Data Curation Techniques

Siddharth R

Data Reduction

- Numerosity reduction vs Dimensionality reduction
  - Sampling
  - Clustering
- Attribute subset selection / feature selection
  - Filter
  - Wrapper
  - Embedded
  - Forward Selection vs Backward Elimination
- Feature Extraction
  - Principal Component Analysis

Sampling

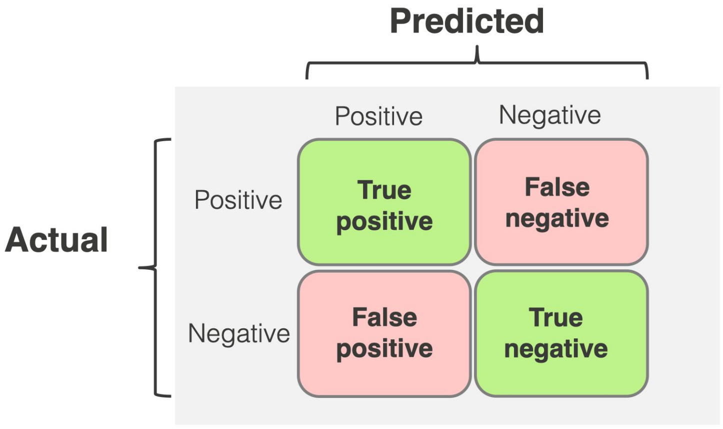
- Importance of sampling
  - The "Literary Digest" disaster (1936) - Selection Bias and Non-Responsive Bias
  - TRP Scam in India (2020)

Common Types:

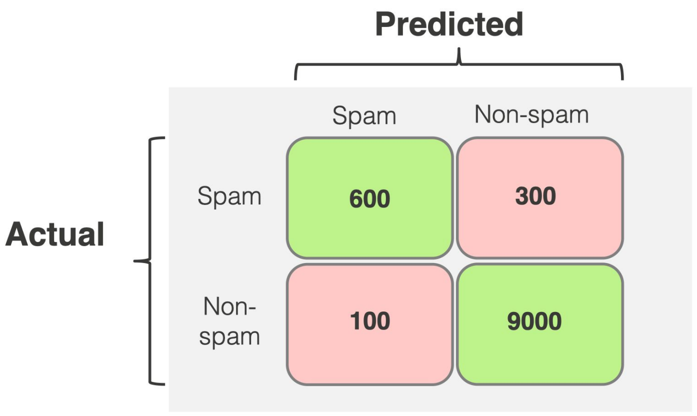
1. Simple random sampling
   - With replacement
   - Without replacement
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling

Imbalanced data & its Impact

- Imbalanced data refers to the situation where the distribution of classes in the target variable is not equal.
- Misleading Performance Metrics

