Machine Learning in Healthcare

#C15 Feature Selection

Technion-IIT, Haifa, Israel

Assist. Prof. Joachim Behar
Biomedical Engineering Faculty
Technion-IIT

AIMLab.

Classification

Unknown
Target Function: $f: x \rightarrow y$

Training Examples
$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
$\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$

Hypothesis Class
$H = \{h\}$

Classification Algorithm

Final Hypothesis
$\hat{h} \approx f$

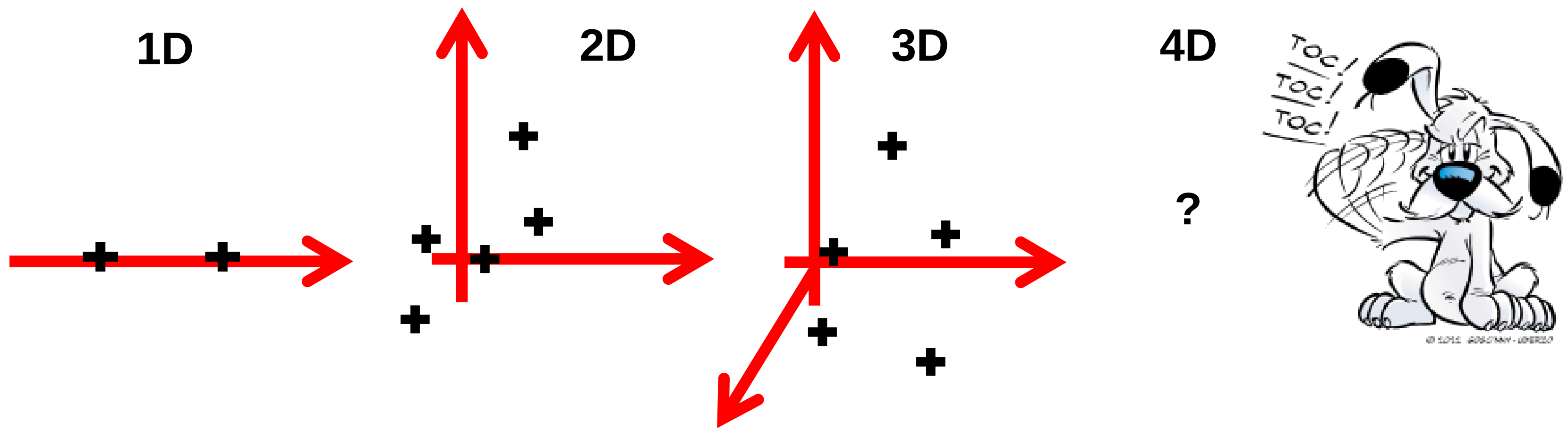
Introducing the Problem

History

- Before the 90^{th} few domains used more than 40 features.
- This has changed dramatically since then with many applications using hundreds to tens of thousands of variables (e.g. DNA microarray).
- In many instances the number of training examples is limited (e.g. cost, technical challenge) and this may cause *problems*.
- How can we identify the feature subset that is the most *adequate* for our learning task?

Many, too many features…

- Main challenges with many features:
  - **Visualization:** how to visualise data in $\mathbb{R}^N, N > 3$?? Too many features may obstruct interpretability.
  - **Curse of dimensionality:** As the number of features increases (equivalently the dimensionality of our problem) we need exponentially more examples in order to ensure our model will generalize well.
- **Other:** computing time, cost for collecting many features.



Feature Selection vs. Feature Transformation

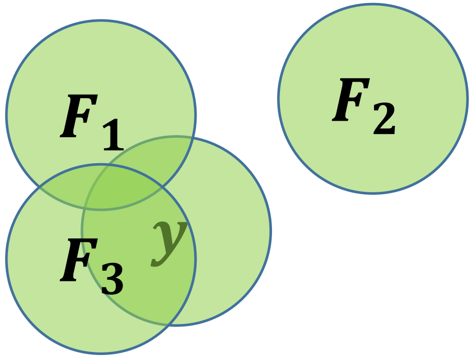
- Introducing the idea
  - $a + b + c + d = e$
  - $ab = a + b$
  - $ab + c + d = e$ Feature Transformation
  - $c = 0$
  - $ab + d = e$ Feature Selection
- Feature transformation
  - Results are not easily interpretable
- Feature selection
  - Discard non-contributing features to the prediction
  - Keep interpretability
- We will focus on **feature selection** in this lecture i.e. selecting subsets of features that are useful to build a good predictor.

