Feature Selection

Ahmet Sacan

Some slides are from:

http://courses.cs.tamu.edu/rgutier/cs790_w02/l16.pdf

https://www.cse.unr.edu/~bebis/CS479/Lectures/FeatureSelection.ppt

Feature Selection vs. Feature Extraction

- Feature Selection: select a subset of features to improve classification (or to reduce the cost of calculating the feature vector).
- Feature Extraction: represent the samples in a new feature space.
  - Dimension Reduction methods. e.g., PCA
  - Calculation of new features from raw data. e.g, histograms or LocalBinaryPatterns extracted from image data.

Why Feature Selection?

- Features may be expensive to obtain
  - You measure all genes in research (expensive), but want to use only a few for the final biomarker set
- Easier interpretation of the classifier rules/decision
  - Measurement units (length, weight, etc.) are preserved in feature selection, but lost in feature extraction.
- Fewer parameters to learn for pattern recognition
  - Improved generalization
  - Reduced time & space requirements

Feature Selection

- Given F features, find the subset that gives the best classification performance.
- Number of possible subsets is exponential in F :
  - Selecting k features: $\binom{F}{k}$
  - Selecting any subset of features: 2^F
- Exhaustive Search impractical.

Objective function

- Objective function: the "goodness" of a candidate subset.
- Filter:
  - Evaluate feature subsets by their information content; e.g. interclass distance, statistical dependence.
- Wrapper:
  - Evaluate feature subsets by their predictive accuracy (cross-validation) when used in a pattern classifier

Filters vs. Wrappers

- **Filters**:
  - Fast execution: a non-iterative computation on the dataset
  - Generality: evaluate intrinsic properties of data. Can use the feature set with many classifier methods
  - Tendency to select large subsets. Objective function is generally monotonic, where selecting all features is optimal. Need to (arbitrarily) decide on a target number of features.
- **Wrappers**:
  - Accuracy: Achieve better prediction accuracy.
  - Lack of generality: The selected feature set is classifier-specific and may not be the best set when another classifier is used.
  - Slow execution: Need to train & test classifier for subsets of features to identify the best performing subset.

Filters

- Distance or Separability Measures
  - Distance between classes: Euclidean, City distance
  - Distance between & within classes: determinant of $S_W^{-1}S_B$ (LDA eigenvalues)
- Correlation and information-theoretic Measures
  - Idea: Good feature sets are highly correlated with the target class and not correlated with each other.
  - Linear: $\frac{\sum_i \rho_{i,class}}{\sum_i \sum_j \rho_{i,j}}$ where $i, j \in featureset\ F$
  - Mutual Information: Decrease in the entropy of the class by knowledge of X_F

Filter Methods

Method		X			Y			Comments
Name	Formula	B	M	C	B	M	C	
Bayesian accuracy	Eq. 3.1	+	s		+	s		Theoretically the golden standard, rescaled Bayesian relevance Eq. 3.2.
Balanced accuracy	Eq. 3.4	+	s		+	s		Average of sensitivity and specificity; used for unbalanced dataset, same as AUC for binary targets.
Bi-normal separation	Eq. 3.5	+	s		+	s		Used in information retrieval.
F-measure	Eq. 3.7	+	s		+	s		Harmonic of recall and precision, popular in information retrieval.
Odds ratio	Eq. 3.6	+	s		+	s		Popular in information retrieval.
Means separation	Eq. 3.10	+	i	+	+			Based on two class means, related to Fisher's criterion.
T-statistics	Eq. 3.11	+	i	+	+			Based also on the means separation.
Pearson correlation	Eq. 3.9	+	i	+	+	i	+	Linear correlation, significance test Eq. 3.12, or a permutation test.
Group correlation	Eq. 3.13	+	i	+	+	i	+	Pearson's coefficient for subset of features.
χ^2	Eq. 3.8	+	s		+	s		Results depend on the number of samples m.
Relief	Eq. 3.15	+	s	+	+	s	+	Family of methods, the formula is for a simplified version ReliefX, captures local correlations and feature interactions.
Separability Split Value	Eq. 3.41	+	s	+	+	s		Decision tree index.
Kolmogorov distance	Eq. 3.16	+	s	+	+	s	+	Difference between joint and product probabilities.
Bayesian measure	Eq. 3.16	+	s	+	+	s	+	Same as Vajda entropy Eq. 3.23 and Gini Eq. 3.39.
Kullback-Leibler divergence	Eq. 3.20	+	s	+	+	s	+	Equivalent to mutual information.
Jeffreys-Matusita distance	Eq. 3.22	+	s	+	+	s	+	Rarely used but worth trying.
Value Difference Metric	Eq. 3.22	+	s		+	s		Used for symbolic data in similarity-based methods, and symbolic feature-feature correlations.
Mutual Information	Eq. 3.29	+	s	+	+	s	+	Equivalent to information gain Eq. 3.30.
Information Gain Ratio	Eq. 3.32	+	s	+	+	s	+	Information gain divided by feature entropy, stable evaluation.
Symmetrical Uncertainty	Eq. 3.35	+	s	+	+	s	+	Low bias for multivalued features.
J-measure	Eq. 3.36	+	s	+	+	s	+	Measures information provided by a logical rule.
Weight of evidence	Eq. 3.37	+	s	+	+	s	+	So far rarely used.
MDL	Eq. 3.38	+	s		+	s		Low bias for multivalued features.