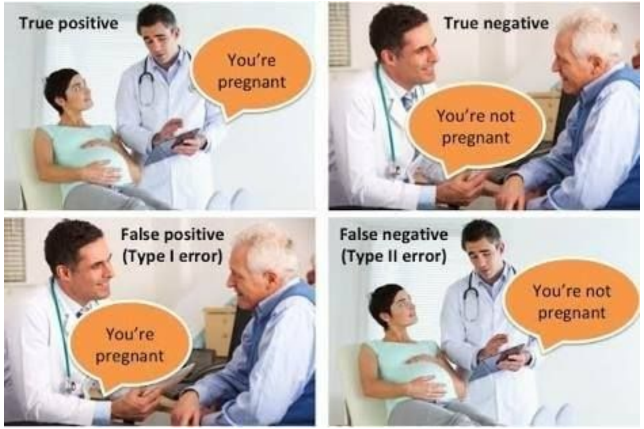


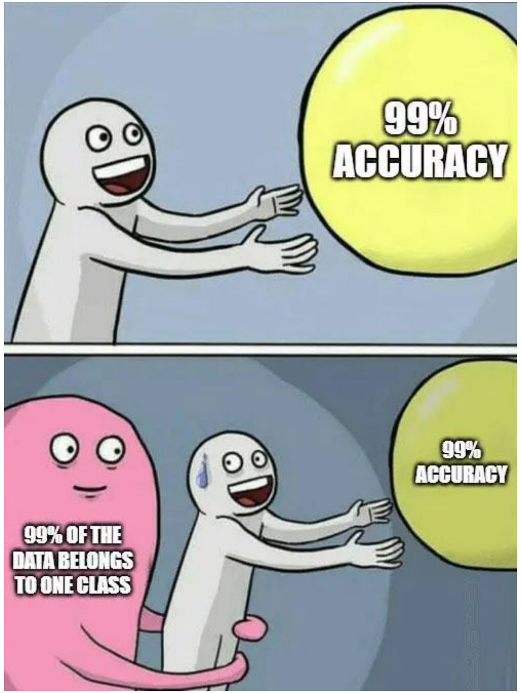
		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative



		Predicted	
		Spam	Non-spam
Actual	Spam	600	300
	Non-spam	100	9000

Handling Imbalanced Data





Ways to Handle Imbalanced Data

- Random Oversampling

  Randomly replicates samples from the minority class to balance the dataset.

- Random Undersampling

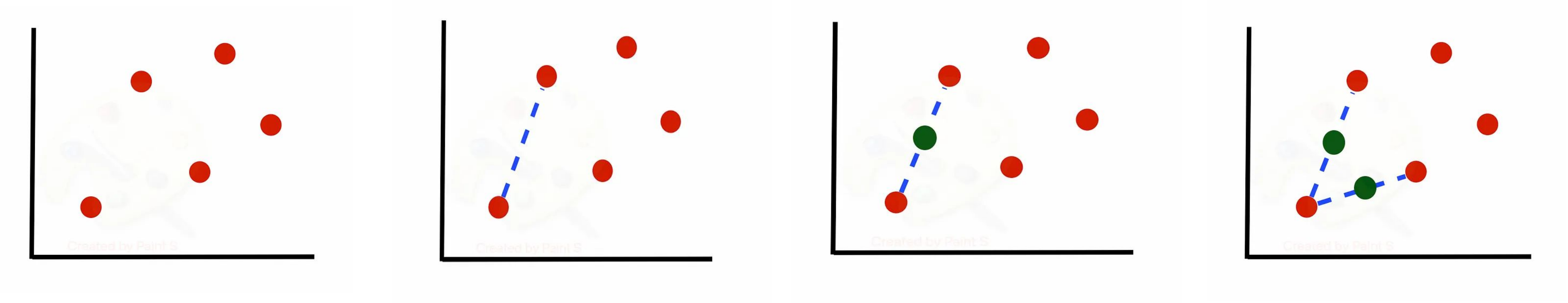
  Randomly removes samples from the majority class to balance the dataset.

- SMOTE (Synthetic Minority Over-sampling Technique)

  Generates synthetic samples for the minority class based on feature space similarities.

How SMOTE works

- Identify k-nearest neighbor
- A synthetic data point is created somewhere between the chosen data point and its neighbor. This is done by interpolation
- Example: New Sample=Sample+Random×(Neighbor−Sample)



Data Reduction

- A large data matrix *D* of *m* rows and *n* columns is reduced to D' with *m' << m* or *n' << n* . Computing on D' should give results similar to computing on D.
- Reducing 'm' is Numerosity reduction and reducing 'n' is Dimensionality Reduction
- Broadly divided into (i) Feature Selection (ii) Feature Extraction
- Feature Selection : Choosing a subset of the most relevant features from the original dataset.
- Feature Extraction : Transforming the original features into a new, reduced set of features while retaining the essential information.

