DLI Accelerated Data Science Teaching Kit

Lecture 3.4 - Feature Selection: Introduction to Filter Methods

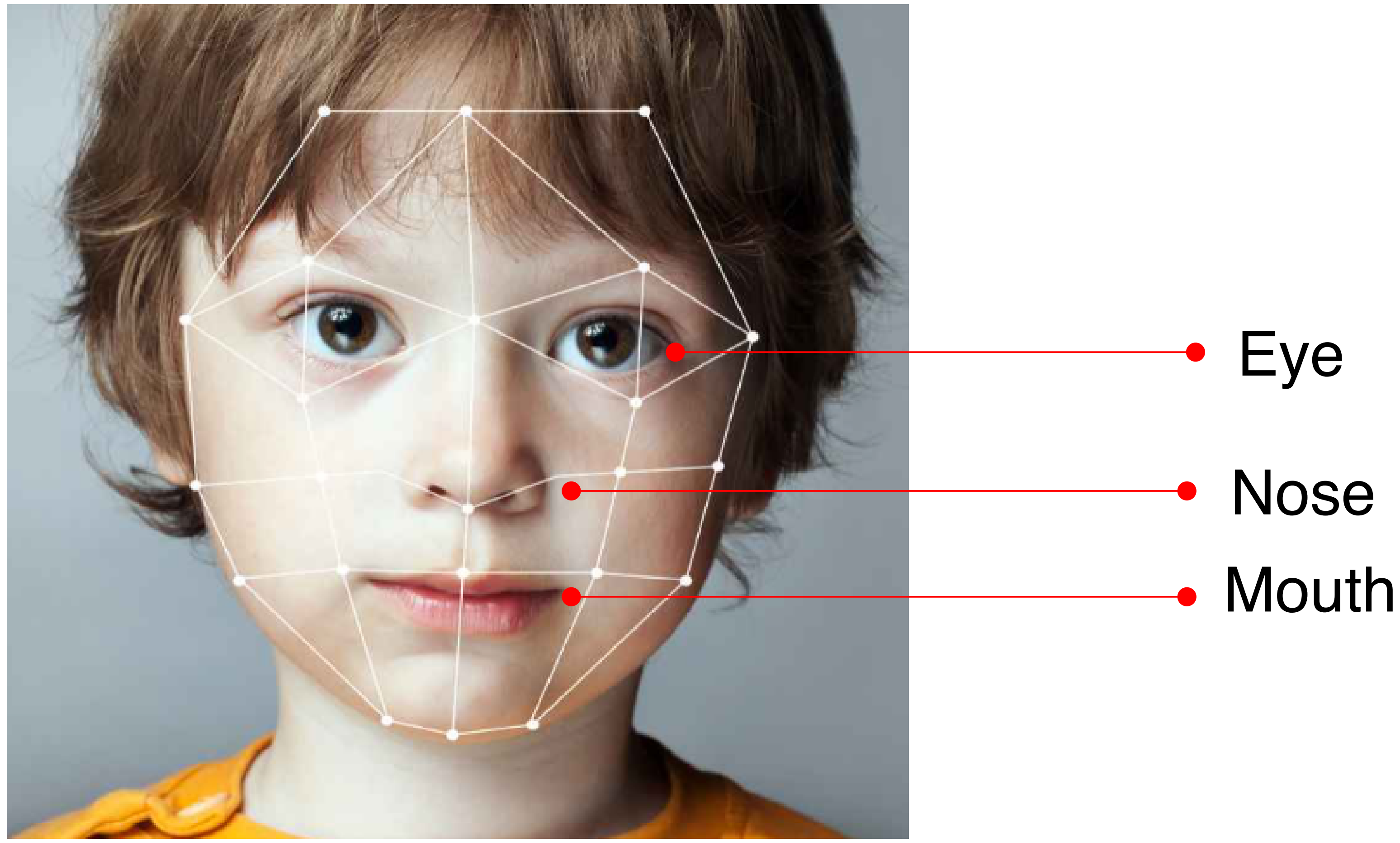
The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the
Creative Commons Attribution-NonCommercial 4.0 International License.

