Undersampling and Oversampling

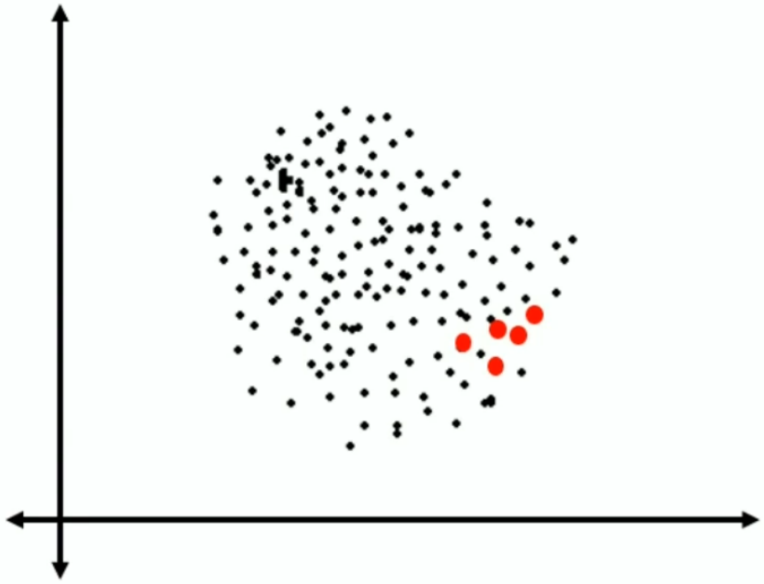
Víctor Acevedo Vitvitskaya

Dealing with imbalanced dataset

- Presence of minority class in the dataset

- Challenges related Imbalanced Dataset
  - Biased predictions
  - Misleading accuracy

- Some Examples
  - Credit card frauds
  - Manufacturing defects
  - Rare diseases diagnosis
  - Natural disasters
  - Enrolment to premier institutes



Two Class Classification

No-Fraud → 99.5%
Fraud → 0.5%

Re-sample the dataset

- Balance the classes by Increasing minority or decreasing majority

- Random Under-Sampling
  - Randomly remove majority class observations
  - Helps balance the dataset
  - Discarded observations could have important information
  - May lead to bias

Total Observations = 1,000
Fraudulent = 10 or 1%
Normal = 990 or 99%

Reduce normal to 90
Fraudulent = 10 or 10%

- Random Over-Sampling
  - Randomly add more minority observations by replication
  - No information loss
  - Prone to overfitting due to copying same information

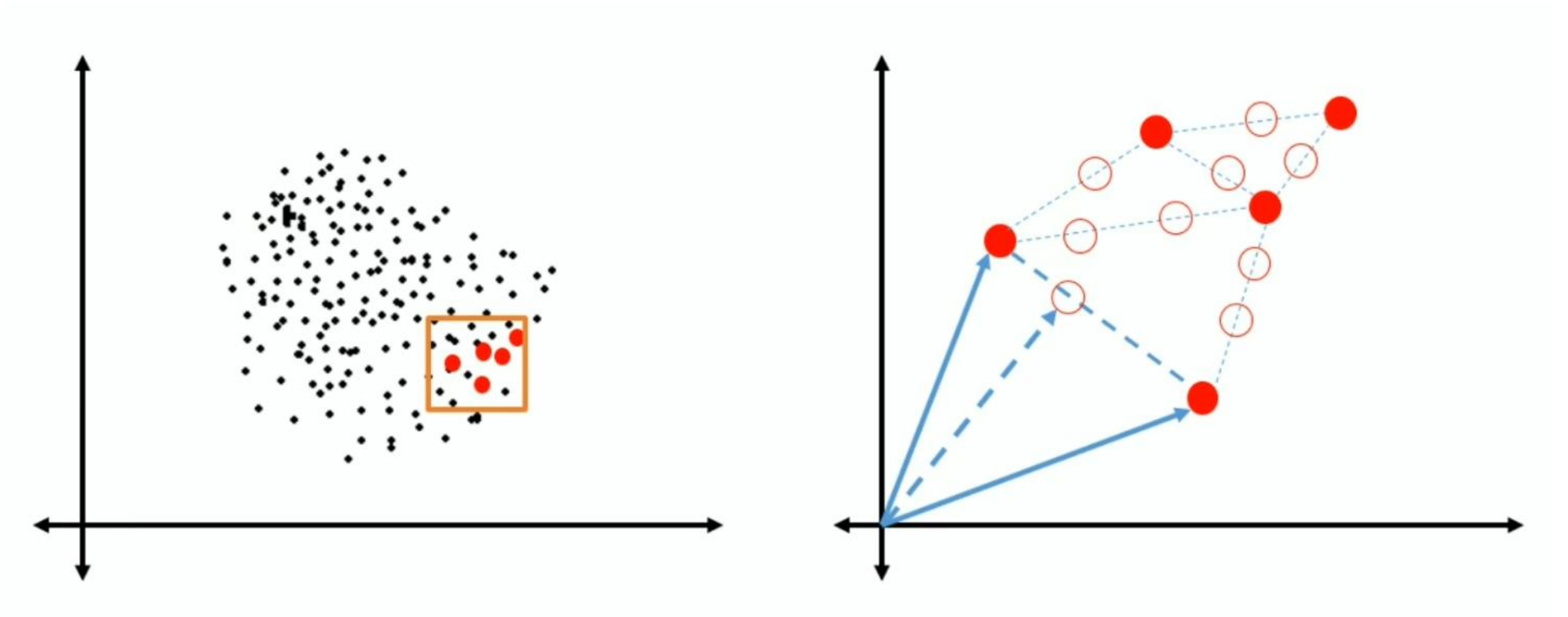
Total Observations = 1,000
Fraudulent = 10 or 1%
Normal = 990 or 99%

Increase fraudulent by 100
Fraudulent 110 or 10%

SMOTE

- Synthetic Minority Oversampling Technique
- Creates new "Synthetic" observations
- SMOTE Process
  - Identify the feature vector and its nearest neighbour
  - Take the difference between the two
  - Multiply the difference with a random number between 0 and 1
  - Identify a new point on the line segment by adding the random number to feature vector
  - Repeat the process for identified feature vectors

SMOTE



Caso Aplicado

Input variables:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

Caso Aplicado

other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')