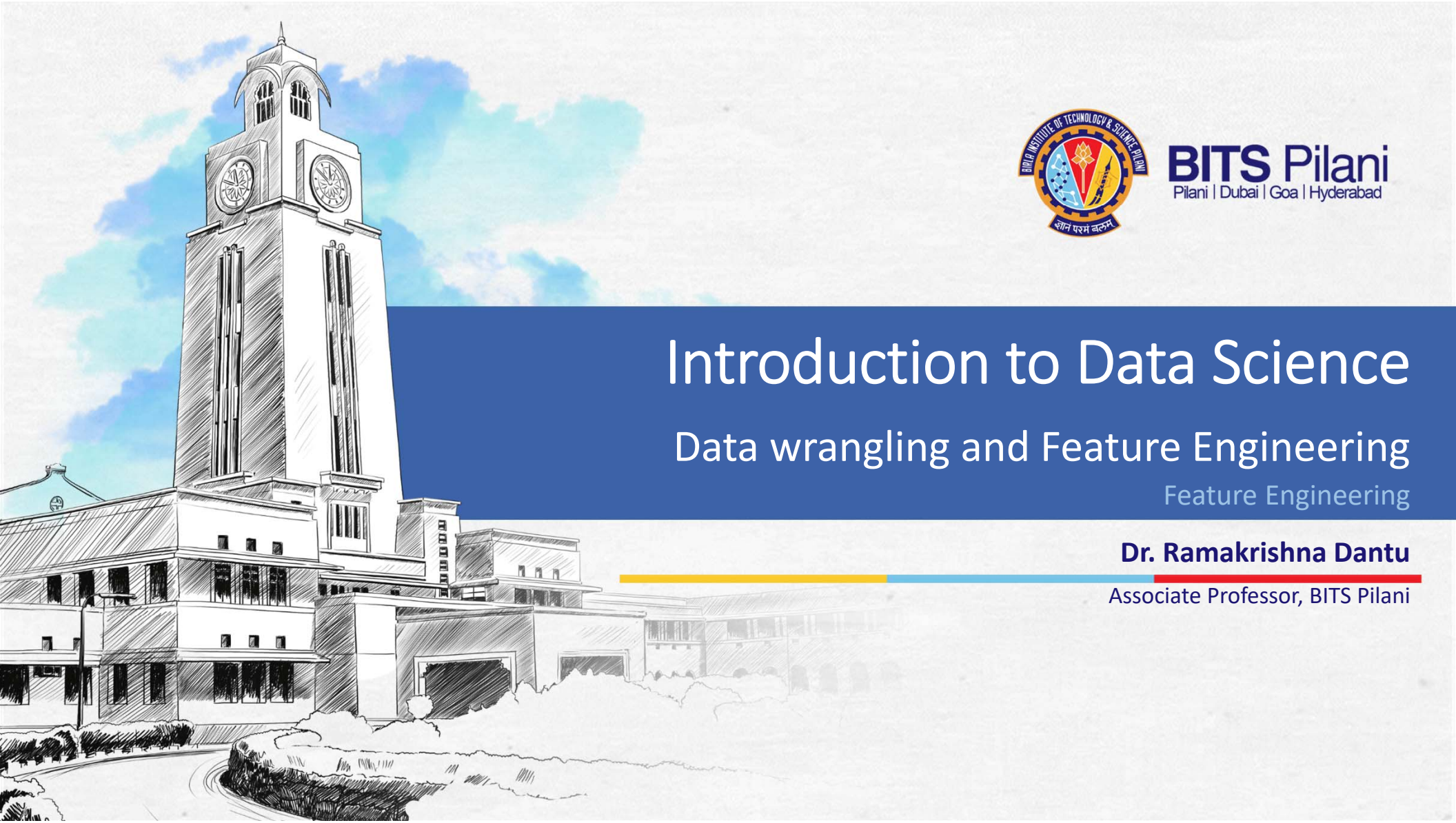


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Data wrangling and Feature Engineering