Feature Selection

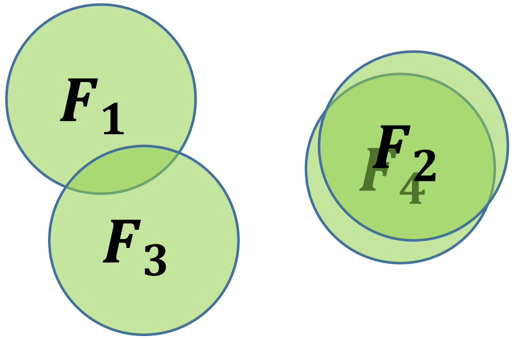
Relevance and Redundancy

- When performing feature selection we assume that some features are either **redundant** or **irrelevant**.
- These are two different notions.
  - A relevant feature may be redundant.
  - Two partially redundant features may both be relevant.
- We will seek maximum relevant and minimal redundancy in selecting our feature set.
- Intuition for both concepts: features F and response y.

Relevance



Redundancy



AIMLab.

Relevance and Redundancy

- Formal definition of **relevance**:
  - Let F be a full set of features, F_i a feature, and $S_i = F - \{F_i\}$.
  - Let C be the target Concept (Boolean)
  - Definition 1 (Strong relevance) A feature F_i is strongly relevant *iff*
    - $P(C|F_i, S_i) \neq P(C|S_i)$.
    - A feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution.
  - Definition 2 (Weak relevance) A feature F_i is weakly relevant *iff*
    - $P(C|F_i, S_i) = P(C|S_i)$, and
    - $\exists S_i' \subseteq S_i$ such that $P(C|F_i, S_i') \neq P(C|S_i')$
    - The feature is not always necessary but may become necessary for an optimal subset at certain conditions.
  - Corollary 1 (Irrelevance): A feature F_i is irrelevant *iff*:
    - $\forall\, S_i' \subseteq S_i, P(C|F_i, S_i') = P(C|S_i')$.
    - The feature is not necessary at all.
- For a formal definition of redundancy see Yu et al.

Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." Journal of machine learning research 5.Oct (2004): 1205-1224.

Categories of Feature Selection Algorithms

- **Filters:** select subsets of features as a pre-processing step, independently of the model. That is filters, do not take into account the classifier that is used.
- **Wrappers:** assess subsets of features according to their usefulness to a given predictor.
- **Embedded:** directly optimize a two-part objective function with a goodness of fit term and a penalty for a large number of features.

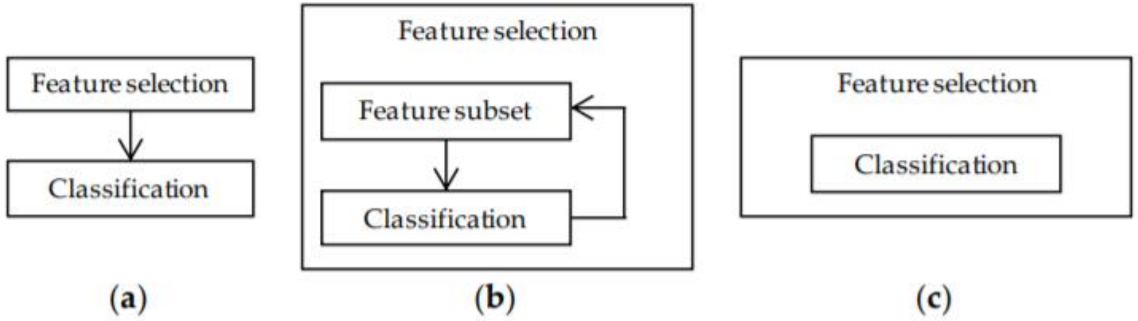


Figure 3. (**a**) Filter, (**b**) wrapper, and (**c**) embedded feature selection methods. Filter methods perform the feature selection independently of construction of the classification model. Wrapper methods iteratively select or eliminate a set of features using the prediction accuracy of the classification model. In embedded methods the feature selection is an integral part of the classification model.

