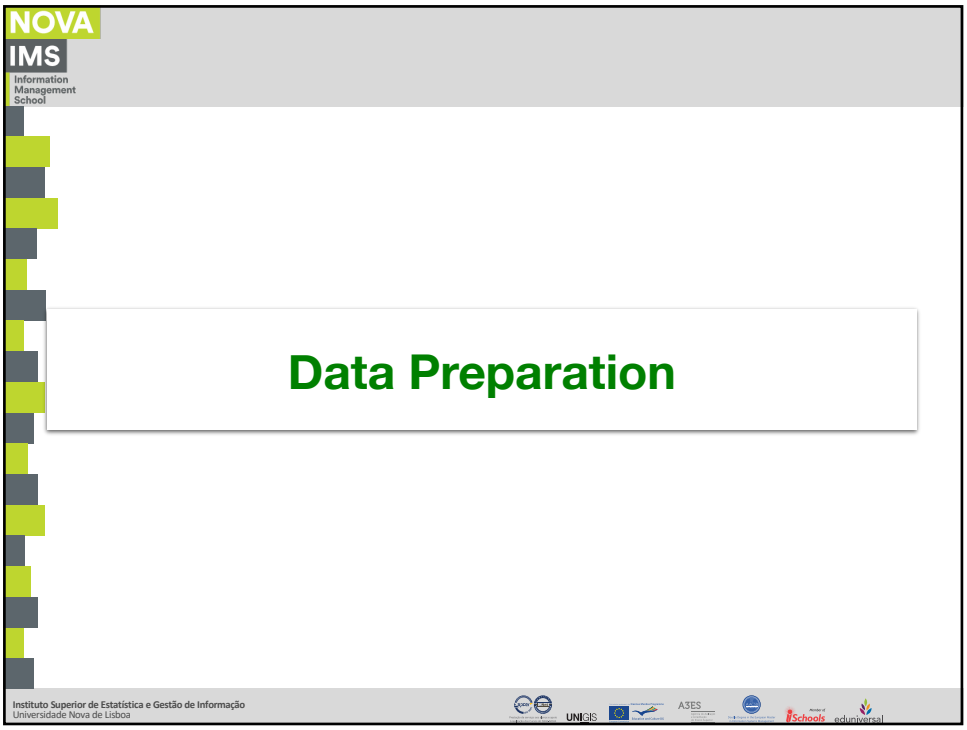
Missing Values → Outliers → Discretization and Encoding → Imbalanced Datasets

Data Pre-processing RoadMap

Feature Selection
- Relevancy analysis
- Redundancy removal

→ Feature Engineering → Feature Scaling and Normalization

2



3

Data Preparation

- **Objective**
- Transform data sets to best expose its information content
- The quality of the models should be better (or at least the same) after preparation
- Good data is a prerequisite for good models
- Some techniques are theoretically based, other are just based on experience

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

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.

5

- **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

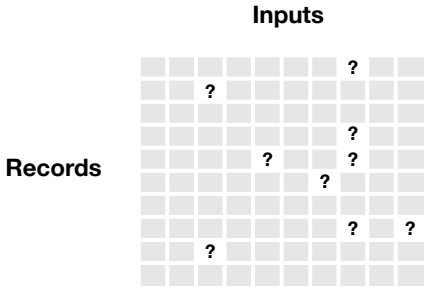
6

Treatment of Missing Data

7

Data Preparation

- **Missing Data**

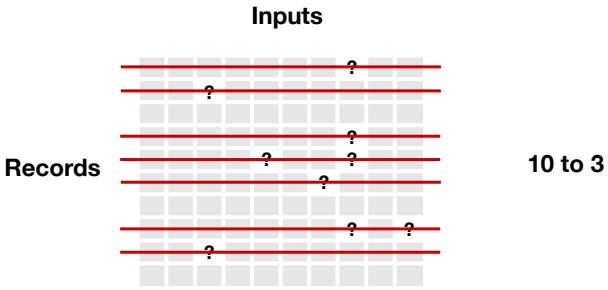


8

NOVA IMS Information Management School

Data Preparation

- **Missing Data**

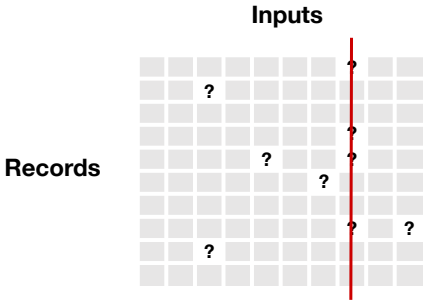


9

NOVA IMS Information Management School

Data Preparation

- **Missing Data**



10

Data Preparation

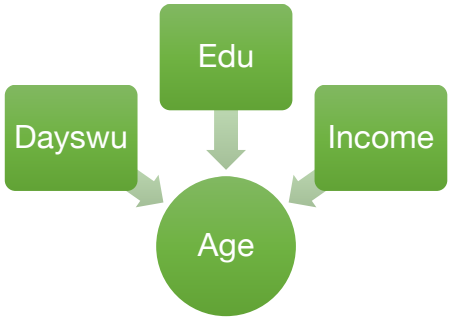
OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