Feature Engineering

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

innovate achieve lead

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Introduction to Data Science

innovate achieve lead

Data wrangling and Feature Engineering – Part-1

- Data cleaning
- Data Aggregation, Sampling,
- Handling Numeric Data
  - Discretization, Binarization
  - Normalization
  - Data Smoothening
- Dealing with textual Data
- Managing Categorical Attributes
  - Transforming Categorical to Numerical Values
  - Encoding techniques

- Feature Engineering
  - Feature Extraction (Dimensionality Reduction)
  - Feature Construction
  - Feature Subset selection
    - Filter methods
    - Wrapper methods
    - Embedded methods
  - Feature Learning
- Case Study involving FE tasks

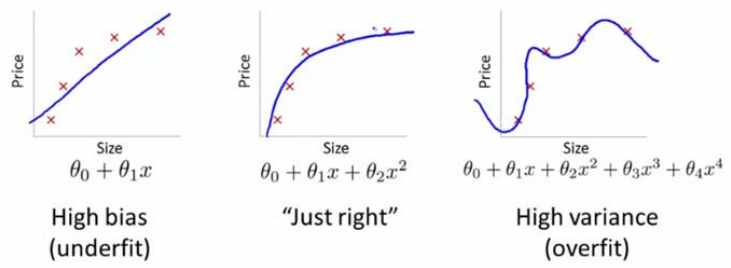
Feature Selection

Feature Selection

innovate achieve lead

Motivation

- Redundant and irrelevant features can reduce:
  - classification accuracy and
  - the quality of the clusters that are found
- Motivation for feature subset selection is:
  - To eliminate noisy features and in the process
    - Reduce the computation load, and
    - Improve the performance (and accuracy) of the model
  - To reduce the training time required to build model
  - To reduce overfitting



[Source: Coursera Andrew Ng Course on Machine Learning]

Feature Selection

Overview

- A feature selection procedure is also known as variable subset selection
- We may select a subset of features in the dataset based on:
  - the knowledge of the application domain
- Feature selection involves identifying relevant and important predictors from irrelevant or redundant variables
  - Basically, a dataset is reduced to having a few attributes that really matter
- Feature selection generally looks to test several subsets of features and find the subset that minimizes the error rate

Source: Big Data Analytics Methods by Peter Ghavami

Feature Selection

innovate achieve lead

Overview

- Given a set of N features $F = \{f_1, f_2, \ldots, f_N\}$ and target class label T
- Find, a minimum subset $F' = \{f_1', f_2', \ldots, f_K'\}$ that achieves maximum classification performance where $F' \subseteq F$ (K << N)
- Need a criteria to decide which subset is the best:
  - Classifier based on these K features has the *lowest classification error*
- However, selecting the best subset of features requires a systematic approach
- The ideal approach would be to try all possible subsets and select the one that produces best results.

Source: Introduction to Data Mining by Tan, Steinbach, & Kumar

Feature Selection

Search Methods

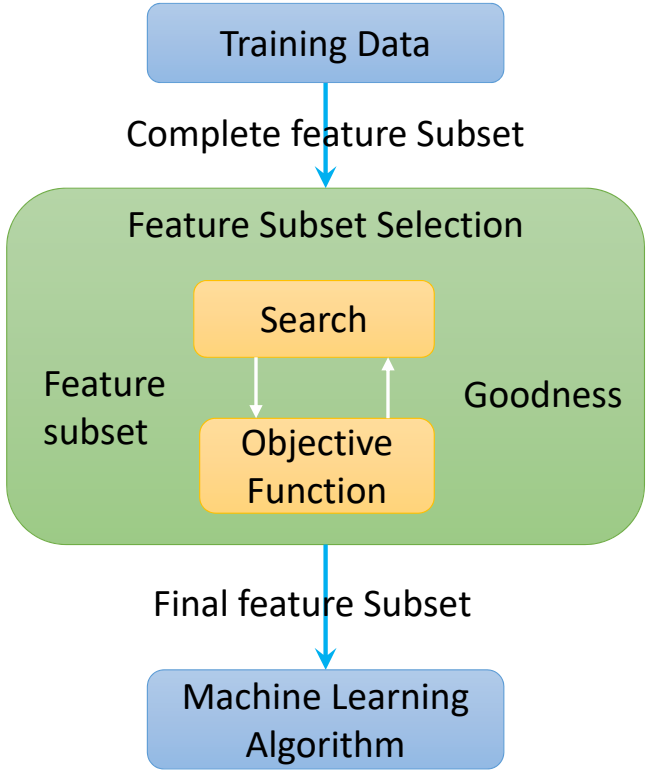
- For a given N initial set of features, there are 2^N-1 ($\approx 2^N$) possible subsets
- Evaluating 2^N-1 possible subsets is impractical because it is time consuming and expensive
  - e.g., evaluation of 10 out of 20 features involves 184,756 feature subsets.
  - e.g., evaluation of 10 out of 100 involves more than 10^{13} feature subsets.
- In practice, heuristics are used to speed-up search but they cannot guarantee optimality

