Machine Learning

FIB, Master in Data Science

Marta Arias, Computer Science @ UPC

On feature selection, multi-objective optimization and some key aspects for understanding Uche's PhD thesis

Outline

1. On Feature Selection
2. On Multi-objective Feature Selection
3. On Uche's project

Feature Selection

Feature Selection

Feature selection in machine learning refers to the process of identifying and selecting a subset of relevant features (variables, attributes) from the original set of features in a dataset.

The goals are to:

Sometimes when there are too many features, overfitting occurs

1. Avoid **curse of dimensionality**
2. Eliminate irrelevant/noisy input features
3. Improve training times by reducing the size of the input
4. Improve performance by reducing overfitting (we obtain simpler models with fewer inputs)
5. Improve interpretability of models

In summary, by selecting the right features, we can simplify models, reduce computational costs, and improve generalization to new data.

Main methods for feature selection

As part of preprocessing step,
ex: find correlation between target and each one of the input

1. **Filter** methods: evaluate the relevance of features based on statistical properties without considering a learning model.

2. **Wrapper** methods: use the predictive performance of a specific machine learning model to evaluate subsets of features.
   sophisticated
   use machine learning model, to see which features will be working
3. **Embedded** methods: incorporate feature selection as part of the model building process; some examples are:
   - LASSO (Least Absolute Shrinkage and Selection Operator)
   - Decision Trees and Random Forests

LASSO - feature selection happening implicitly

whatever is not in tree is not looked at

Filter methods

Can be seen as a preprocessing step for later learning/modelling, and thus it is done *independently* from later learning procedure.

These methods typically assign a score or rank each feature **individually** according to some statistical or information-theoretic criterion.

- Main advantage: computational efficiency
- Main limitation: they do not take into account **interactions** among features

Limitation
Sometimes, individual feature may not be relevant
but combination of features might be useful
This is not considered in filter methods

Filter methods, examples

1. Correlation Analysis: Look at correlation between target values and input features if is large, keep it else remove it
   - Features with high correlation coefficients with the target (either positive or negative) are considered more relevant.
   - Pearson correlation coefficient is commonly used for linear relationships, while other measures like Spearman rank correlation can handle non-linear relationships.
2. Information Gain:
   - Derived from information theory, information gain quantifies the amount of information obtained about the target variable by knowing the value of a particular feature.
   - Features with higher information gain are considered more informative and relevant.
3. Chi-Square Test:
   - Used for feature selection in classification tasks with categorical target variables.
   - Measures the independence between each feature and the target variable using the chi-square statistic.
4. Variance Thresholding:
   - Removes features with low variance, assuming they contain little information.

Wrapper methods

iterate over model built in different subset of features
and see how these features selected perform

Adv - unlike filter method, interaction between features is looked at

Used in combination of a machine learning algorithm, they use the algorithm's predictive performance to assess performance of subsets of features.

Wrapper methods involve training and evaluating the model multiple times on different subsets of features, which can be computationally expensive but often result in better feature subsets tailored to the specific model.

- Main advantages: take into account **interactions** among features, feature selection tailored to learning task
- Main limitation: computationally expensive due to large search space

Wrapper methods can be conceptualized as a search problem, where the goal is to find the optimal feature subset that maximizes model performance.

Various search algorithms like hill climbing, simulated annealing, and genetic algorithms can be used to explore the feature space efficiently.

Wrapper methods, examples

1. Forward Selection
   - Start with an empty set of features.
   - Iteratively add one feature at a time, evaluating the performance of the model at each step.
   - Choose the feature that improves the model's performance the most and add it to the set.
   - Repeat until a predefined stopping criterion is met.

2. Backward Elimination
   - Start with the full set of features.
   - Iteratively remove one feature at a time, evaluating the performance of the model at each step.
   - Remove the feature that causes the least deterioration in performance.
   - Repeat until a predefined stopping criterion is met.

3. Exhaustive search
   - Try out all subsets, keep the best (only with small number of features. . .)

Multi-objective Feature Selection

Multi-objective Feature Selection (MOFS)

Multi-objective feature selection is a technique used to select subsets of features that optimize **multiple objectives simultaneously**.