Feature

Feature is a distinctive attribute or aspect of something.

In machine learning based data analytics, a feature is an individual measurable property or characteristic of a phenomenon being observed.

- **Face Recognition**
  - **Nose**
  - **Mouth**
  - **Eye**
  - **Ear**
  - **… …**



Source: https://www.nyclu.org/en/news/facial-recognition-cameras-do-not-belong-schools

Feature Extraction

Feature extraction is to extract features from raw data or even create new features on the raw data.

- **Creating new features** is a process to generate new variables/features based on existing variables/features.

New Features

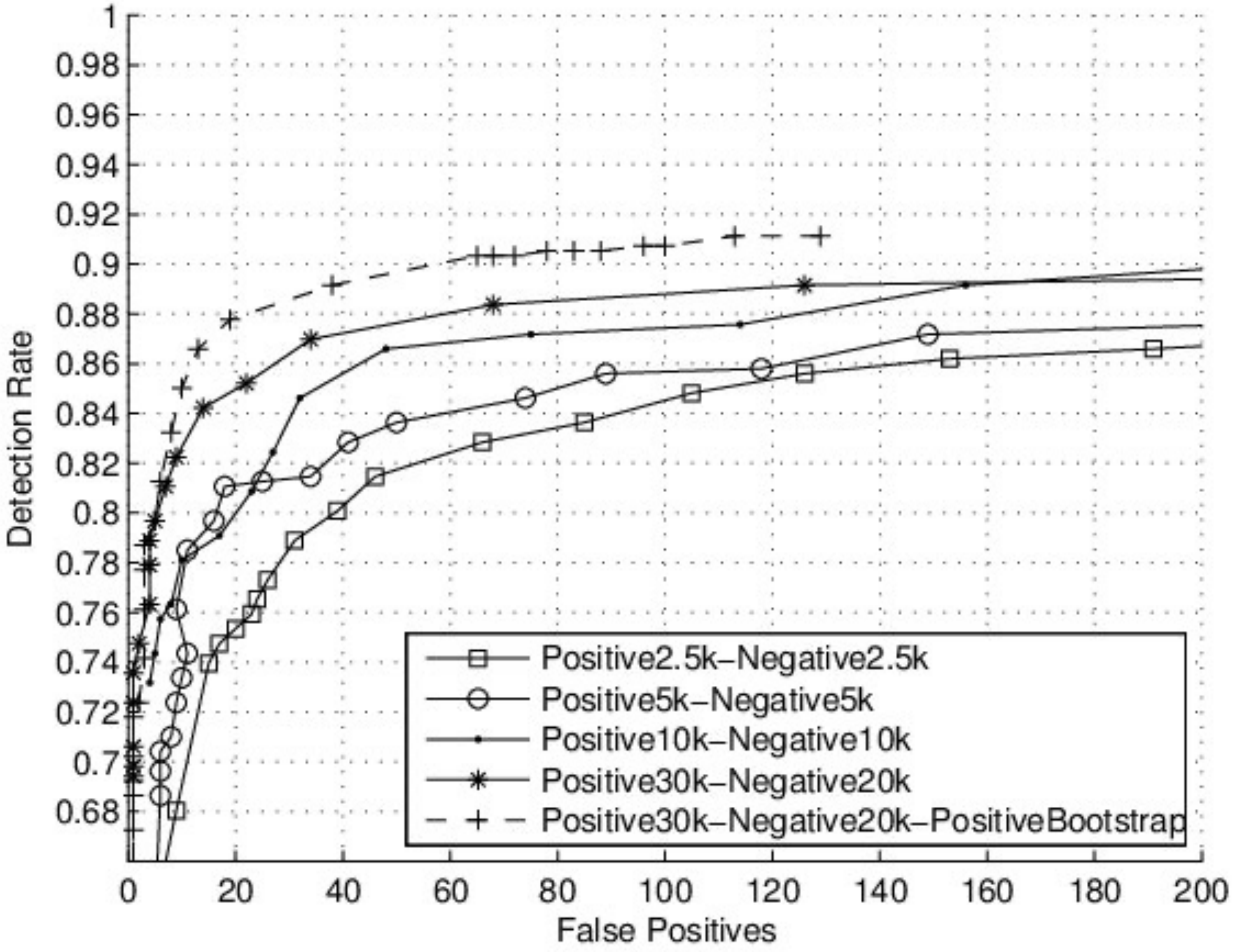
#	First Name	Last Name	Date	New Day	New Month
1	John	Boo	1/12/2020	12	1
2	Mary	Brown	11/24/2019	24	11
3	James	Mooray	5/13/2019	13	5