Source: Introduction to Data Mining by Tan, Steinbach, & Kumar

Feature Selection

Feature Selection Main Steps

- Feature selection is an optimization problem having the following steps:
  - Step 1:
    - Search the space of possible feature subsets.
  - Step 2:
    - Pick the subset that is optimal or near-optimal with respect to some objective function
  - ABCDE ($2^5 - 1 = 31$)
    - A B C D E
    - AB AC AD AE BC BD BE CD CE DE
    - ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
    - ABCD ABCE ABDE ACDE BCDE
    - ABCDE



Feature Selection

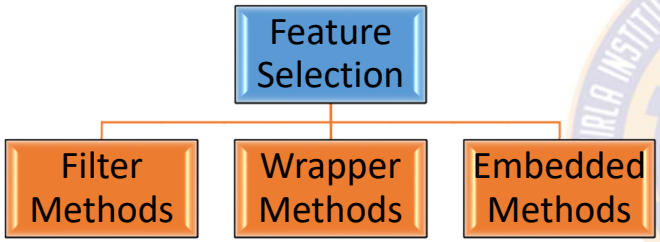
Feature Selection – Sample Dataset

- **id** : Unique Identifier for mobile device
- **battery_power** : Total energy a battery can store in one time measured in mAh
- **blue** : Has bluetooth or not
- **clock_speed** : speed at which microprocessor executes instructions
- **dual_sim** : Has dual sim support or not
- **fc** : Front Camera mega pixels
- **four_g** : Has 4G or not
- **int_memory** : Internal Memory in Gigabytes
- **m_dep** : Mobile Depth in cm
- **mobile_wt** : Weight of mobile phone
- **n_cores** : Number of cores of processor
- **pc** : Primary Camera mega pixels
- **px_height** : Pixel Resolution Height
- **px_width** : Pixel Resolution Width
- **ram** : Random Access Memory in Megabytes
- **sc_h** : Screen Height of mobile in cm
- **sc_w** : Screen Width of mobile in cm
- **talk_time** : longest time that a single battery charge will last when you are
- **three_g** : Has 3G or not
- **touch_screen** : Has touch screen or not
- **wifi** : Has wifi or not
- **price** : Actual market price of the device

