Feature Learning

Finding the diamonds in the rough

High Dimensionality

Higher Dimensionality

More Data

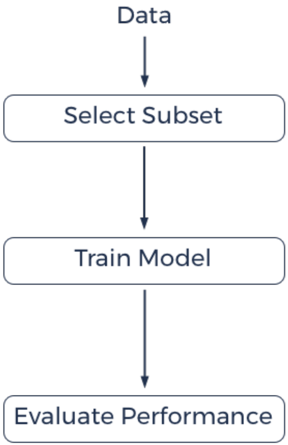
Feature 1	Feature 2	Feature 3	Feature 4
Value 1, 1	Value 1, 2	Value 1, 3	Value 1, 4
Value 2, 1	Value 2, 2	Value 2, 3	Value 2, 4
Value 3, 1	Value 3, 2	Value 3, 3	Value 3, 4
Value 4, 1	Value 4, 2	Value 4, 3	Value 4, 4
Value 5, 1	Value 5, 2	Value 5, 3	Value 5, 4
Value 6, 1	Value 6, 2	Value 6, 3	Value 6, 4

Feature Learning

- Feature Selection
- Feature Extraction

Feature Selection

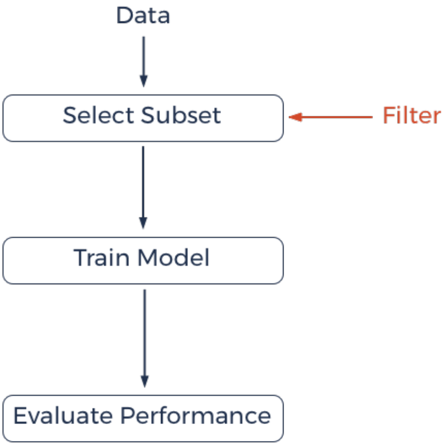
Feature Selection



Feature Selection Methods

- Filter methods
- Wrapper methods
- Embedded methods

Filter Methods

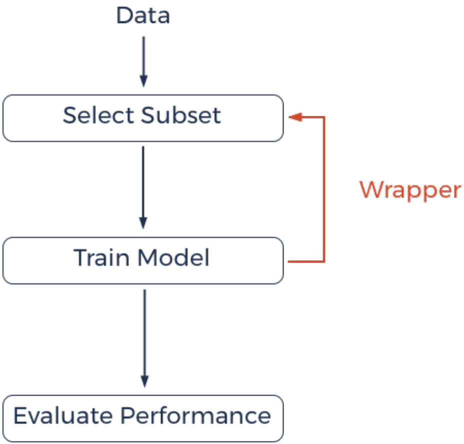


Mutual Information Score

Mutual information between two vectors of discrete values, **x** and **y** is,

$$MI(\mathbf{x},\mathbf{y}) = \sum_{x_i \in \mathbf{x}} \sum_{y_j \in \mathbf{y}} p(x_i, y_i) \log \left(\frac{p(x_i, y_j)}{p(x_i) \times p(y_j)} \right).$$

Wrapper Methods

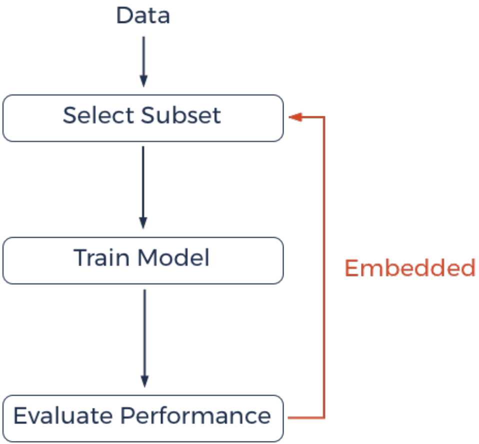


Recursive Feature Elimination

Algorithm 1: Recursive Feature Elimination

Input: Features from data
Parameter: r features to select from the data
Parameter: k features to remove per iteration
Output: Subset of r features from the data
Train a model on all features and obtain feature coefficients
while *selected features count* $> r$ **do**
    Remove up to k features with the lowest feature coefficients
    Train a model on the new subset of features

Embedded Methods



Regularization Based methods

$$\min_{\mathbf{w}} f(\mathbf{x}, \mathbf{w}) + g(\mathbf{w}),$$

where f is a core objective function with learned parameters $\mathbf{w}$ and g is a regularization function applied to the learned parameters.

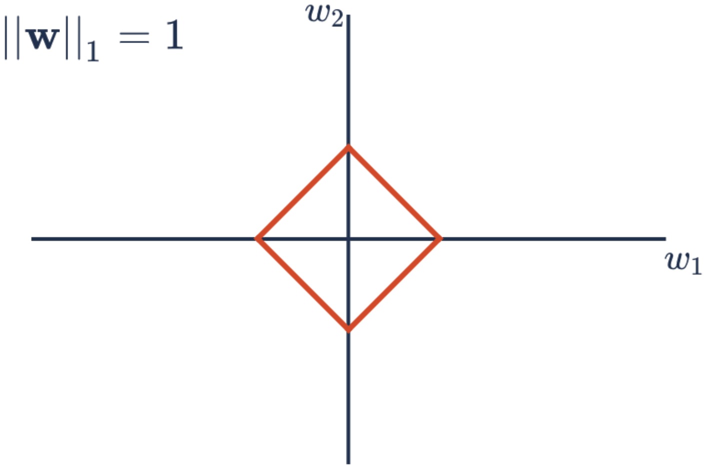
Linear Regression Based Regularization

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2 + g(\mathbf{w}).$$

Lasso

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2 + \alpha||\mathbf{w}||_1.$$

ℓ_1-norm



Feature Selection Summary

- Filter methods
  - **Pros:** Low computation time
  - **Cons:** May select redundant data, not as effective as other methods, greedy
- Wrapper methods
  - **Pros:** Incorporate information from learned model
  - **Cons:** Potentially high computation time, prone to overfitting, greedy
- Embedded methods
  - **Pros:** Improve on both Filter and Wrapper in terms of performance
  - **Cons:** High computation time

Feature Extraction

PCA

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2,$$
$$s.t.\ \mathbf{U}^T\mathbf{U} = \mathbf{I}.$$

Demo

Questions