Feature Extraction

Different sets of features will lead to different performance of data analytics.

- Face Detection

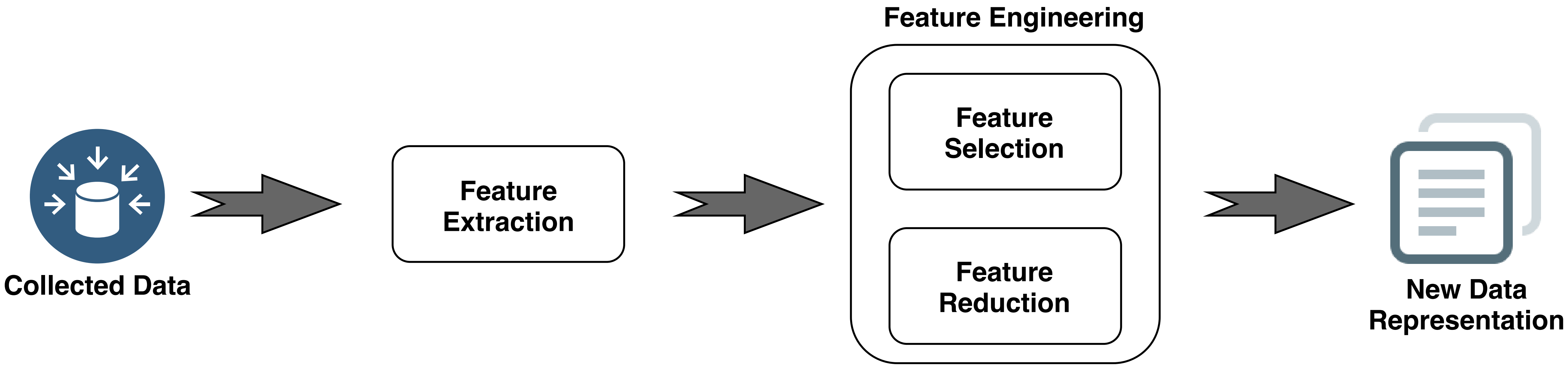
- Performance comparison of detectors using multiple feature sets on the CMU+MIT frontal face data set.



Source: Xiao, R., Zhu, H., Sun, H., & Tang, X. (2007, October). Dynamic cascades for face detection. In 2007 IEEE 11th International Conference on Computer Vision (pp. 1-8).

Feature Engineering

Feature engineering is a process that is to represent the features by feature selection or feature reduction for data analytics.