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

11

Data Preparation

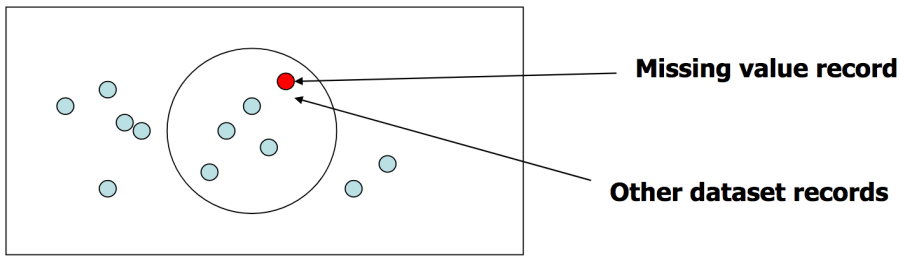
CUSTID	DAYSWUS	AGE	EDU	INCOME
1138	951	29	17	55637
6313	1231	22	16	32795
9185	1186	31	18	40717
7735	1097	27	18	29184
7918	1196	19	13	16479
1585	849	34	19	55403
7897	1120	46	15	63603
7389	792	39	16	57402
6764	1060	missing	15	69803
8416	1217	31	18	56028
6541	1074	54	15	91934
8263	1185	46	18	78262
8052	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



12

NOVA IMS Information Management School

Data Preparation



13

NOVA IMS Information Management School

Data Preparation

- **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.

14

NOVA IMS Information Management School

Data Preparation

- **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 an 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.

15

NOVA IMS Information Management School

Outlier treatment

16

NOVA IMS Information Management School

Data Preparation

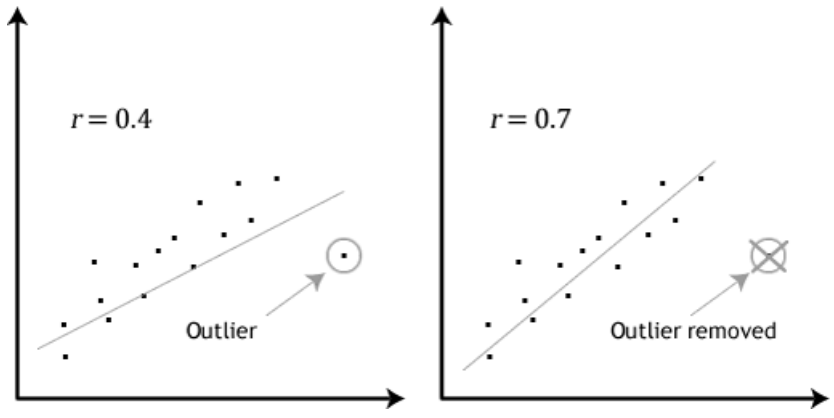
- **Outliers**
  - In statistics, an outlier is an observation point that is distant from other observations.
  - An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.
  - Extreme cases in one or more variables 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.

17

NOVA IMS Information Management School

Data Preparation

- **Outliers (leverage effect)**

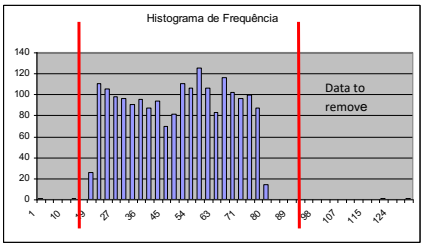


18

NOVA IMS Information Management School

Data Preparation

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

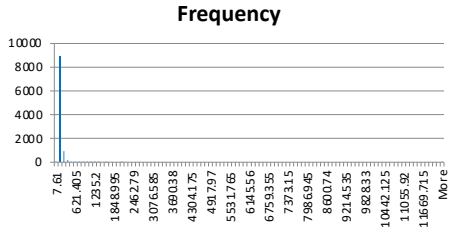


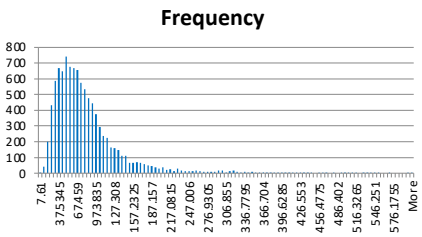
19

NOVA IMS Information Management School

Data Preparation

- **Detect Data Outliers**





20

NOVA IMS Information Management School

Data Preparation

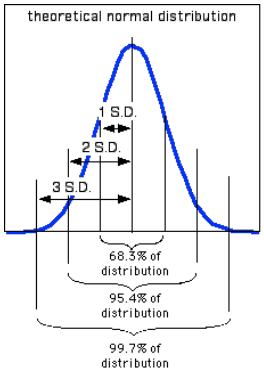
- **Detect Data Outliers**
  - **When Examining Your Histogram:**
    - Are there isolated bars separated from the main distribution?
    - Is there a long tail stretching much farther than expected?
    - Do some bins have very few observations far from the center?
    - Is the shape asymmetric with one side much longer?
    - Are there gaps between the main data and extreme values?
  - **Next Steps After Detection:**
    - Investigate the outlier values - are they errors or real?
    - Check domain knowledge - do these make sense?
    - Decide on treatment - remove, cap, transform, or keep?
    - Document your decisions for reproducibility