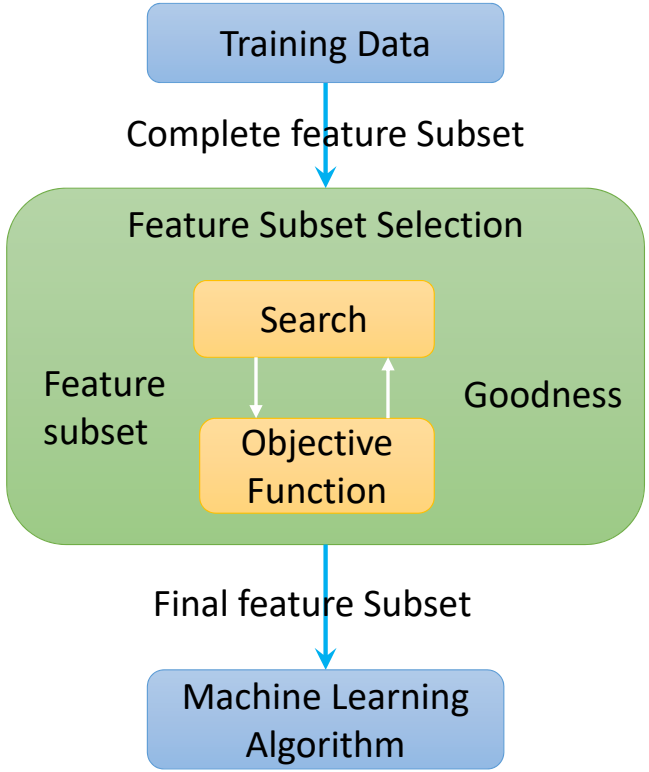
Feature Selection

Three Types of Feature Selection

- Three types of feature selection approaches to evaluate the possible subsets



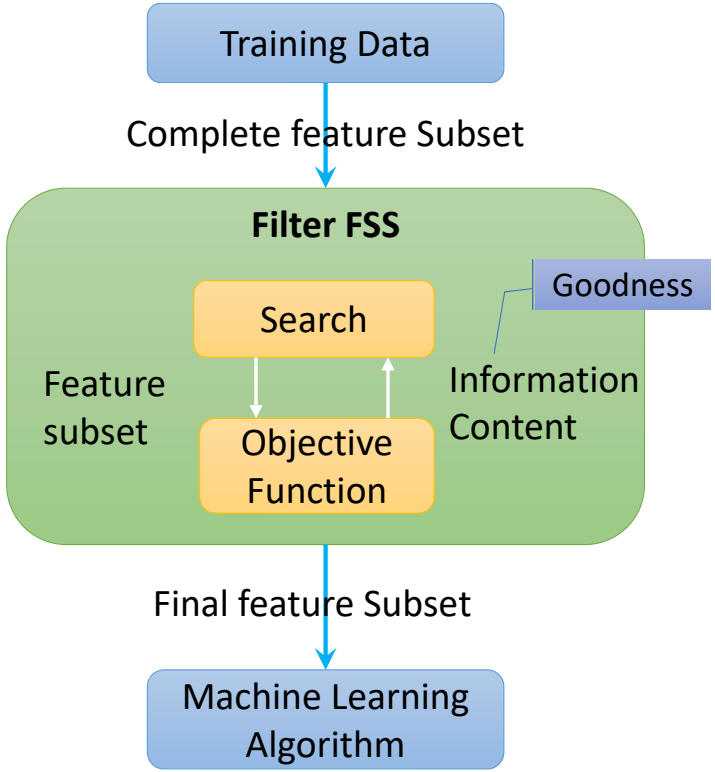
- Search methods
  - Exhaustive
  - Heuristic
  - Randomized



Three Types of Feature Selection

Unsupervised: Filter Method

- Generally used as a preprocessing step
- Feature selection is independent of the machine learning algorithm
- The objective function is based on the information content of the feature subsets, e.g.:
  - interclass distance
  - statistical dependence
  - information-theoretic measures (e.g., mutual information)
- Features are selected on the basis of their scores in various statistical tests or their correlation with the outcome variable