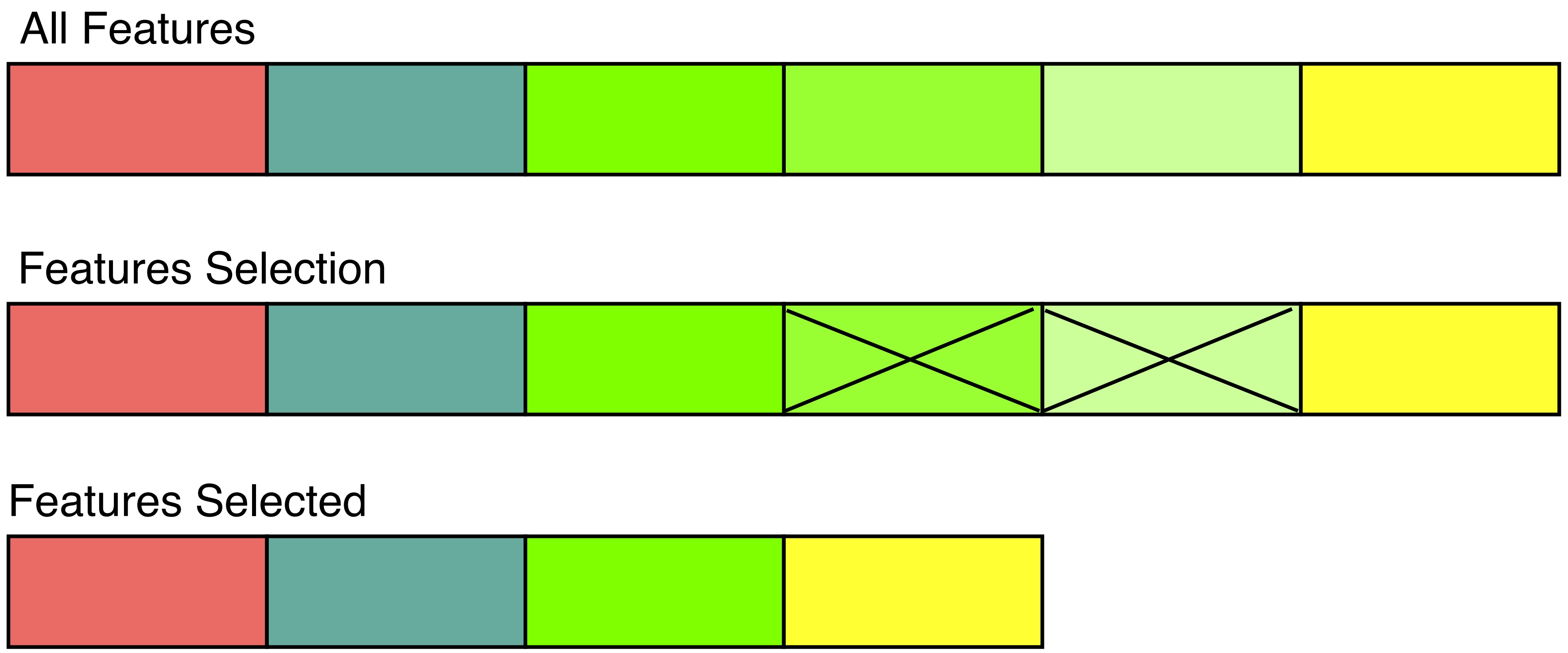


- Requiring collaborations between domain experts and data scientists
- A trade-off between including more informative features and avoiding too many unrelated features
    - Informative features improve model performance.
    - Unrelated features reduce model performance.

Feature Selection

Feature selection is able to remove features that are either redundant or irrelevant without loss of information of data.