21

NOVA IMS Information Management School

Data Preparation

- **Detect Data Outliers**
  - Normal data distribution
    - 3σ +/- Average

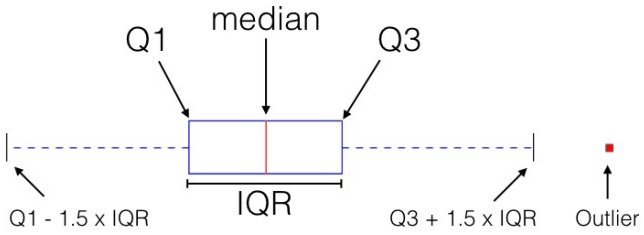


House Prices in a Neighborhood:

- Mean price: $500,000
- Standard deviation: $100,000
- Normal range: $500,000 +/- $300,000 = $200,000 to $800,000
- Outliers: Houses below $200,000 or above $800,000

22

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*, Q3 - Q1

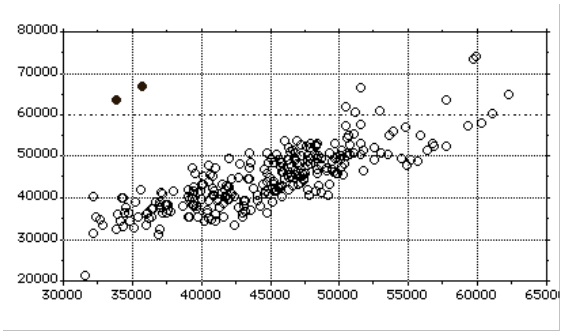
Outliers: Data points are considered to be outliers if
value < *Q1 - 1.5 x IQR* or
value > *Q3 + 1.5 x IQR*

fmi:
http://sebastianraschka.com/Articles/2014_dixon_test.html
http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

23

Data Preparation

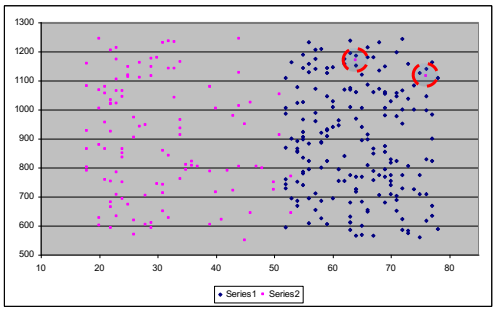
- **Detect Data Outliers**
  - In two dimensions…



24

Data Preparation

- **Detect Data Outliers**
  - In three dimensions…



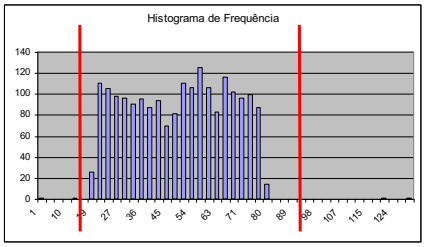
25

Data Preparation

- **Detect Multidimensional Outliers**
  - **k-NN Distance**: Points with large average distance to their k nearest neighbors
  - **Isolation Forest**: Points that require fewer random splits to isolate from the rest of the data
  - **Clustering Methods**: Points that are far from cluster centers or belong to very small clusters
  - **Self-Organizing Maps**: Points that map to neurons with large quantization errors or low activation frequencies
  - **Dimensionality Reduction Tools**: Points that appear separated or distant from the main data when visualized in reduced dimensions

26

- **Outlier Treatment: Capping (Winsorizing)**
  - Instead of removing outliers, put a "ceiling" and "floor"
  - Example: If most salaries are $30k-$100k, but you have one at $1 million, cap it at $200k

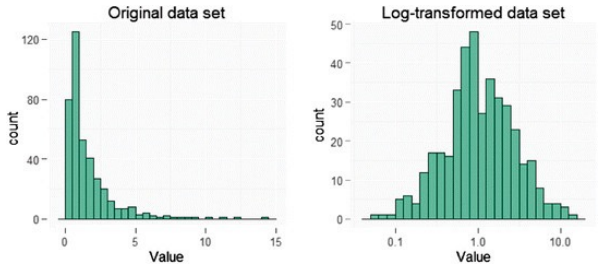


27

- **Outlier Treatment: Transformation Methods**
  - **When to Use Transformations:**
    - When you want to keep all data points but reduce outlier impact
    - For right-skewed data with extreme high values
    - When building linear models that assume normality