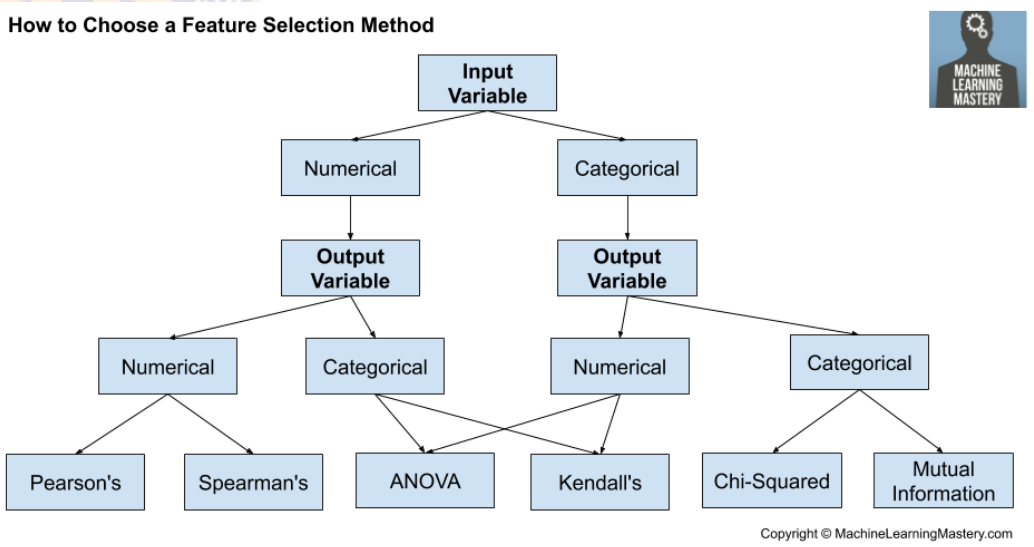


Three Types of Feature Selection

Unsupervised: Filter Method

- Table shows various statistical tests based on the type of input and output variables

Response → Feature↓	Continuous	Categorical
Continuous	Pearson's Correlation	Linear Discriminant Analysis
Categorical	Anova	Chi-Square



Three Types of Feature Selection

Unsupervised: Filter Method

- Information Content of Feature Subset
  - *Univariate filters* evaluate each feature independently with respect to the target variable
  - *Multivariate filters* evaluate features in context of others
- Correlation-based
  - Pearson product-moment correlation
  - Spearman rank correlation
  - Kendall concordance
- Statistical/probabilistic independence metrics
  - Chi-square statistic
  - F-statistic
  - Welch's statistic
- Information-theoretic metrics
  - Mutual Information (Information Gain)
  - Gain Ratio
- Others
  - Fisher score
  - Gini index
  - Cramer's V