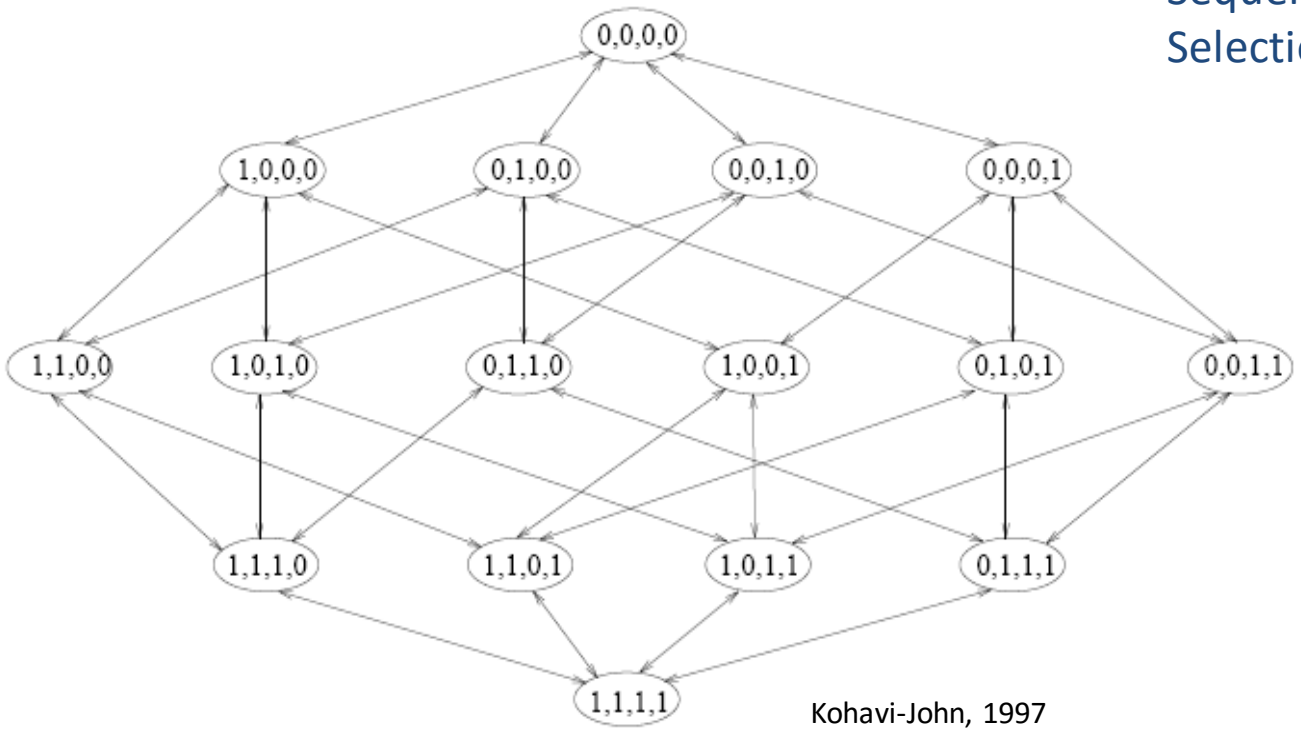
Feature Extraction - Foundations and Applications. Guyon et.al., (eds.)

Wrapper Methods: Search Strategies

- Naïve: order features by their individual goodness and keep top k.
  - Drawback: Worse features may be better in combination.
- Exponential algorithms
  - Exhaustive search
  - Branch & Bound
- Sequential algorithms
  - Iteratively add or remove features
  - Tendency to be trapped in local minima
- Randomized algorithms
  - Randomize to escape local minima
  - Genetic algorithms

Sequential Search

Sequential Forward Selection (SFS)