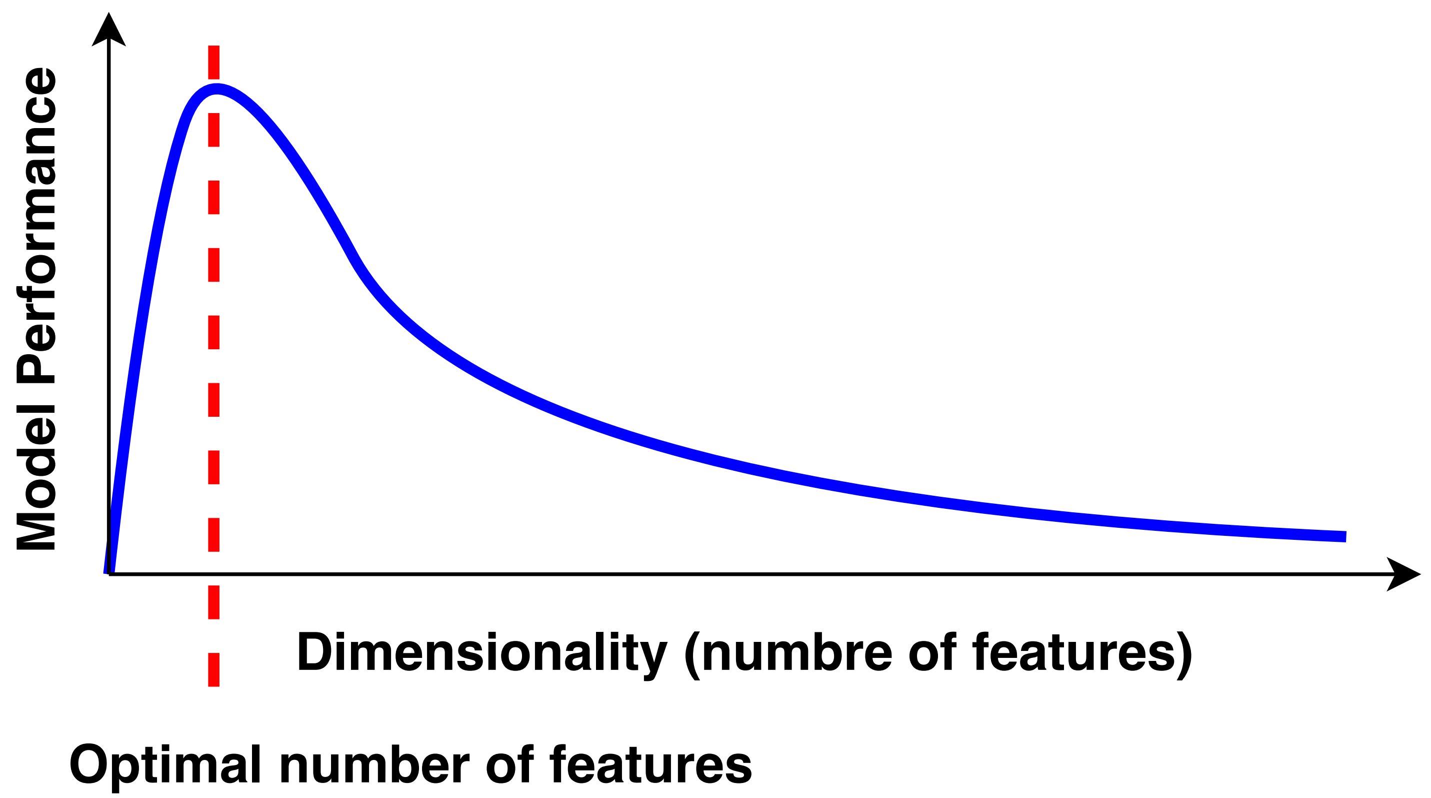
- Simplifying the models
- Shortening time for model construction
- Avoiding curse of dimensionality
- Enhancing model generalization



Curse of Dimensionality

It refers to phenomena that arise when analyzing and organizing data in high-dimensional spaces.

- Certain models might perform poorly in high-dimensional space.
- More features might need more samples to fill out the high-dimensional space.
- Samples are difficult to be obtained.