Image: Suppers, Anouk, Alain Gool, and Hans Wessels. "Integrated chemometrics and statistics to drive successful proteomics biomarker discovery." *Proteomes* 6.2 (2018): 20.

Categories of Feature Selection Algorithms

- **Filters:** select subsets of features as a pre-processing step, independently of the model.
  - (+) Computation time.
  - (+) Robust to overfitting.
  - (-) Tend to select redundant variables.
  - (+/-) The set of feature selected is not tuned to a particular model.
- **Wrappers:** assess subsets of features according to their usefulness to a given predictor.
  - (+) Detect possible interactions between variables.
  - (+) Accuracy.
  - (-) Tuned for a specific classifier.
  - (-) Increased overfitting risk when a low number of examples.
  - (-) Computation time.
- **Embedded:** directly optimize a two-part objective function with a goodness of fit term and a penalty for a large number of features. Thus feature selection is performed simultaneously with classification.
  - (+) Usually provide the best performing feature set **for the type of model chosen**.
  - (-) Computation time.

AIMLab.

Examples of Filters, Wrappers and Embedded Algorithms

- We will explore more in details one example for each category. Examples of algorithms for each category is given below.
- Filters
  - Pearson correlation coefficient
  - Statistical test
  - Minimum redundancy and maximum relevance (mRMR)
  - Mutual information
  - Relief-based algorithms
- Wrappers
  - Genetic algorithm
- Embedded
  - LASSO
  - Recursive feature elimination (RFE)
  - Other popular: Ridge regression, Boruta.

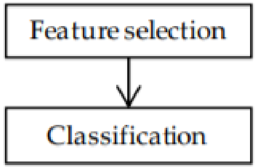
Filters: Pearson correlation coefficient