28

- **Outlier Treatment: Transformation Methods**
  - **Log Transformation** **(others like Square Root)**
    - Best for: Income, prices, population counts
    - Effect: Compresses large values, expands small values



29

- **Outlier Treatment: Transformation Methods**
  - **Pros:**
    - Preserves all data points
    - Can improve model performance
    - Helps meet statistical assumptions
  - **Cons:**
    - Reduces interpretability (log dollars?)
    - Must reverse-transform for predictions
    - Doesn't work for all distributions

30

- **Outlier Treatment: Transformation Methods**
  - **When NOT to Transform:**
    - Outliers represent important rare events (fraud, defects)
    - Using tree-based models (robust to outliers)
    - Interpretability is more important than model performance

31

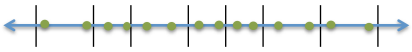
Discretization

32

- **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
  - Divided into unsupervised (no target) or supervised (target)

- **Discretization (unsupervised)**
  - 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 (unsupervised)**
  - 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



35

- **Discretization (unsupervised)**
  - How to choose number of bins
    - Square Root Rule:
      - Count how many data points you have
      - Take the square root
      - Example: 1000 data points > sqrt 1000 ≈ 32 bins
    - Rice Rule:
      - Use: $2 \times n^{1/3}$ (cube root)
      - Example: 1000 data points > $2 \times 10 = 20$ bins
    - Business Common Sense:
      - Ages: Child (0-12), Teen (13-19), Adult (20-64), Senior (65+)
      - Income: Low (<$30K), Medium ($30K-70K), High ($70K-150K), Very High (>$150K)
      - Grades: A (90-100), B (80-89), C (70-79), D (60-69), F (<60)

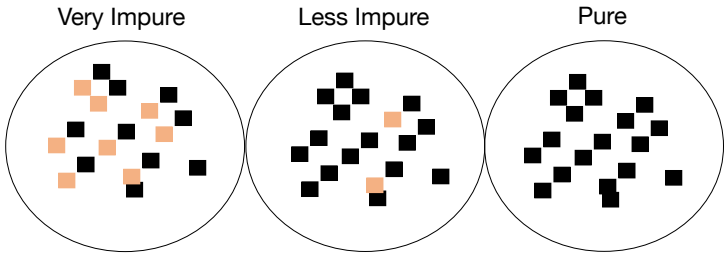
36

- **Discretization (unsupervised)**
  - How to choose number of bins
  - Very general guidelines

Dataset Size	Recommended Bins
< 100	5-7 bins
100-1,000	10-15 bins
1,000-10,000	15-25 bins
> 10,000	20-30 bins

- **Discretization (unsupervised)**
  - EQUAL-WIDTH:
    - Simple and fast
    - Good for uniformly distributed data
    - Sensitive to outliers
  - EQUAL-DEPTH:
    - Handles outliers better
    - More intuitive bins
    - May break natural clusters

- **Discretization (supervised)**
  - Entropy (also called Expected Information) based discretization

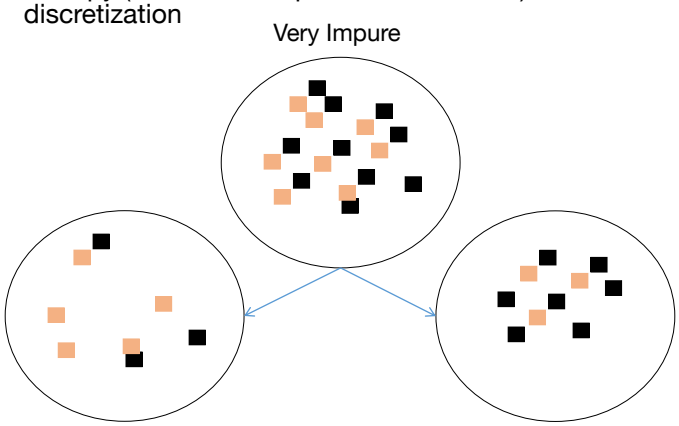


39

- **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

40

- **Discretization**
  - Entropy (also called Expected Information) based discretization

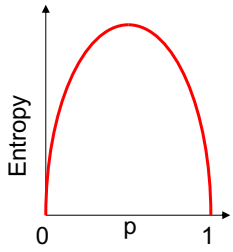


41

- **Discretization**
  - Entropy based discretization
  - In information theory, entropy is a measure of uncertainty associated with a random variable or a probability distribution.
    - High entropy > high degree of uncertainty, randomness, or unpredictability,
    - Low entropy > high degree of certainty or predictability.
  - Entropy is typically measured using the formula:

$$E = - p \log2(p)$$

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



42

- **Discretization**
  - Entropy based discretization

	Income <= 50K	Income > 50K
Age	13	7

	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

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

- **Discretization**
  - Entropy based discretization

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

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

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

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

- **Discretization**
  - Entropy based discretization

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

$$\sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v) = \frac{1}{2}(0.469) + \frac{1}{2}(0,971) = 0.72$$

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

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

47

Encoding

48

- **Encoding**
  - Converting categorical data (text, categories) into numerical format that machine learning algorithms can understand.
    - One-Hot Encoding (Dummy Encoding)
    - Creating binary (0/1) columns for each category.

