# Introduction to the basics of AI

## Session 5

Z. TAIA-ALAOUI

# Outline

- Definitions
- Linear Discriminant Analysis
- Filter-based feature selection (Fisher Score and Mutual Information)
- Wrapper-based Feature Selection

# Statistical Tools - Dataset

- **Sample**

$$X_{k \in [1,N]} \in R^p$$

- **Set of samples**

$$\mathbf{X} = \{X_k \in \mathrm{R}^p\}_{k \in [1,\mathrm{N}]}$$

**IRIS DATASET**

$p$ variables

$N$ samples

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | [illegible] | [illegible] | [illegible] | [illegible] |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 |

Row = Sample = Observation = Measurement

# Statistical Tools - Dataset

- **Set of N samples expressed in a space of p Variables**

$$\mathbf{X} = \begin{array}{c c} & \begin{array}{ccccc} \phantom{x_{11}} & \phantom{x_{12}} & \phantom{x_{13}} & V_j & \phantom{x_{1p}} \end{array} \\ \begin{array}{c} \\ \\ \\ X_i \\ \\ \\ \end{array} & \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N1} & . & . & . & x_{Np} \end{pmatrix} \end{array} = [V_1, V_2, \ldots, V_p] = \begin{bmatrix} X_1^{\mathrm{T}} \\ X_2^{\mathrm{T}} \\ . \\ . \\ . \\ X_N^{\mathrm{T}} \end{bmatrix}$$

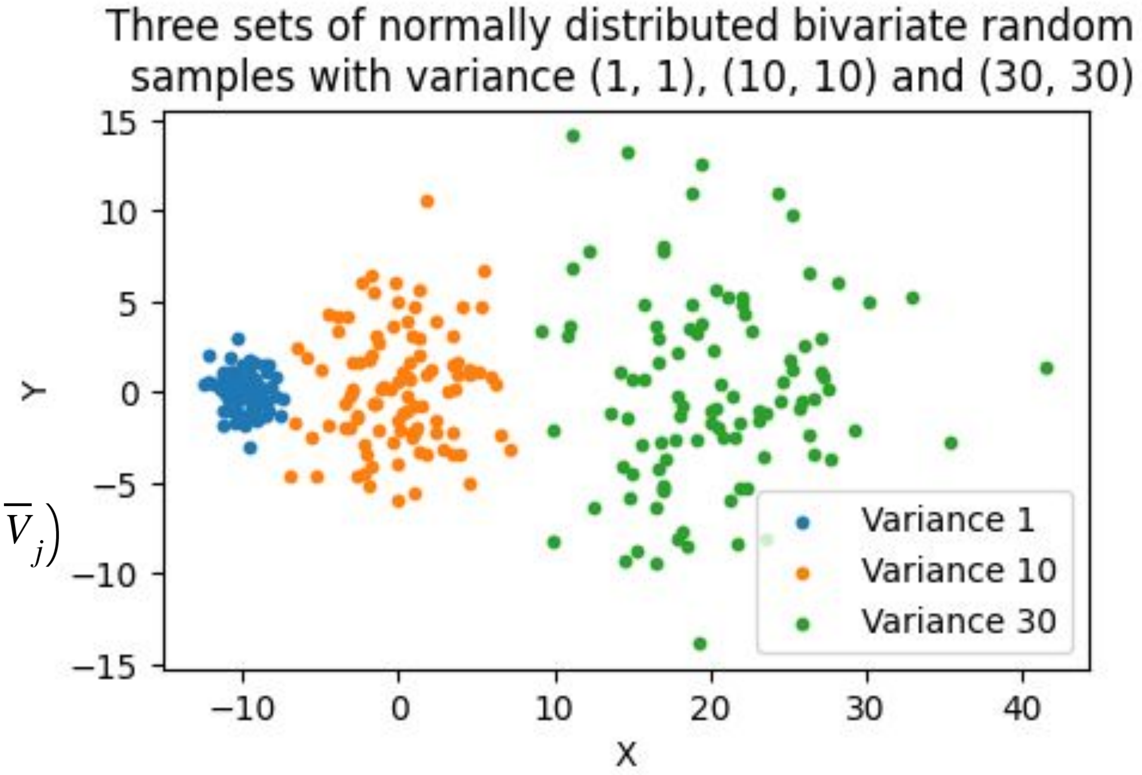
$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip} \end{bmatrix} \qquad V_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ . \\ x_{Nj} \end{bmatrix}$$