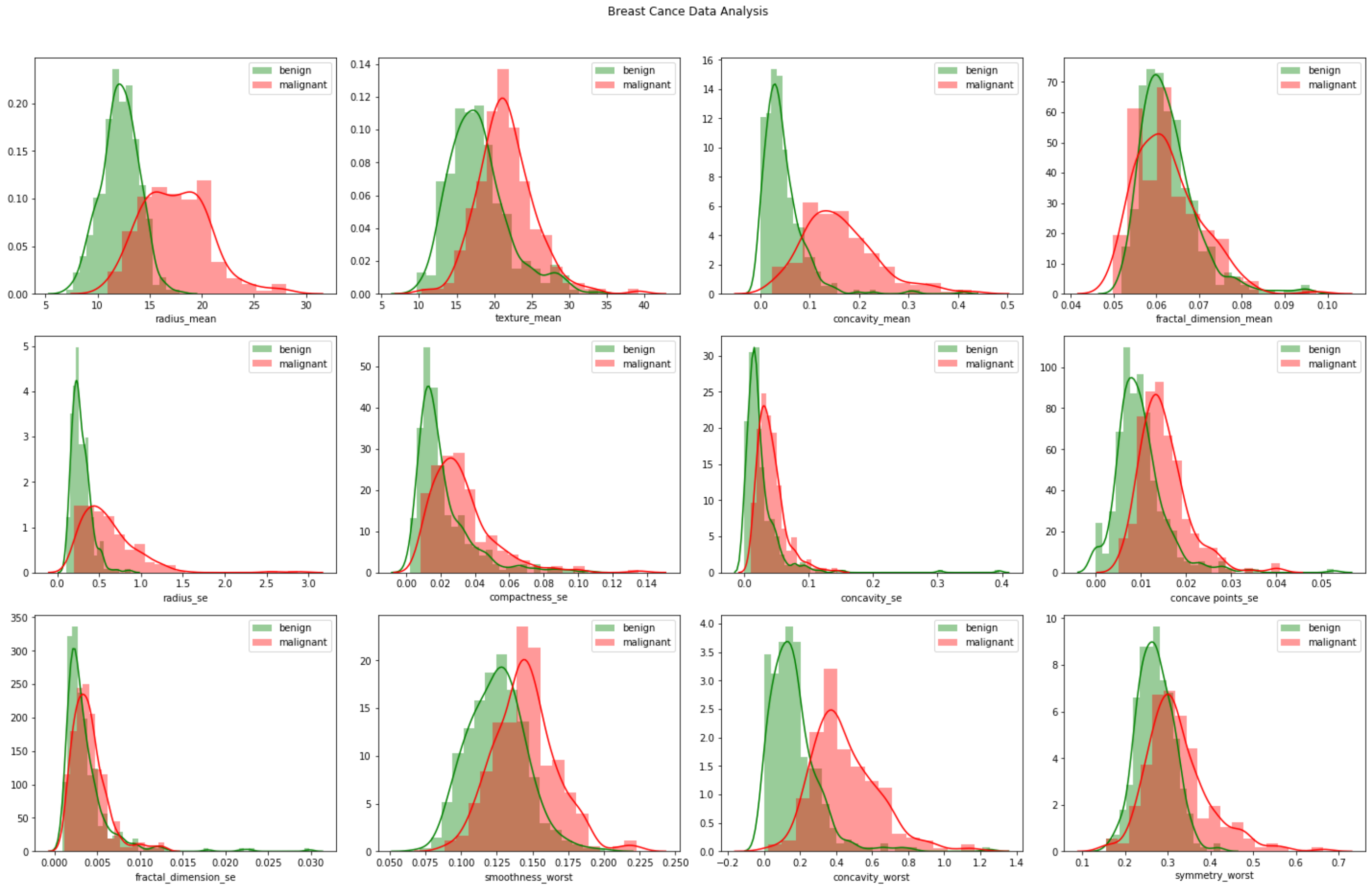
Feature selection → Classification

- Pearson correlation coefficient,
  - $R(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i) var(Y)}}$
  - The estimate of $R(i)$ between a given feature x_i and the response outcome variable y:
  - $R(i) = \frac{\sum_{k=1}^{m} (x_i^{(k)} - \bar{x}_i)(y^{(k)} - \bar{y})}{\sqrt{\sum_{k=1}^{m} (x_i^{(k)} - \bar{x}_i)^2 \sum_{k=1}^{m} (y^{(k)} - \bar{y})^2}}$
  - k: example number.
- In linear regression, the coefficient of determination $R(i)^2$ corresponds to the total variance around the mean $\bar{y}$ that is explained by the linear relation between x_i and y.
- In this context, using $R(i)^2$ as the variable ranking criterion will enforce a ranking according to the goodness of linear fit.

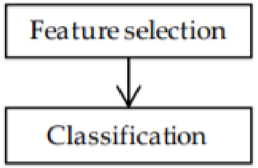
Filters: Statistical test

- Get the p-value
- Rank them
- Remove features where both groups come from the same distribution according to the statistical test.