Typically one has several goals that one wants to achieve, for example:

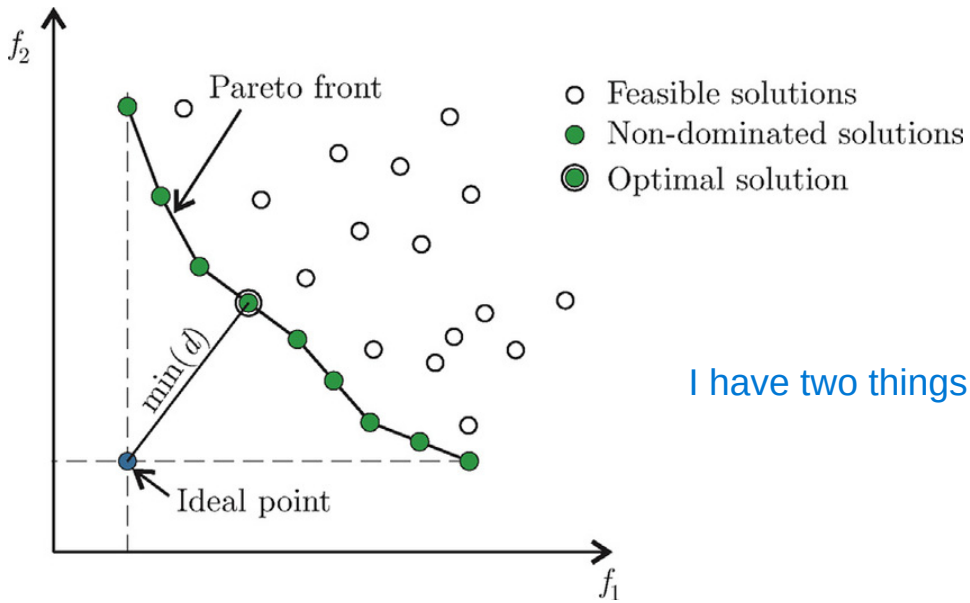
1. Maximize predictive performance
2. Minimize number of features
3. Maximize interpretability

And when we consider *several* objectives (conflicting with each other) in optimization, the notion of *optimality* may no longer apply since improving one is typically at the expense of the other. if we improve one, another degrade

In these scenarios we consider **Pareto optimality**.

Pareto optimality

A solution is **Pareto-optimal** if no other solution is superior to it in all objectives. There are many Pareto-optimal solutions, which are called **dominant solutions**.



I have two things that I want to optimize

Conceptualizes trade-offs between conflicting objectives through the **Pareto front** (Pareto frontier, Pareto set):

- Collection of all Pareto-optimal solutions.
- Represents the best achievable trade-offs between objectives.

Uche's project

Uche's project

- Given: a set of non-dominated (Pareto-optimal) solutions
- Challenge: narrow it down to exactly **one solution**

Uche's project, metrics used

1. Number of selected features (↓)
2. Balanced accuracy (↑) % of times got correct value
   - takes into account class imbalance
   - mean of accuracy of predictions for positive and negative examples
3. F1-Score (↑)
   - takes into account class imbalance
   - harmonic mean of precision and recall
4. Variance inflation factor (↓)
   - measures correlation among input features
5. Statistical parity (↑)
   - related to fairness
   - measures whether predictions are affected by sensitive features (gender, race, ..)
6. Equalised odds (↑)
   - related to fairness
   - measures equality of predictive performance (in terms of both true positive and true negative rates) for all demographic groups

Uche's project, steps

Before lab class on **March 19th**:

1. Watch her video on the **Diabetes Dataset**
2. Make sure you understand what is explained here, if not ask.

During the lab session on **March 19th**:

3. Listen to her instructions
4. You will help her by doing the selection yourselves, on several Pareto-optimal solutions on the *Diabetes Dataset* (binary classification problem)
5. Answer some questions

After gathering and studying your answers:

6. You will get the results along some explanations