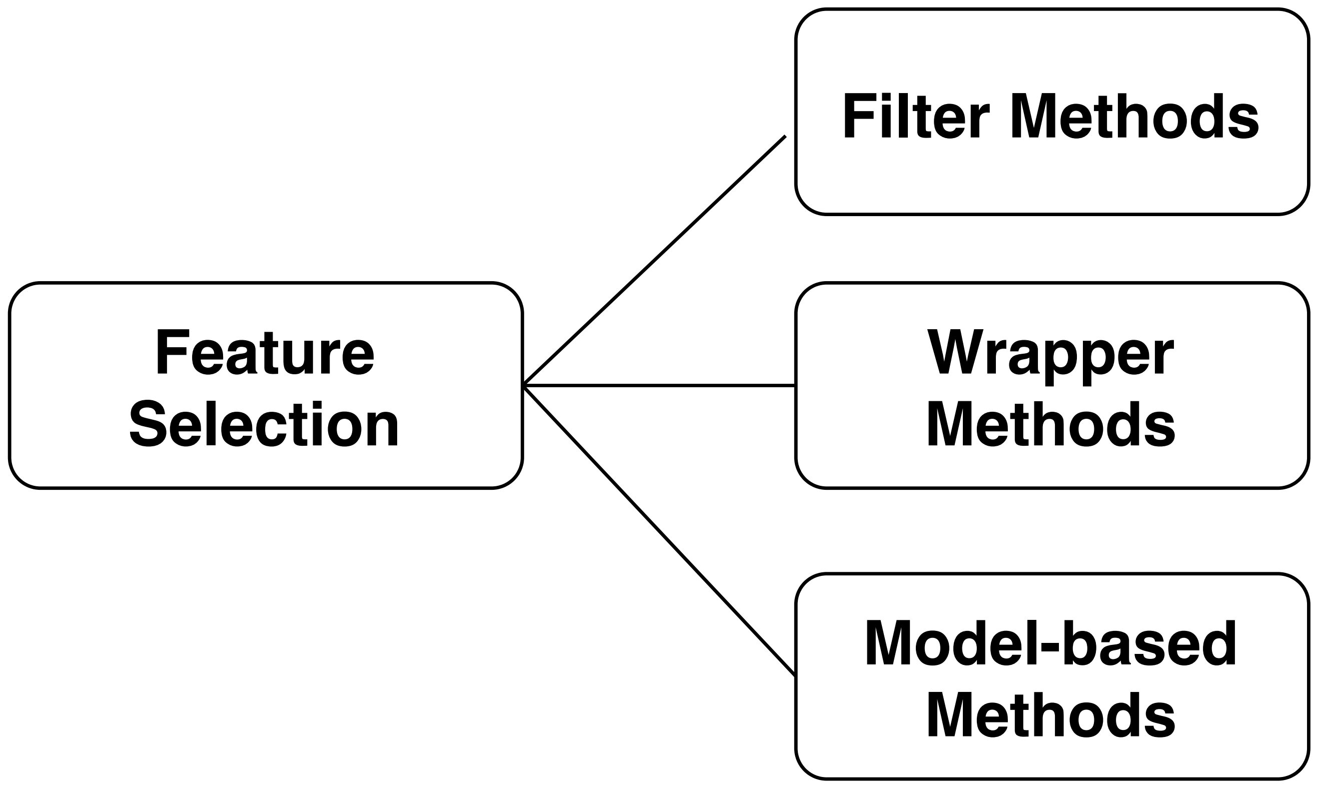


NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

Approached for Feature Selection

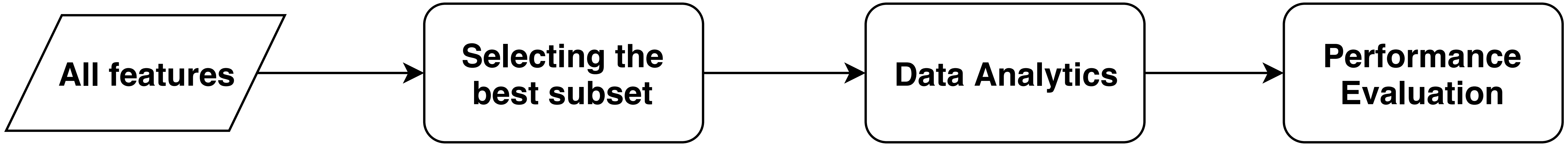
There are three categories of feature selection methods.