# Statistical Tools - Variance

$$Var(V_j) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \overline{V_j}\right)^2$$

$$\overline{\mathrm{V}}_{\mathrm{j}} = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$$

$$Cov(V_i, V_j) = \frac{1}{N-1}\sum_{k=1}^{N}\left(x_{ki} - \overline{V}_i\right)\left(x_{kj} - \overline{V}_j\right)$$



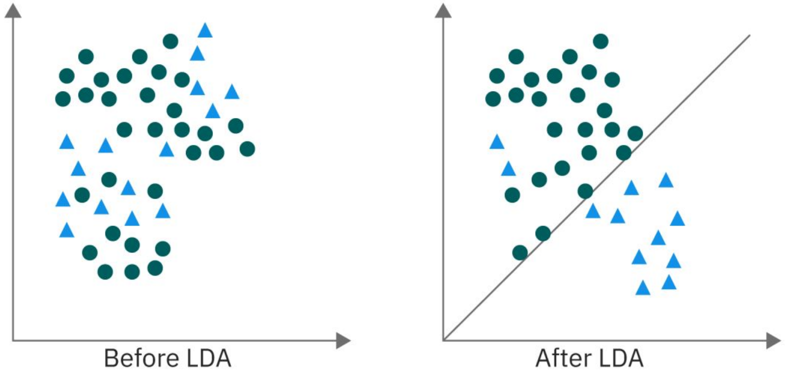
# Linear Discriminant Analysis

- LDA is an equivalent for PCA that is well suited for **supervised** classification problems.

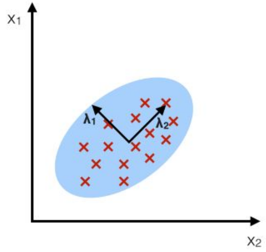
- Unlike ANOVA, LDA has continuous independent variable (measurements) and categorical dependent variables (class labels).

- LDA is a feature **extraction** method that makes linear combination of input features in order to optimize class separability.

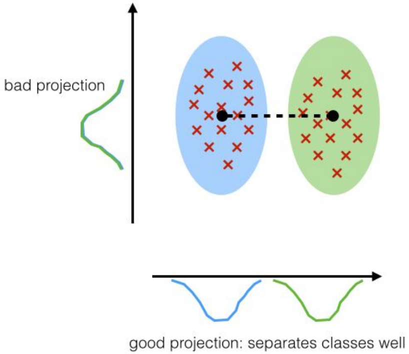
# Linear Discriminant Analysis




**PCA:**
component axes that maximize the variance




**LDA:**
maximizing the component axes for class-separation



# Linear Discriminant Analysis

- LDA projects data into a new space in which between-class separation is maximized.
- Separation means maximizing the distance between the projected means and minimizing the projected variance within classes. (Fisher method again !).
- Assumptions:
    - Normal Distribution
    - Covariance Homogeneity

# Linear Discriminant Analysis

1. Compute the $d$-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$) and corresponding eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose $k$ eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\boldsymbol{W}$ (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$ (where $\boldsymbol{X}$ is a $n \times d$-dimensional matrix representing the $n$ samples, and $\boldsymbol{y}$ are the transformed $n \times k$-dimensional samples in the new subspace).