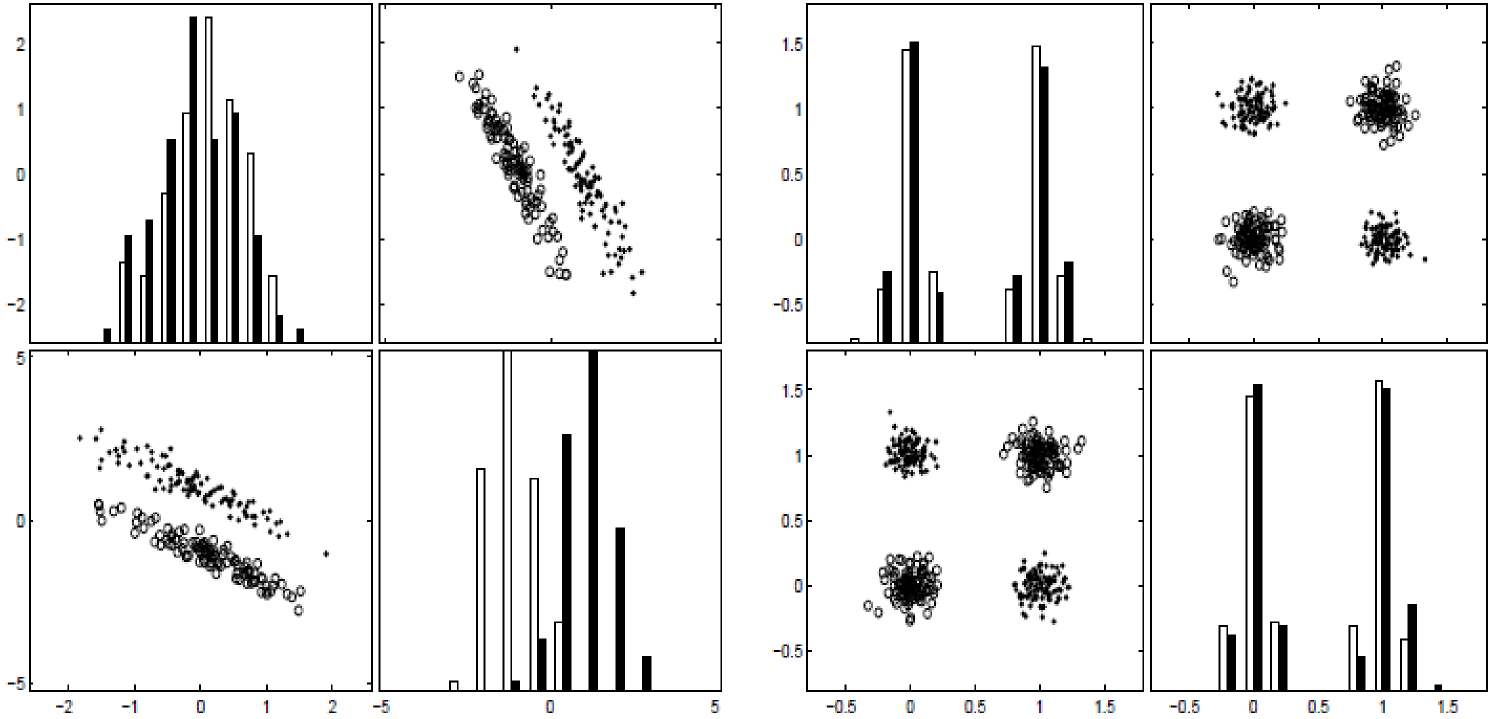




Filters Limitations

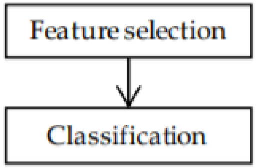


- A variable useless by itself can be useful together with others.
- This example highlight the importance of **features interaction**.



Guyon, Isabelle, and André Elisseeff. J. Mach Learn. Res. 3 (2003).

Filters Limitations



- Filters may have important limitations such as:
  - Correlation does not capture non-linear relationship between feature and target.
  - A variable *useless* by itself can be useful together with others.
  - If only looking at the feature to target - traditional filters (e.g. correlation, mutual information) do not consider relationships among features. Thus the selected features may be correlated and information redundant.
- We will see how to alleviate some of these limitation with a popular algorithm called: **minimum redundancy and maximum relevance (mRMR).**

Filter: mRMR

Feature selection → Classification

- Minimize redundancy:
  - $\min_S W(S), \quad W(S) = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j)$
  - S is the set of features.
  - $I(f_i, f_j)$ is mutual information between feature f_i and f_j.
- Maximize relevance:
  - $max_S V(S), \quad V(S) = \frac{1}{|S|} \sum_{f_i \in S} I(c, f_i)$
  - c: target classes (e.g. types of different cancers, atrial fibrillation or not.)
- mRMR criterion:
  - $mRMR = max_S \left[\frac{1}{|S|} \sum_{f_i \in S} I(f_i, c) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j) \right]$
- In practice mRMR shows to perform well.

Peng, Hanchuan, et al. IEEE Trans. Pattern Analysis & Machine Intelligence 8 (2005).

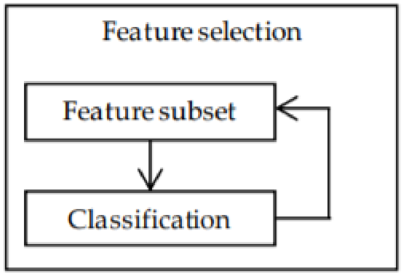
AIMLab.