Three Types of Feature Selection

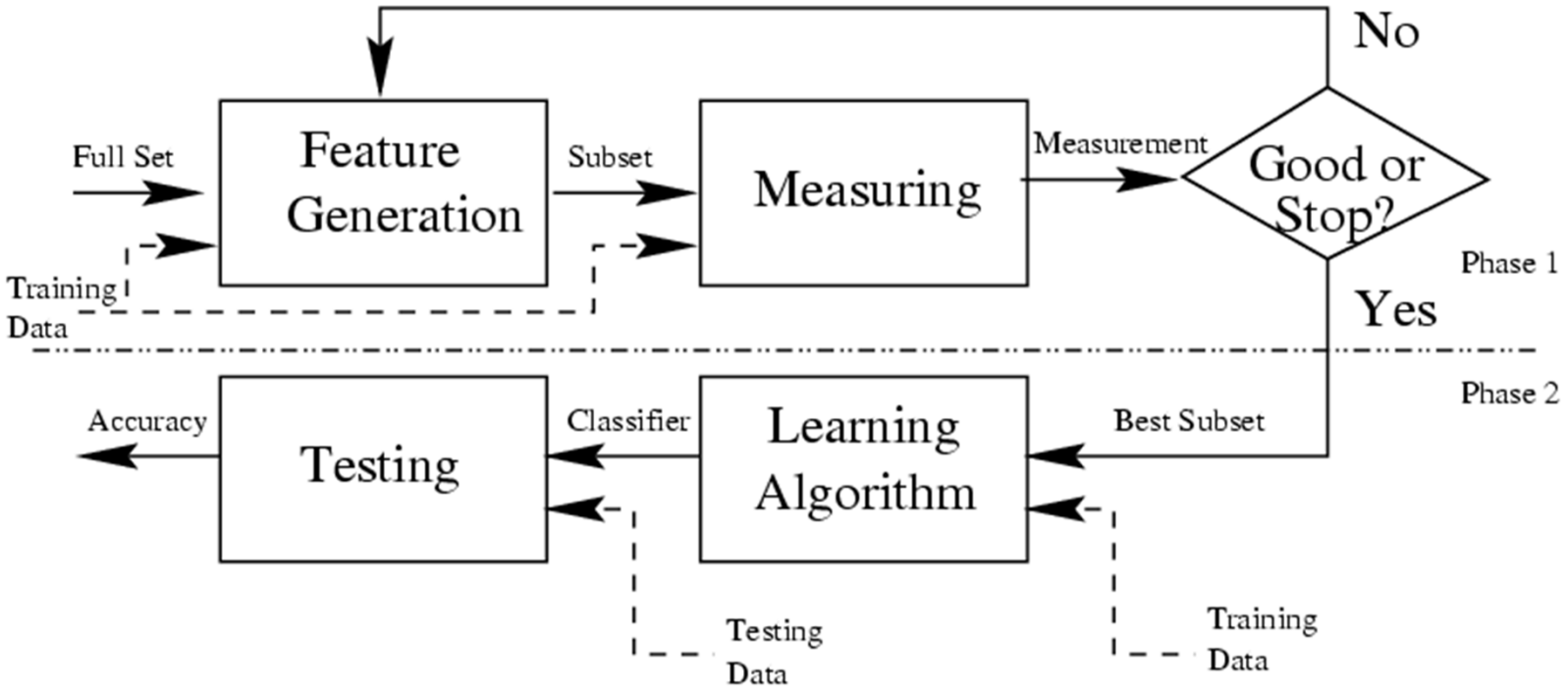
Unsupervised: Filter Method

- The scikit-learn library provides useful statistical measures:
- For example:
  - f_regression()
    - Pearson's Correlation Coefficient
  - f_classif()
    - ANOVA
  - chi2()
    - Chi-Squared
  - mutual_info_classif() and mutual_info_regression()
    - Mutual Information:
- The SciPy library provides implementations of other useful statistics, such as
  - kendaltau
    - Kendall's tau
  - spearmanr
    - Spearman's rank correlation

Three Types of Feature Selection

innovate achieve lead

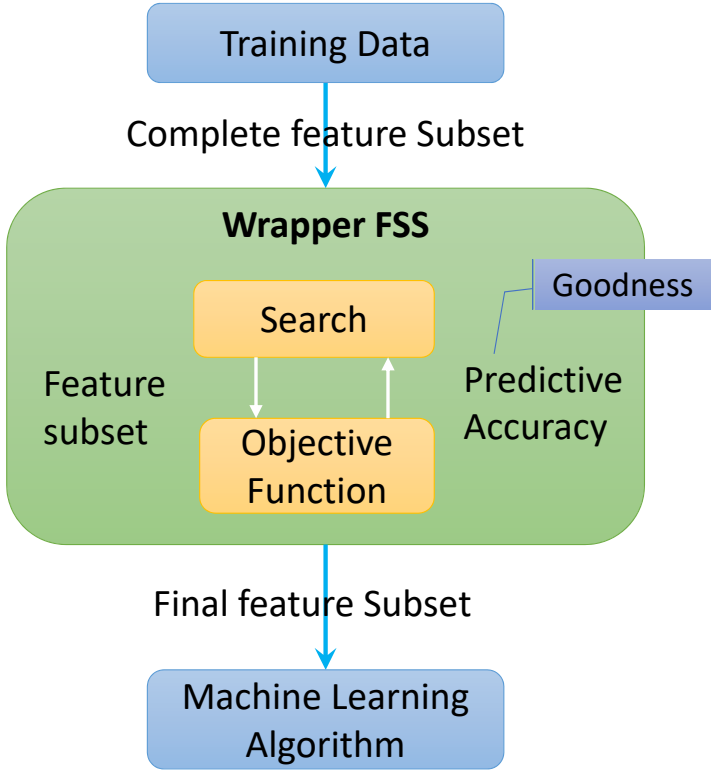
Unsupervised: Filter Method



Three Types of Feature Selection

Supervised: Wrapper Method

- Evaluates a subset of features using a specific machine learning algorithm
- It follows a search strategy by evaluating all the possible combinations of features against the evaluation criterion
- These methods are called greedy algorithms because they aim to find the best possible feature combination for the model
- The evaluation criterion depends on the type of the problem
- For example:
  - For regression, evaluation criterion can be p-values, adjusted R-square
  - For classification, evaluation criterion can be accuracy, precision, recall, f1-score, etc.
- The process selects a combination of features that gives the optimal results for the specified machine learning algorithm