Attribute Subset / Feature Selection

1. **Filter Methods**
   - Selects features independently of the model. Uses statistical techniques to evaluate feature importance.
   - Techniques: Correlation: Remove highly correlated features, Chi-Square Test: For categorical data, Mutual Information: Measures dependency between variables.
   - Cons: Ignores feature interactions, Independent of the model’s performance
2. **Wrapper Methods**
   - Uses the model’s performance (like accuracy) to select features. Tries different feature subsets and evaluates results.
   - Forward Selection: Add features one by one.
   - Backward Elimination: Start with all features, remove one at a time.
   - Cons: Computationally expensive, Prone to overfitting
   - For n attributes, how many subsets will be there?

Attribute Subset Selection

3. **Embedded Method**

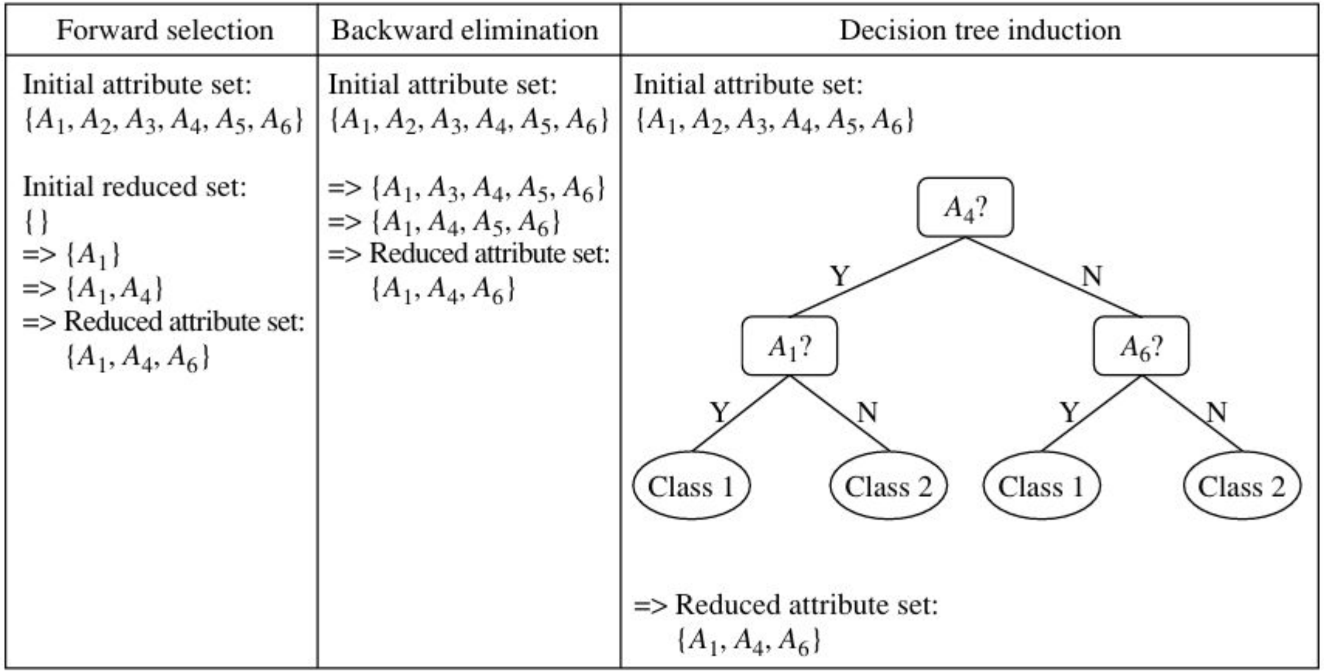
- Feature selection happens during model training. Models like decision trees have built-in feature selection.

- Cons: Only works with models having built-in feature selection

When to Use What?

- Filter: Quick, when you have many features and want to reduce dimensionality fast.
- Wrapper: When model performance is critical and you have time/resources to evaluate subsets.
- Embedded: When using models like Lasso, Decision Trees, or Boosting that naturally handle feature importance.