Filter: mRMR

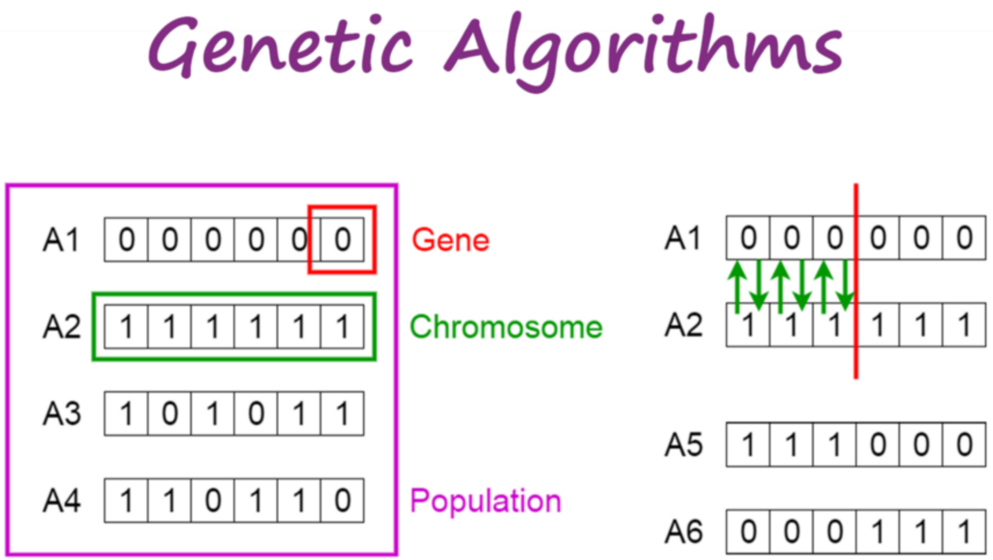
Feature selection → Classification

- The purpose of mRMR is to select features that are mutually far away from each other while still having high “correlation” to the classification variable.
- Usually in mRMR mutual information is used and not “correlation”.
- mRMR is an approximation to maximizing the dependency between the joint distribution of the selected features and the classification variable.
- In practice mRMR performs well!
- However:
  - Does not account for non-pairwise redundancy.
    - $W(S) = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j)$
  - Does not account for interactions between features and the target.
    - $V(S) = \frac{1}{|S|} \sum_{f_i \in S} I(c, f_i)$

Wrappers: Genetic Algorithm



- Optimization algorithm.
- Type of evolutionary algorithm (EA).
- Mimicking evolutive biology techniques.
- robust, adaptive search techniques.



https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3

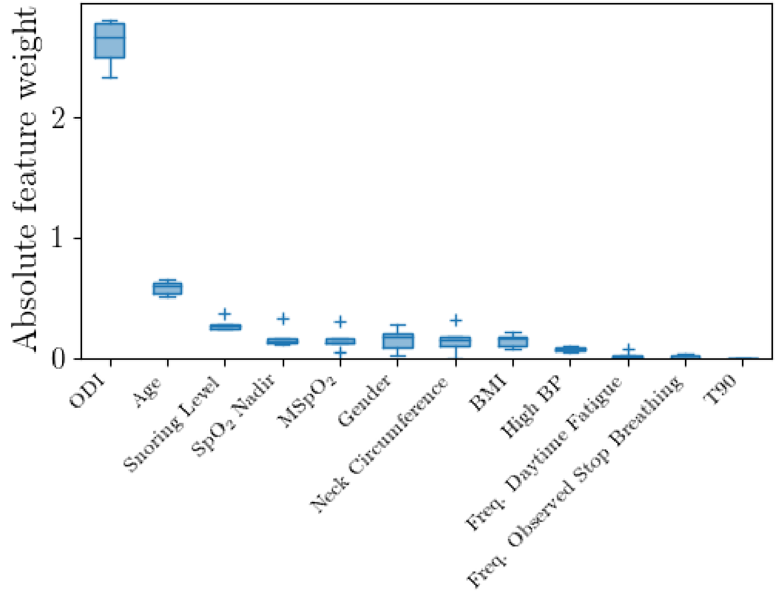
AIMLab.

Feature selection

Classification

Embedded Methods: LASSO

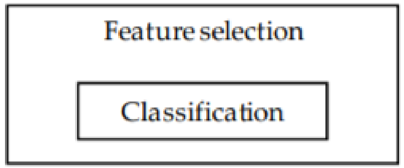
- LASSO Regularization:
  - Reminder, the cost function in LASSO regularized LR:
    - $J(w) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}log\left(h_w\left(x^{(i)}\right)\right) - (1 - y^{(i)})\log(1 - h_w\left(x^{(i)}\right))\right] - \frac{\lambda}{2m}\sum_{j=1}^{n}|w_j|$
  - LASSO is a form of regularization.
  - In practice it makes some coefficients tend to zero so it is essentially doing feature selection.
- Example: using LASSO for estimating feature importance in obstructive sleep apnea diagnosis.



