Lecture 13: Feature selection

Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2022-2023

Motivation

Question

Do you remember what is the definition of feature reduction ?

Motivation

Question

Do you remember what is the definition of feature reduction ?

- Find the subset of variables which bears the most information

Motivation

Question

Do you remember what is the definition of feature reduction ?

- Find the subset of variables which bears the most information
- Use only these variables to perform the learning task

Motivation

Question

Why do we want to perform feature reduction ?

Motivation

Question

Why do we want to perform feature reduction ?

- Simplification of models to make them easier to interpret by users

Motivation

Question

Why do we want to perform feature reduction ?

- Simplification of models to make them easier to interpret by users
- Shorter training times

Motivation

Question

Why do we want to perform feature reduction ?

- Simplification of models to make them easier to interpret by users
- Shorter training times
- Avoid the curse of dimensionality

Motivation

Question

Why do we want to perform feature reduction ?

- Simplification of models to make them easier to interpret by users
- Shorter training times
- Avoid the curse of dimensionality
- Drop variables that do not respect models hypothesis

Motivation

Features can be either:

- **Redundant**: the information is already given by other features (highly correlated) (for example, late arrival and late departure).

Motivation

Features can be either:

- **Redundant**: the information is already given by other features (highly correlated) (for example, late arrival and late departure).
- **Unrelevant**: the feature does not bring any information (for example, the weather for image recognition)

Feature selection

Feature selection

Feature selection consists in finding the subset of variables which bear the most information for the learning task and only using those for problem solving.

Feature selection

Feature selection

Feature selection consists in finding the subset of variables which bear the most information for the learning task and only using those for problem solving.

Feature selection can be done using:

Feature selection

Feature selection

Feature selection consists in finding the subset of variables which bear the most information for the learning task and only using those for problem solving.

Feature selection can be done using:

- Field specific insight

Feature selection

Feature selection

Feature selection consists in finding the subset of variables which bear the most information for the learning task and only using those for problem solving.

Feature selection can be done using:

- Field specific insight
- Automatic rules based on objective metrics

Feature selection

Feature selection

Feature selection consists in finding the subset of variables which bear the most information for the learning task and only using those for problem solving.

Feature selection can be done using:

- Field specific insight
- Automatic rules based on objective metrics
- Embedded within the algorithm itself (can be used as part of the data processing step)

Possible approaches

Approaches can be:

Possible approaches

Approaches can be:

- **Filter based**: ignore the learning model and select a subset of variables.

Possible approaches

Approaches can be:

- **Filter based**: ignore the learning model and select a subset of variables.
- **Wrapper based**: iteratively build a model with a subset of feature and select the best performing one according to a learning score.

Variance analysis

Variance threshold selection

Variance based selection consists in **removing features with low variance**.

Variance analysis

Variance threshold selection

Variance based selection consists in **removing features with low variance**.

In practice:

- Set a minimum variance threshold (careful about bias !)

Variance analysis

Variance threshold selection

Variance based selection consists in **removing features with low variance**.

In practice:

- Set a minimum variance threshold (careful about bias !)
- Remove all features not matching this criterion.

Variance analysis

Variance threshold selection

Variance based selection consists in **removing features with low variance**.

In practice:

- Set a minimum variance threshold (careful about bias !)
- Remove all features not matching this criterion.

Question

Is this a filter-based or wrapper based approach ?

Advantages and limits

Advantages:

- Model independant
- Limitted risk of overfitting

Advantages and limits

Advantages:

- Model independant
- Limitted risk of overfitting

Limits:

- Does not take into account correlation between variables

Wrapper based selection

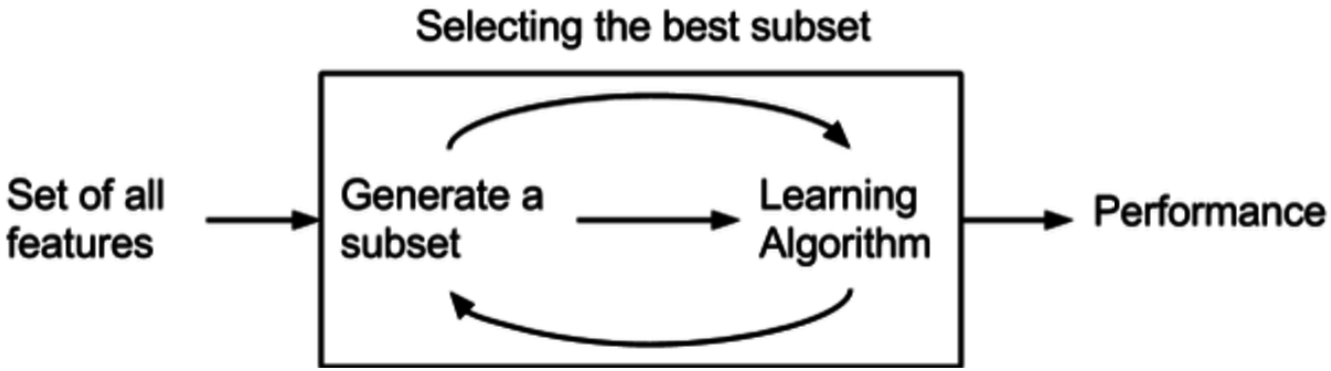
Wrapper based selection

Wrapper based selection (or **model-based selection**) consists in iteratively building a model on a subset of features and selecting the best performing subset of features.

Wrapper based selection

Wrapper based selection

Wrapper based selection (or **model-based selection**) consists in iteratively building a model on a subset of features and selecting the best performing subset of features.



Wrapper-based selection

Selection can be:

- **Forward**: Start without any feature and add at each round the one which maximizes the score

Wrapper-based selection

Selection can be:

- **Forward**: Start without any feature and add at each round the one which maximizes the score
- **Backward**: Start with every set of feature and remove at each round the one which maximizes the score.

Wrapper-based selection

Selection can be:

- **Forward**: Start without any feature and add at each round the one which maximizes the score
- **Backward**: Start with every set of feature and remove at each round the one which maximizes the score.

. . . and stop algorithm once a threshold is reached.

Advantages and limits

Advantages:

Advantages and limits

Advantages:

- Takes into account the specificity of the learning task

Advantages and limits

Advantages:

- Takes into account the specificity of the learning task
- Takes into account correlation between variables

Advantages and limits

Advantages:

- Takes into account the specificity of the learning task
- Takes into account correlation between variables

Limits:

Advantages and limits

Advantages:

- Takes into account the specificity of the learning task
- Takes into account correlation between variables

Limits:

- Model dependent

Advantages and limits

Advantages:

- Takes into account the specificity of the learning task
- Takes into account correlation between variables

Limits:

- Model dependent
- Correlation depends on subset

Questions

Questions ?