Example

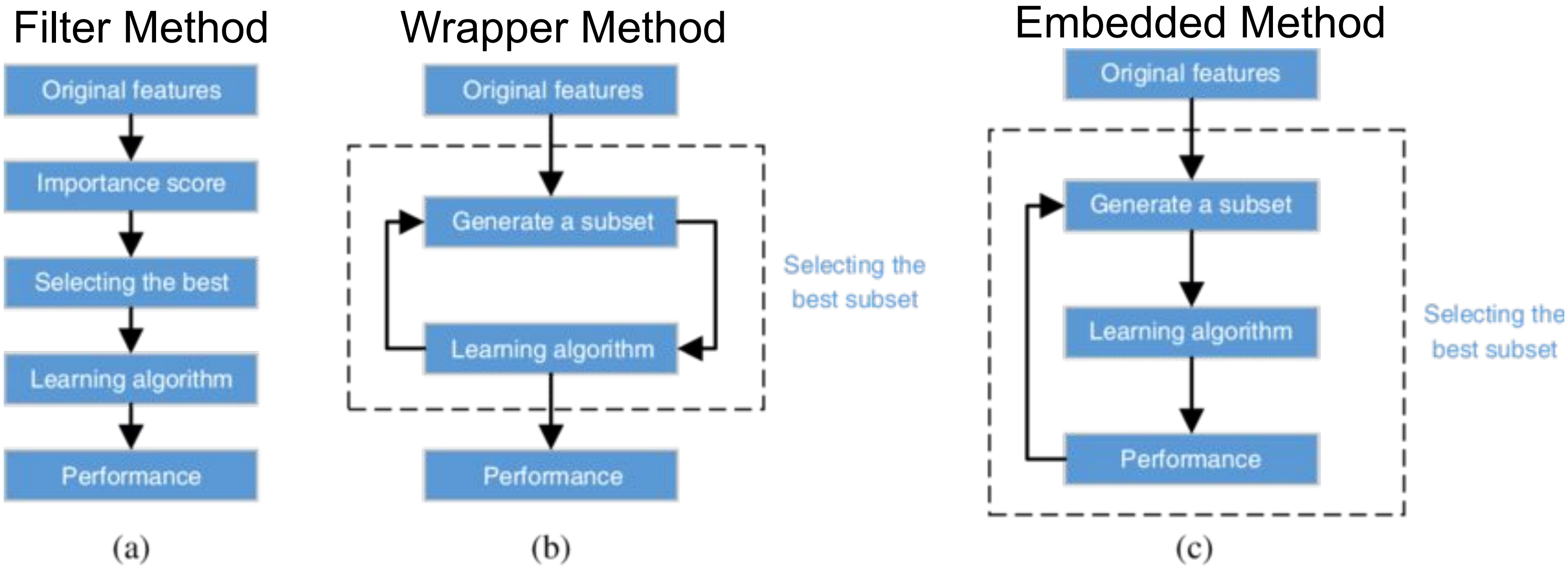


Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: { } => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$