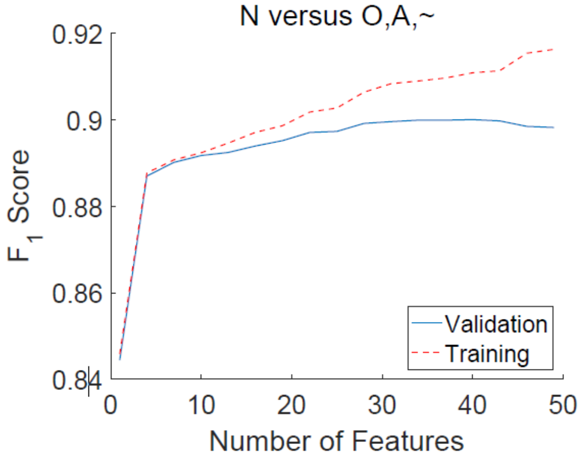
Behar J, Palmius N et al. 11 EclinicalMedicine (2019).

AIMLab.

Embedded Methods: RFE

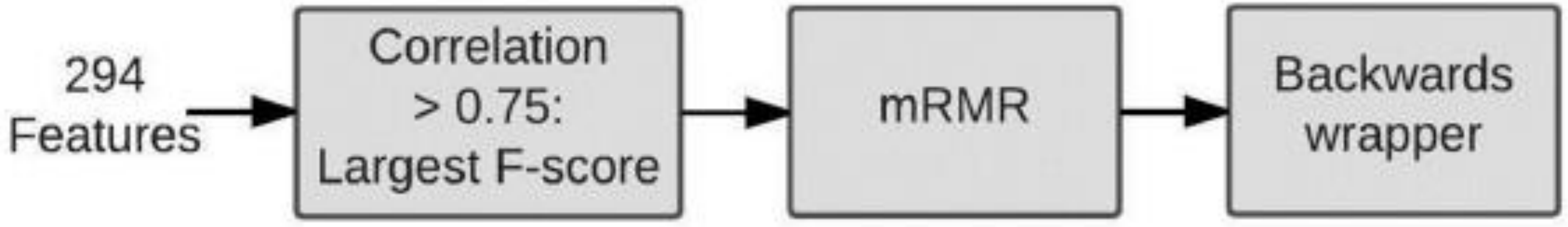


- Recursive feature elimination (RFE)
- Select features by recursively considering smaller and smaller sets of features:
  - Train on the initial set of features and obtain the importance of each feature
  - The least important features are pruned
  - Procedure is recursively repeated.
- Example: RFE in SVM for atrial fibrillation prediction.
  - Recall the dual Lagrangian:
    - $\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y^{(n)} y^{(m)} \phi\left(x^{(n)}\right)^T \phi\left(x^{(m)}\right)$
  - Cost function for features ranking:
    - $J(k) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y^{(n)} y^{(m)} \phi\left(x^{(n)}\right)^T \phi\left(x^{(m)}\right) - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y^{(n)} y^{(m)} \phi\left(x_{-k}^{(n)}\right)^T \phi\left(x_{-k}^{(m)}\right)$
    - The $-k$ means that the feature k has been removed.
    - We look at the difference in the cost function between including and excluding feature k for all k.
    - We can use that for features ranking.