Three Types of Feature Selection

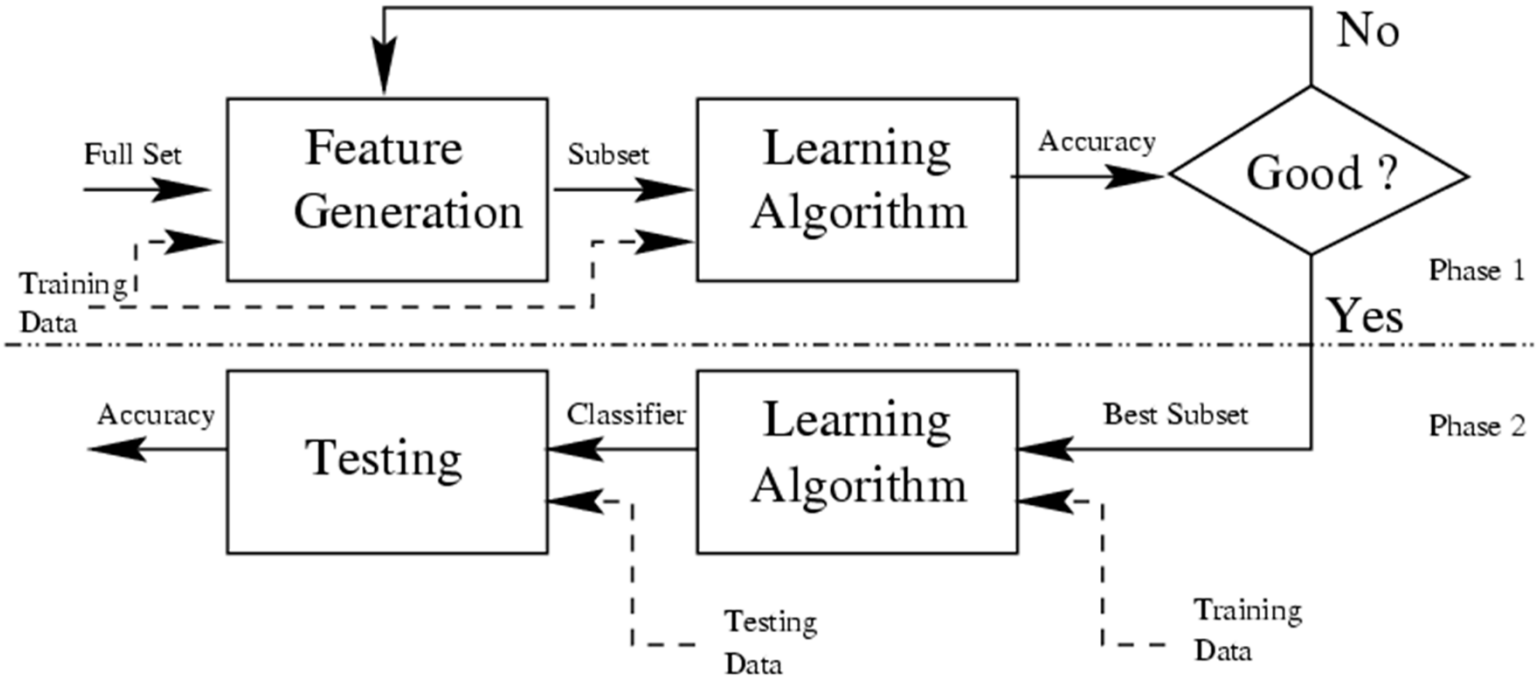
Supervised: Wrapper Method – General Process