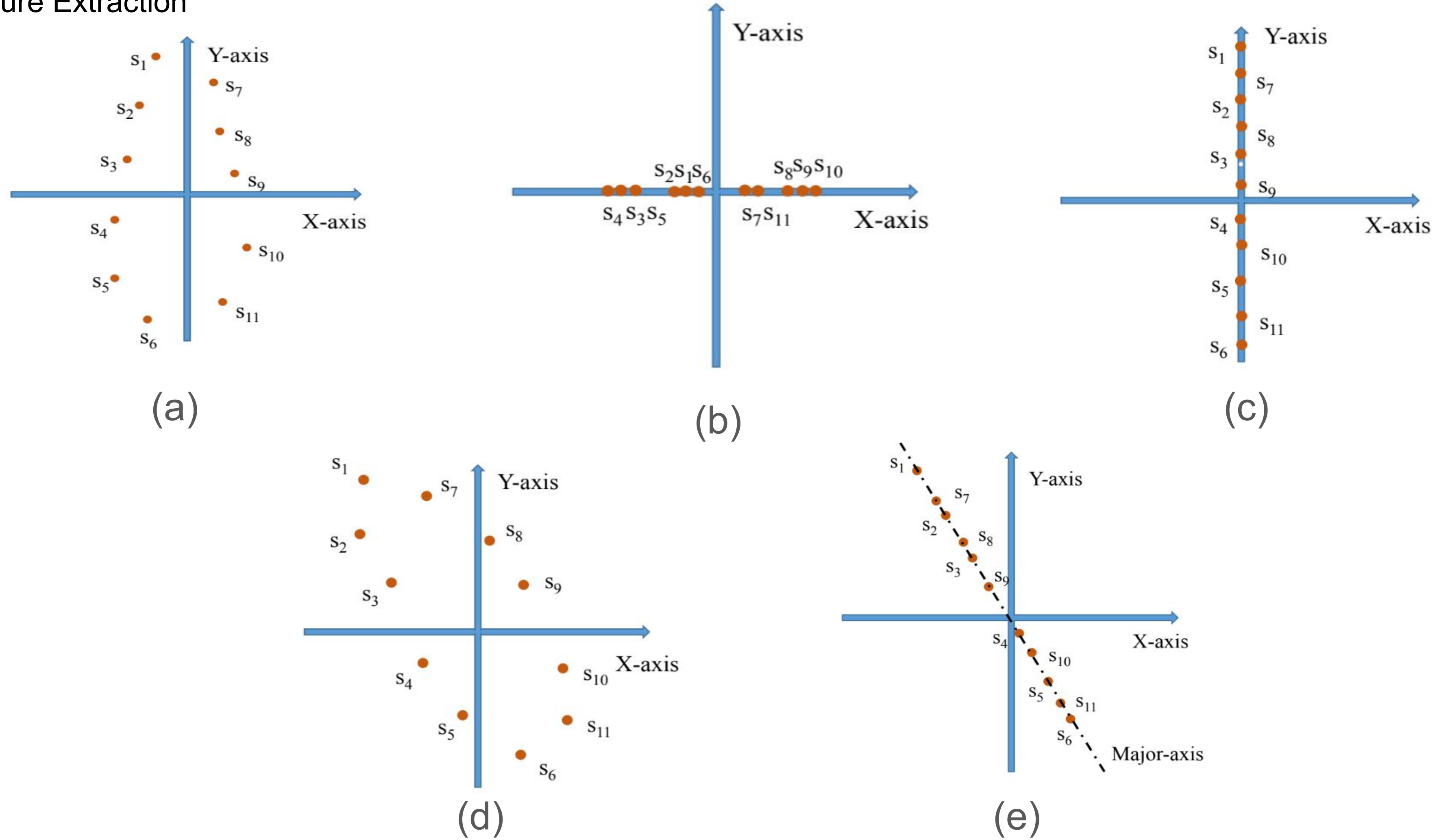
A_4?
Y
N
A_1?
A_6?
Y
N
Y
N
Class 1
Class 2
Class 1
Class 2

Source: *Data Mining: Concepts and Techniques”* by Jiawei Han, Jian Pei, Hanghang Tong , Morgan Kaufmann, 2022, ISBN: 9780128117613

Attribute Subset Selection



Feature Extraction



Feature Extraction - PCA

- The first principal component contains the maximum variance of the data, and then each successive component has the remaining variance with the subsequent maximum value.

Steps for PCA:

1. Standardize the data to zero mean and unit variance
2. Obtain the covariance matrix for the input features
3. Calculate Eigen value and Eigen vector for the covariance matrix
4. Sort the Eigen value and its corresponding Eigen vectors
5. Choose ‘d’ Eigen values and generate matrix of Eigen vectors
6. Project the input features to low dimensional space by multiplying with the matrix

Thank You !!!