# Linear Discriminant Analysis

- Within class scatter matrix

$$S_W = \sum_{i=1}^{c} S_i \qquad S_i = \sum_{\boldsymbol{x} \in D_i}^{n} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

- Between class scatter matrix

$$S_B = \sum_{i=1}^{c} N_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

# Linear Discriminant Analysis

- Eigenvalue and Eigenvectors computation

$$\boldsymbol{A} = S_W^{-1} S_B$$
$$\boldsymbol{v} = \text{ Eigenvector}$$
$$\lambda = \text{ Eigenvalue}$$

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

- Final Projection of data

$$\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$$

# Feature Selection Based on score thresholds

- **Fisher Score**

$$F(\mathbf{x}^j) = \frac{\sum_{k=1}^{c} n_k(\mu_k^j - \mu^j)^2}{(\sigma^j)^2}$$

$\mu^j, \sigma^j$ are mean and variance of j-th feature

$\mu_k^j$ is the mean of the j-th feature for group k

# Feature Selection Based on score thresholds

- **Mutual Information**

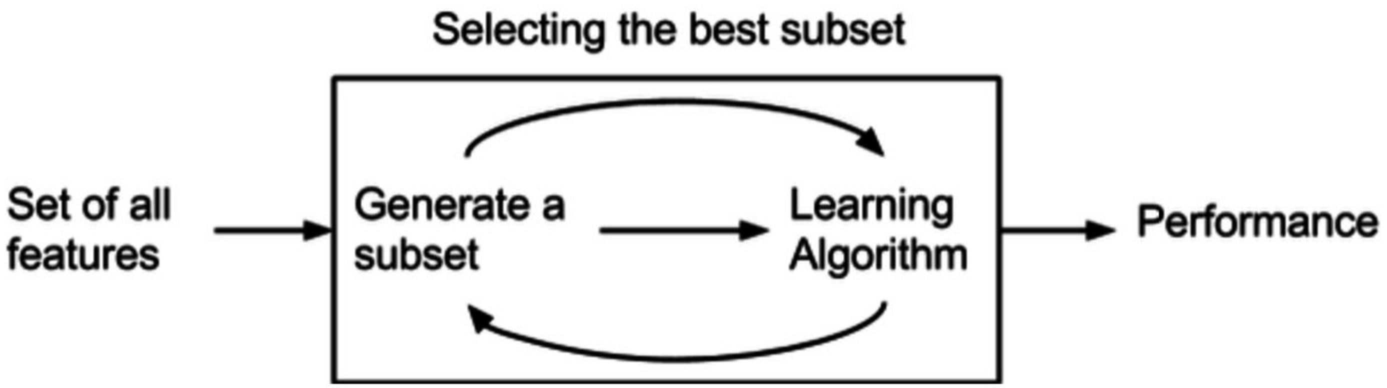
  - The Mutual Information is a measure of the similarity between two labels of the same data, so the input feature must be first categorized.

  - Qualitatively, entropy is a measure of uncertainty – the higher the entropy, the more uncertain one is about a random variable.

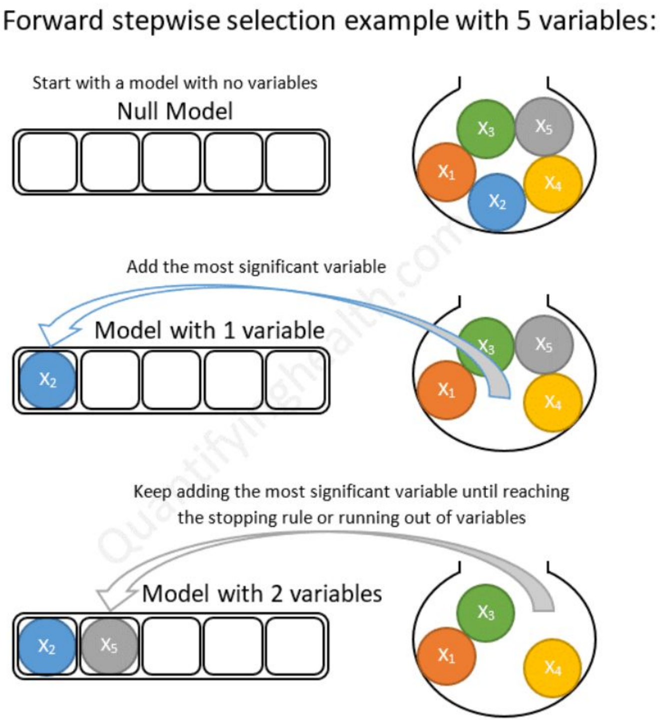
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