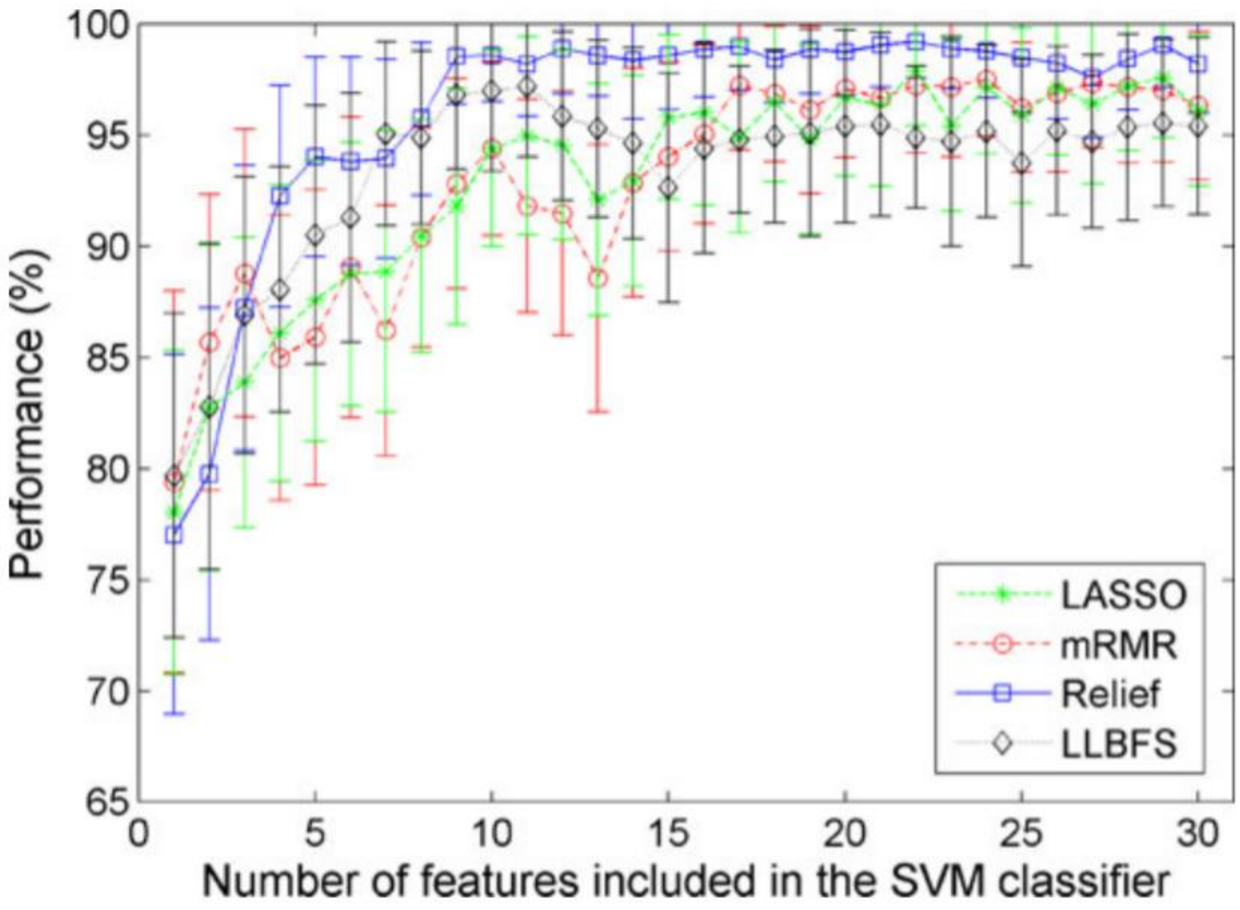
sklearn.feature_selection.RFE
Behar J et al. Computing in Cardiology 2017.

Combining Different Flavors



Vandecasteele, Kaat, et al. Sensors 17.10 (2017).

Comparing Algorithms



Tsanas, Athanasios, et al. IEEE TBME 59.5 (2012).

Take home

- Feature transformation versus feature selection.
- **Relevance, redundancy** and **feature interaction**. Need to find a trade-off.
- Filters, Wrappers and Embedded methods for feature selection.
- Simple filters do not take into account the relationship between features.
  - To alleviate that use mRMR.
- Wrappers and Embedded methods may give better results for a given classifier but it will be more computational.
- Examples of popular algorithms which works well:
  - LASSO.
  - mRMR.
  - Recursive feature elimination.
  - Genetic algorithm.

References

[1] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research3.Mar (2003): 1157-1182.
[2] Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." Journal of machine learning research 5.Oct (2004): 1205-1224.
[3] Athanasios Tsanas. Information Driven Healthcare. Course notes on Feature selection I – Concepts.
[4] Rokach, Lior. "Genetic algorithm-based feature set partitioning for classification problems." Pattern Recognition 41.5 (2008): 1676-1700.