- Search for a subset of features:
  - Using a search method (forward selection, backward elimination, or exhaustive selection), select a subset of features from the available ones
- Build a machine learning model
  - Train the chosen ML algorithm on the selected subset of features
- Evaluate model performance:
  - Evaluate the newly-trained ML model with a chosen metric
- Repeat:
  - Repeat the whole process with a new subset of features, a new ML model trained, and so on.
- Stop when the desired condition is met, and then choose the best subset with the best result in the evaluation phase

Three Types of Feature Selection

innovate achieve lead

Supervised: Wrapper Method



Three Types of Feature Selection

Supervised: Wrapper method

- Algorithm

Input: large feature set Ω

Identify candidate subset $S \subseteq \Omega$

While !stop_criterion()

Evaluate error of a classifier using S.

Adapt subset S.

Return S

- Stopping Criteria
  - Increase/Decrease in model performance (predictive accuracy)
  - A predefined number of features is reached

Three Types of Feature Selection

Supervised: Wrapper method

- Search Methods
  - Forward Feature Selection:
    - Consider a *significance level* (Say, 0.05) and *confidence level* (say, 95%)
    - Start with no features to begin with and iteratively add each feature that improves the performance of the model
      - Select the feature with minimum *p-value*
    - Now, fit a model with two features at a time (combinations of earlier selected feature with the remaining features)
      - Select the feature combination with minimum p-value
    - Now, fit a model with three features at a time (combinations of two earlier selected features with the remaining features)
      - Select the feature combination with minimum p-value
    - Repeat the process until we get a feature subset with a p-value < significance level

Three Types of Feature Selection

Supervised: Wrapper method

- Search Methods
  - Backward Feature Elimination:
    - Consider a *significance level* (Say, 0.05) and *confidence level* (say, 95%)
    - Create a full model by including all the features
    - Remove a feature that is least significant *(p-value > significance level)*
    - Build a model with the remaining features
    - Remove a feature that is least significant *(p-value > significance level)*
    - Repeat the process until there is no more improvement in the performance

Three Types of Feature Selection

Supervised: Wrapper method

- Search Methods
  - Bi-directional Elimination
    - Consider a *significance level* (Say, 0.05) and confidence level (say, 95%)
    - Uses a combination of both forward selection and backward elimination methods
    - While adding a new feature, checks the significance of previously added features
      - If the previously added feature is insignificant (p-value > 0.05) then, remove the particular feature through backward elimination
    - Newly added feature must have p-value < 0.05 to be considered
    - Repeat the process until all features are exhausted

Three Types of Feature Selection

Supervised: Wrapper method

- Search Methods
  - Exhaustive Feature Selection:
    - Uses a brute-force evaluation of feature subsets
    - This method tries all possible feature combinations.
    - Repeatedly creates models with different feature subsets and keeps aside the best or the worst performing feature at each iteration
    - ABCDE
      - A B C D E
      - AB AC AD AE BC BD BE CD CE DE
      - ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
      - ABCD ABCE ABDE ACDE BCDE
      - ABCDE

Three Types of Feature Selection

Embedded Method

- Embedded methods combine the qualities of filter and wrapper methods
- Selects features that provide a non-zero regression coefficient (meaning the regression line is sloped)
- They are implemented by algorithms that have their own built-in feature selection methods
- Most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting
- Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.
- Ridge regression performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

Three Types of Feature Selection

Embedded Method

Filter Methods	Wrapper Methods	Embedded Methods
Generic set of methods Do not incorporate any particular ML algorithm	Evaluates features based on a specific ML algorithm	Embeds features during the model building process Feature selection is done by observing each iteration of model training phase
Faster compared to Wrapper methods in terms of time complexity	Higher computation time for datasets with larger number of features	Sits between filter and wrapper methods in terms of time complexity
Less prone to overfitting	High possibility of overfitting because it involves training of ML models with different combination of features	Reduces overfitting by penalizing the coefficients of a model being too large
Examples: Correlation, Chi-Sqr, ANOVA, Information Gain, etc.	Examples: Forward Selection, Backward Selection, Embedded, etc.	Examples: LASSO, Elastic Net, Ridge Regression, etc.


innovate
achieve
lead

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Thank You!