Original: Red, Blue, Green

One-Hot Encoded:

Color_Red	Color_Blue	Color_Green
1	0	0
0	1	0
0	0	1

49

Imbalanced Learning

50

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?		

51

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	

52

NOVA IMS Information Management School

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	

53

NOVA IMS Information Management School

- 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%**

54

Imbalanced Learning

- 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)

55

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.

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

56

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**

57

Imbalanced Learning

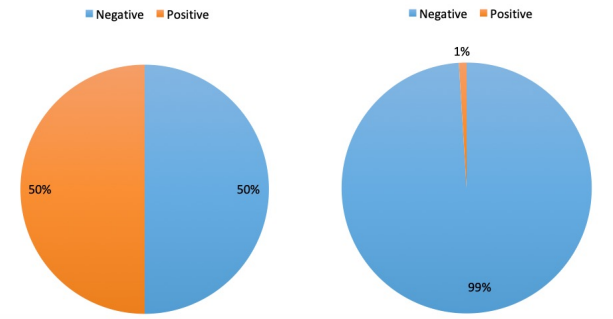
- Imbalanced Learning
  - **Standard learning methods induce a bias** in favor of the majority class during training.
  - This happens because the **minority classes contribute less to the maximization** of the objective function, which is usually accuracy.

$$Accuracy = \frac{0 + 999}{1000} = 0{,}999$$

58

Imbalanced Learning

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



59

Imbalanced Learning

- Imbalanced Learning
  - Another inherent **assumption** of many classication algorithms is the **uniformity of misclassication costs**

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

  - Frequently the costs of misclassifying the **minority class are much higher** than the costs of misclassification of the majority class

60

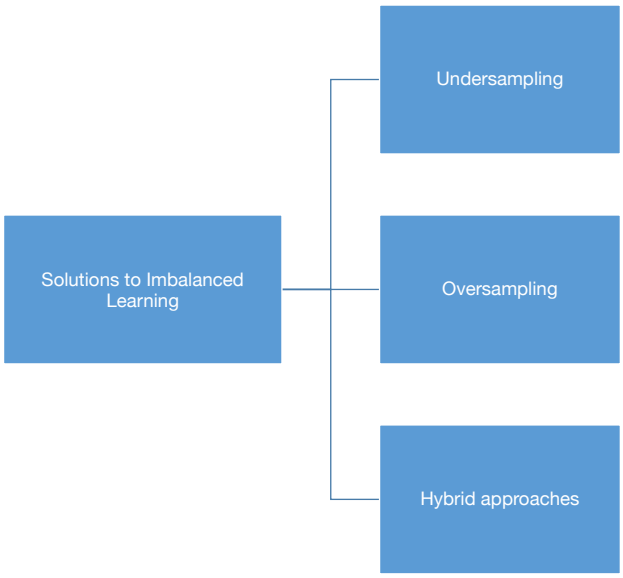
Imbalanced Learning

- Imbalanced Learning
  - **Diseases screening tests are a typical situation** in in which false negatives involve a much higher cost than the false positives.