- **Filter methods** select features regardless of the model.
- **Wrapper methods** select features based on performance of a specific model with a greedy search in a forward/backward manner.
- **Model-based methods** select features during the procedure of model construction.



Filter Methods

Generally, filter methods are to calculate the correlations between the features and target attributes.



- Statistical measurements
  - Chi squared test
  - Mutual information
  - Pearson correlation
  - … …

Chi Squared Test

Mathematically, a **Chi-Square test** is done on two distributions to determine the level of similarity of their respective variances.

In its null hypothesis, it assumes that the given distributions are independent.

This test can be used to determine the best features for a given dataset by determining the features on which the target attribute is most dependent on.

For each feature in the dataset, the χ^2 is calculated and then ordered in descending order according to the χ^2 value.

Chi Squared Test

The higher the value of χ^2, the more dependent the output label is on the feature and higher the importance the feature has on determining the output.

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij}: Observed frequency

E_{ij}: Expected frequency

Chi Squared Test

Predicting **Play Tennis**

- Two features
  - Outlook
  - Wind
- Target Attribute
  - Play Tennis

Day	Outlook	Wind	Play Tennis
D1	Sunny	Weak	No
D2	Sunny	Strong	No
D3	Overcast	Weak	Yes
D4	Rain	Weak	Yes
D5	Rain	Weak	Yes
D6	Rain	Strong	No
D7	Overcast	Strong	Yes
D8	Sunny	Weak	No
D9	Sunny	Weak	Yes
D10	Rain	Weak	Yes
D11	Sunny	Strong	Yes
D12	Overcast	Strong	Yes
D13	Overcast	Weak	Yes
D14	Rain	Strong	No

Chi Squared Test

The contingency table for the feature "Outlook" is constructed as below.

	Yes	No	
Sunny	2 (3.21)	3 (1.79)	5
Overcast	4 (2.57)	0 (1.43)	4
Rain	3 (3.21)	2 (1.79)	5
	9	5	14

- The expected value for the cell (Sunny, Yes) is calculated as $\frac{5}{14}\times 9 = 3.21$ and similarly for others.

$$\chi^2_{outlook} = \frac{(2-3.21)^2}{3.21} + \frac{(3-1.79)^2}{1.79} + \frac{(4-2.57)^2}{2.57} + \frac{(0-1.43)^2}{1.43} + \frac{(3-3.21)^2}{3.21} + \frac{(2-1.79)^2}{1.79}$$

$$\Rightarrow \chi^2_{outlook} = 3.129$$

Chi Squared Test

The contingency table for the feature “Wind” is constructed as below.

	Yes	No	
Strong	3 (3.86)	3 (1.14)	6
Weak	6 (5.14)	2 (2.86)	8
	9	5	14

$$\chi^2_{wind} = \frac{(3-3.86)^2}{3.86} + \frac{(3-1.14)^2}{1.14} + \frac{(6-5.14)^2}{5.14} + \frac{(2-2.86)^2}{2.86}$$

$$\Rightarrow \chi^2_{wind} = 3.629$$

- On comparing the two scores, we can conclude that the feature “Wind” is more important to determine the output than the feature “Outlook”.

NVIDIA. DEEP LEARNING INSTITUTE Georgia Tech PRAIRIE VIEW A&M UNIVERSITY

Mutual Information

In information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables.

- **Entropy**

$$\boldsymbol{H(X) = -\sum_{x \in \mathcal{X}} p(x) log(p(x))}$$

- **Joint Entropy**

$$\boldsymbol{H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) log(p(x,y))}$$

- **Mutual Information**

$$\boldsymbol{I(X;Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) log(\frac{p(x,y)}{p(x)p(y)})}$$

NVIDIA. DEEP LEARNING INSTITUTE Georgia Tech PRAIRIE VIEW A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You