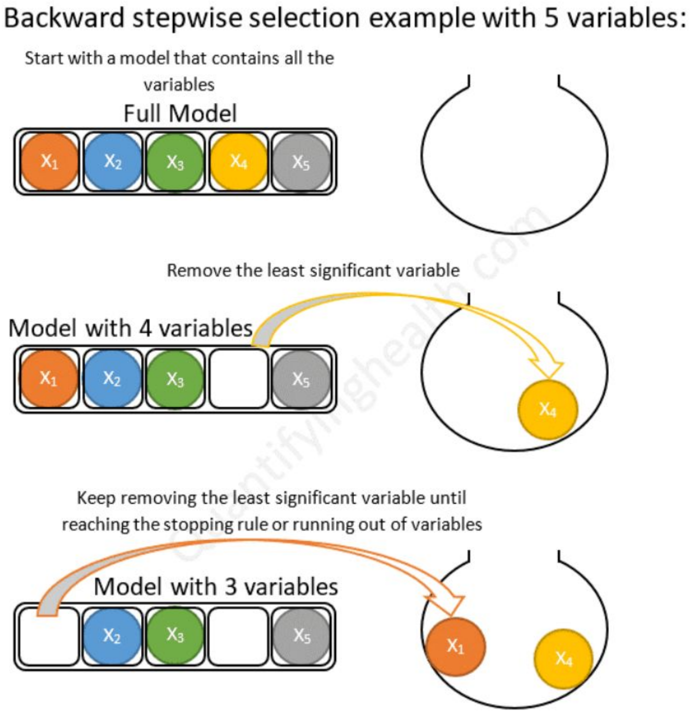
$$H(X) = -\sum_i p(x_i)\log_2 p(x_i)$$

# Wrapper Methods:

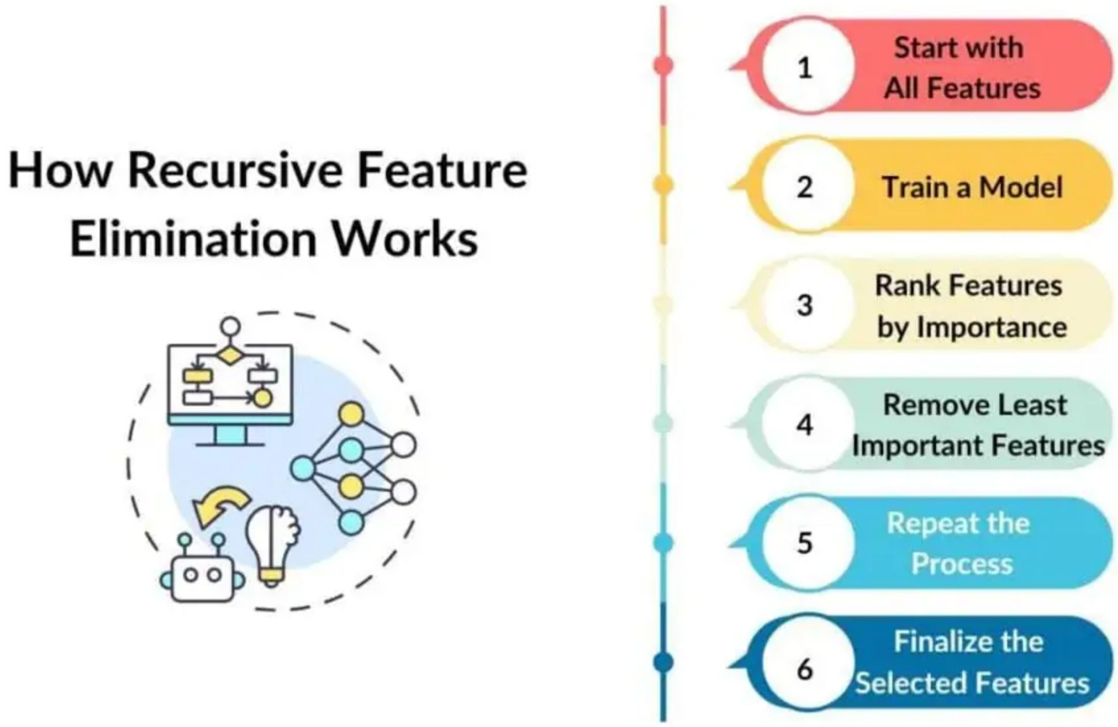


# Wrapper Methods: **Forward Selection, Backward Selection, Exhaustive Selection**

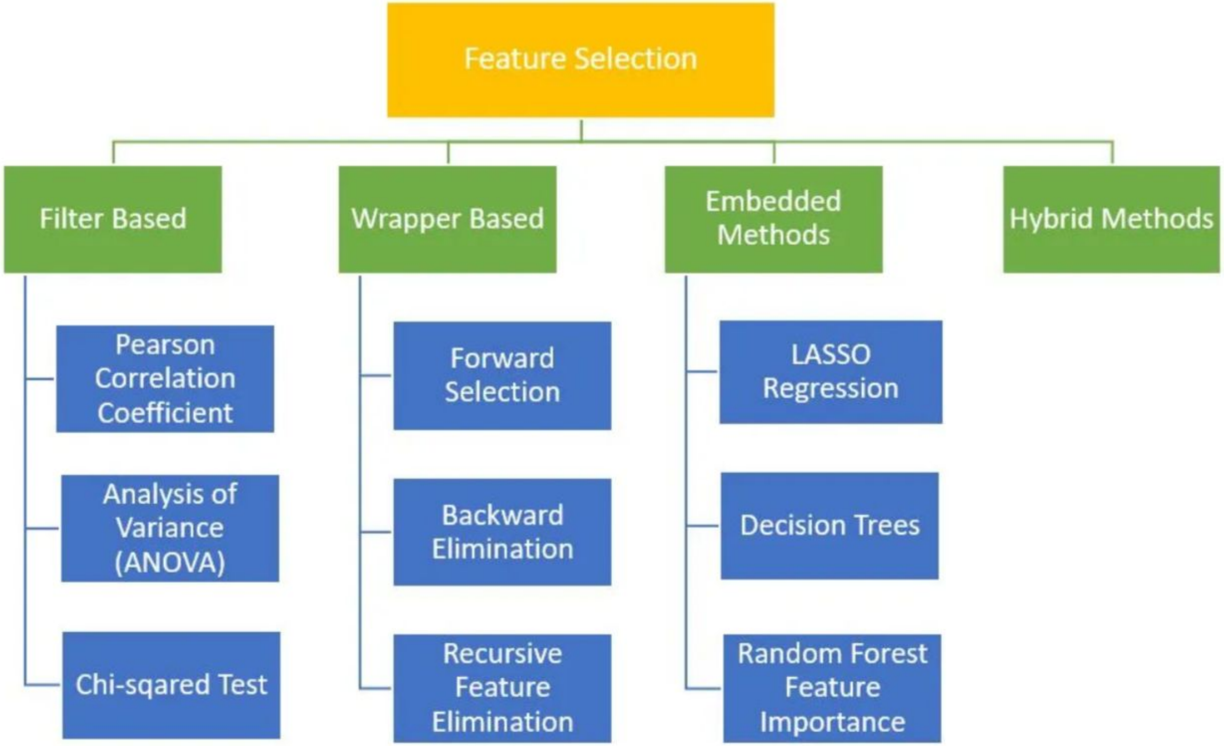


# Wrapper Methods: Recursive Feature Elimination



# Summary


Feature Selection
Filter Based
Wrapper Based
Embedded Methods
Hybrid Methods
Pearson Correlation Coefficient
Analysis of Variance (ANOVA)
Chi-sqared Test
Forward Selection
Backward Elimination
Recursive Feature Elimination
LASSO Regression
Decision Trees
Random Forest Feature Importance

# Resources

https://www.ibm.com/topics/linear-discriminant-analysis

https://sebastianraschka.com/Articles/2014_python_lda.html#normality-assumptions

https://spotintelligence.com/2024/11/18/recursive-feature-elimination-rfe/

https://www.kdnuggets.com/2018/06/step-forward-feature-selection-python.html

https://dsdojo.medium.com/wrapper-based-feature-selection-techniques-dd3ff6d1b79f

https://www.stratascratch.com/blog/feature-selection-techniques-in-machine-learning/