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

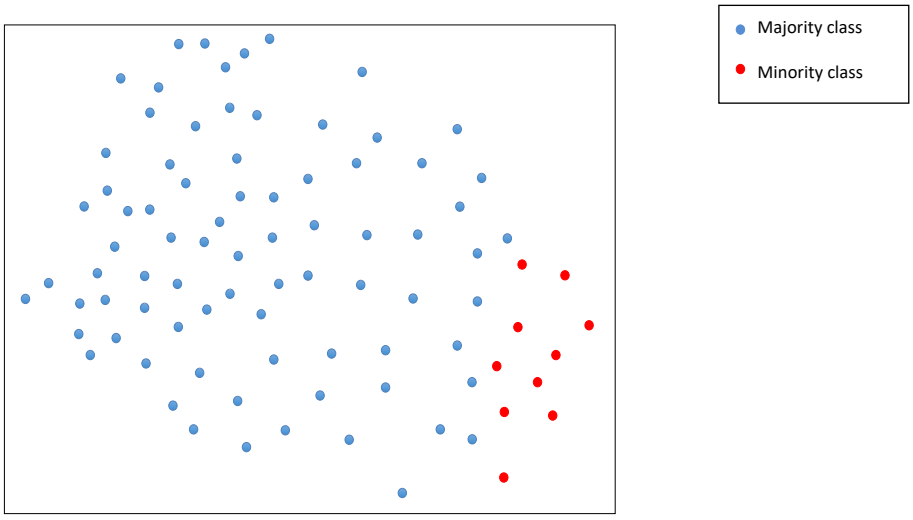
61

Approaches to the Imbalanced Learning Problem



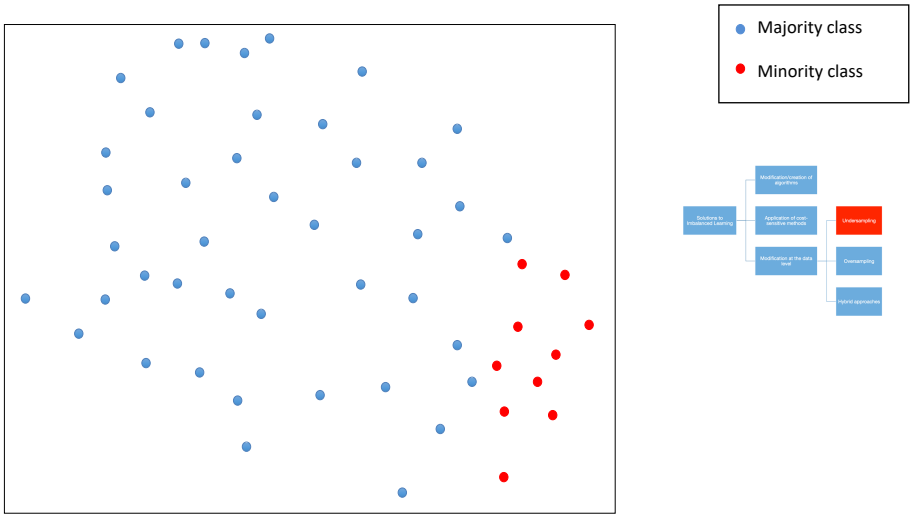
62

Approaches to the Imbalanced Learning Problem



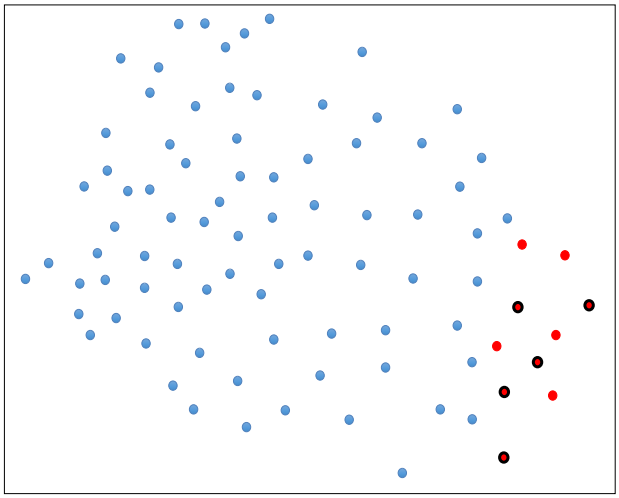
63

Approaches to the Imbalanced Learning Problem
Random Undersampling

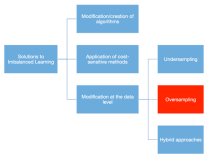


64

Approaches to the Imbalanced Learning Problem
Random Oversampling

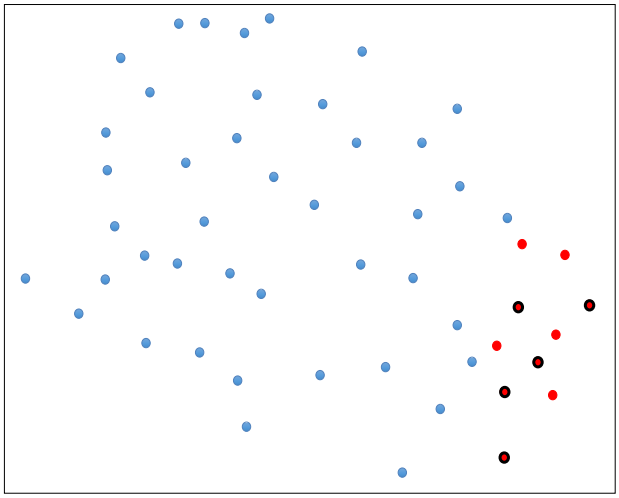


Majority class
Minority class
Generated sample

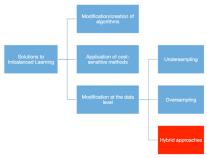


65

Approaches to the Imbalanced Learning Problem
Hybrid Approach



Majority class
Minority class
Generated sample



66

NOVA IMS Information Management School

SMOTE: Synthetic Minority Over-sampling TEchnique

67

NOVA IMS Information Management School

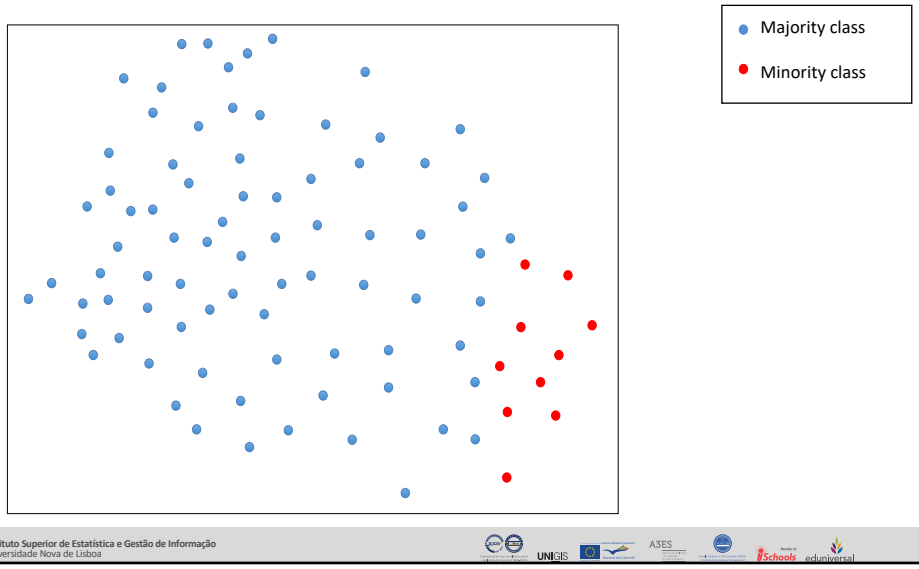
Imbalanced Learning

- 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:

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

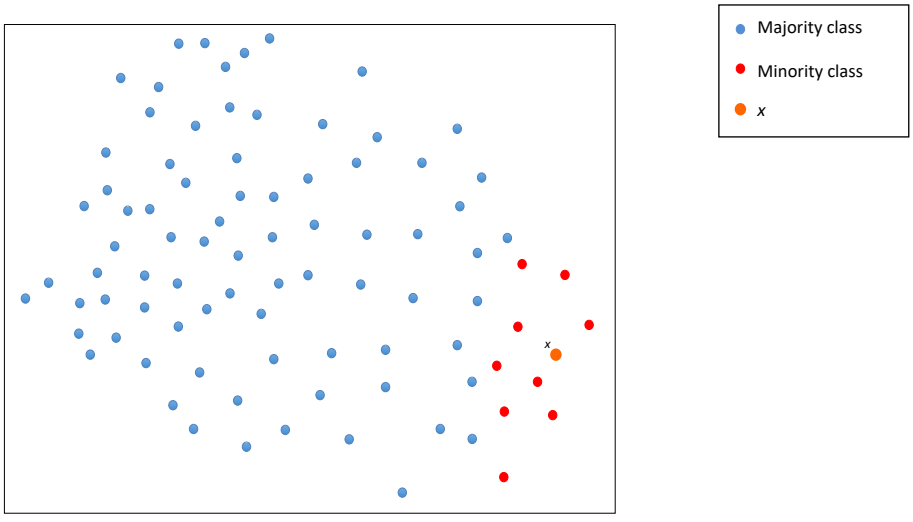
68

General aspects of data collection – acquiring knowledge



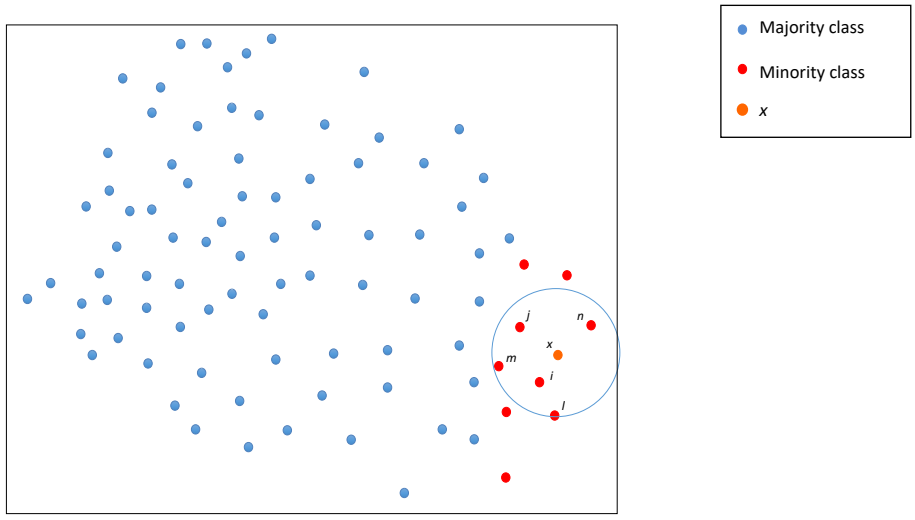
69

Informed Oversampling



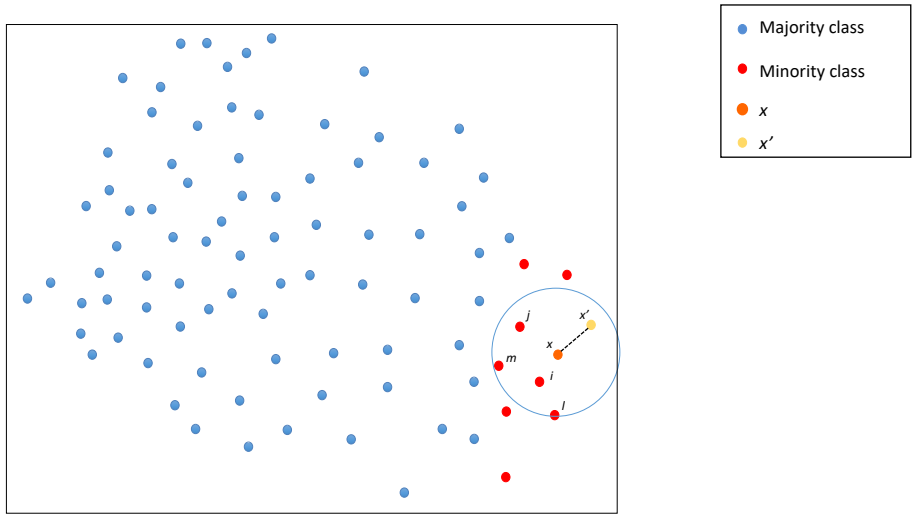
70

Informed Oversampling



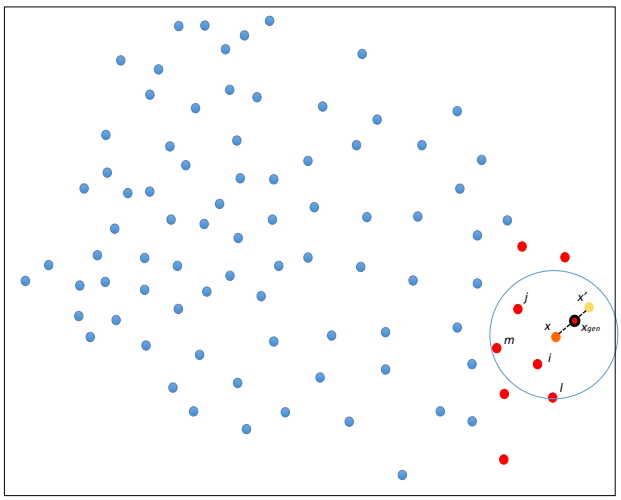
71

Informed Oversampling



72

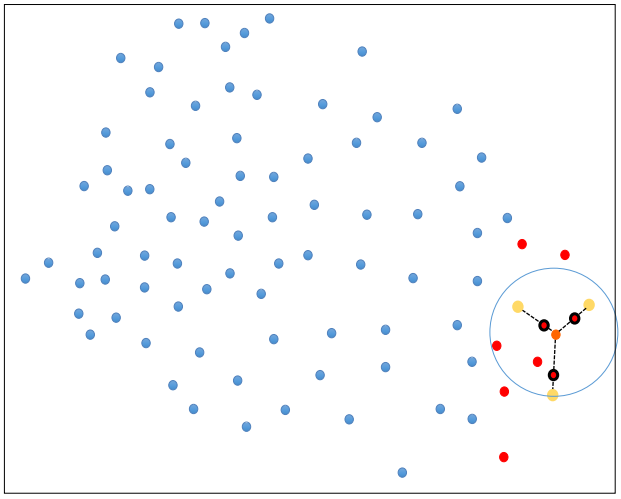
Informed Oversampling



Majority class
Minority class
x
x'
x_{gen}

73

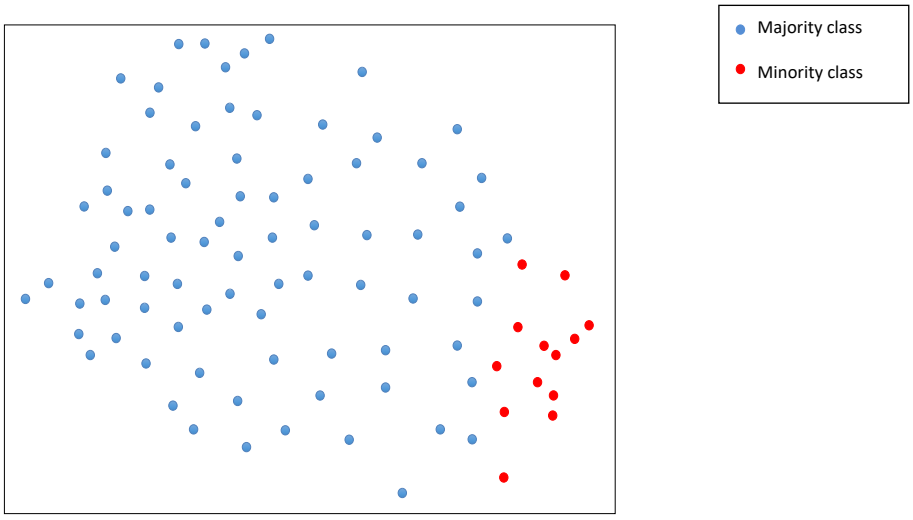
Informed Oversampling



Majority class
Minority class
x
x'
x_{gen}

74

Informed Oversampling



75

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.

76

77