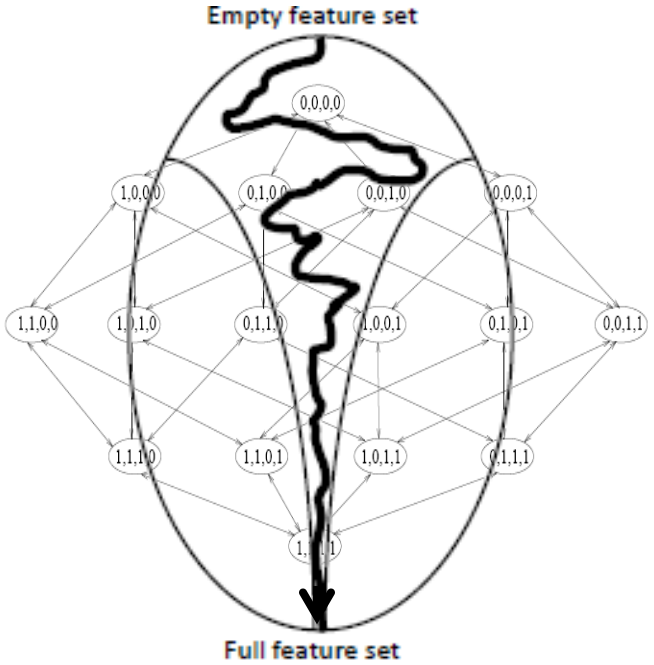


Kohavi-John, 1997

Sequential Backward Selection (SBS)
aka. Recursive Elimination (RE)

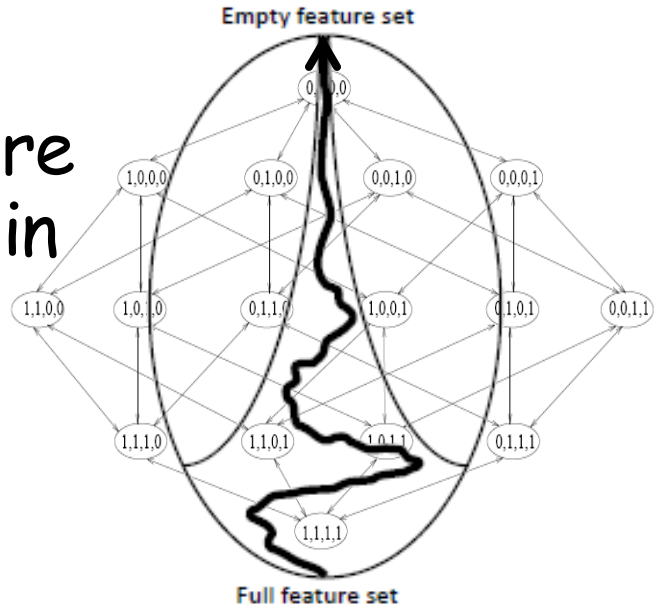
Sequential Forward Selection

- Start with empty set $F = \{\}$
- Iterate:
  - Find (and add) the feature that when added to F results in the best objective function value.
  - Optional: Stop if objective function value cannot be improved by adding a new feature.



Sequential Backward Selection

- Start with $F = \{all\ features\}$
- Iterate:
  - Find (and remove) the feature that when removed, results in the best objective function value.
  - Optional: Stop if objective function value cannot be improved by removal of a feature.

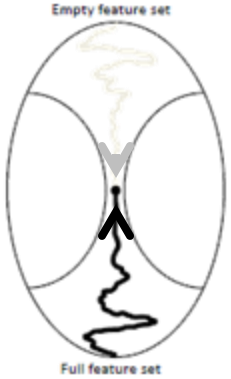


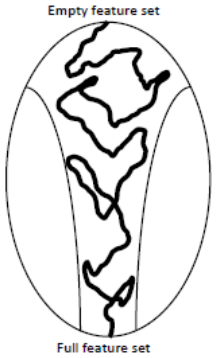
SFS and SBS

- When to use SFS vs SBS:
  - SFS works well when the optimal subset is small.
    - SFS is computationally more efficient if you stop before considering the full feature set.
  - SBS works well when the optimal subset is large.
- Limitations of SFS and SBS:
  - SFS is unable to remove features that become non-useful with the addition of additional features.
  - SBS is unable to re-evaluate the usefulness of a feature once it has been discarded.
- Variations:
  - Bidirectional Search
  - Plus-L, minus-R selections (LRS)
  - Sequential floating forward/backward selection

SFS/SBS Variants

- Bidirectional Search
  - Perform SFS and SBS simultaneously
    - Features selected by SFS are not removed by SBS
    - Features removed by SBS are not added by SFS
- Plus-L, minus-R selection (LRS)
  - L>R: F={}; iterate: add L, remove R.
  - L<R: F=all; iterate: remove R, add L.
- Sequential Floating Forward/Backward Selection (SFFS, SFBS)
  - Similar to LRS
  - SFFS: After each forward step, perform backward steps as long as the objective function improves.





Genetic Algorithms (GAs)

- Model the search for the best solution as evolution of a population of candidate solutions.
- Requirements:
  - Genetic representation of the solution domain
    - Commonly: bit string representation
  - Fitness function
    - f(candidate solution) → fitness value

Feature Selection by Genetic Algorithm

- Each chromosome:
  - Bit string specifying whether or not a feature is used.
- Population of encoded Solutions:
  - E.g., 4 samples in population, 32 features:

00010101 00111010 11110000

00010001 00111011 10100101

00100100 10111001 01111000

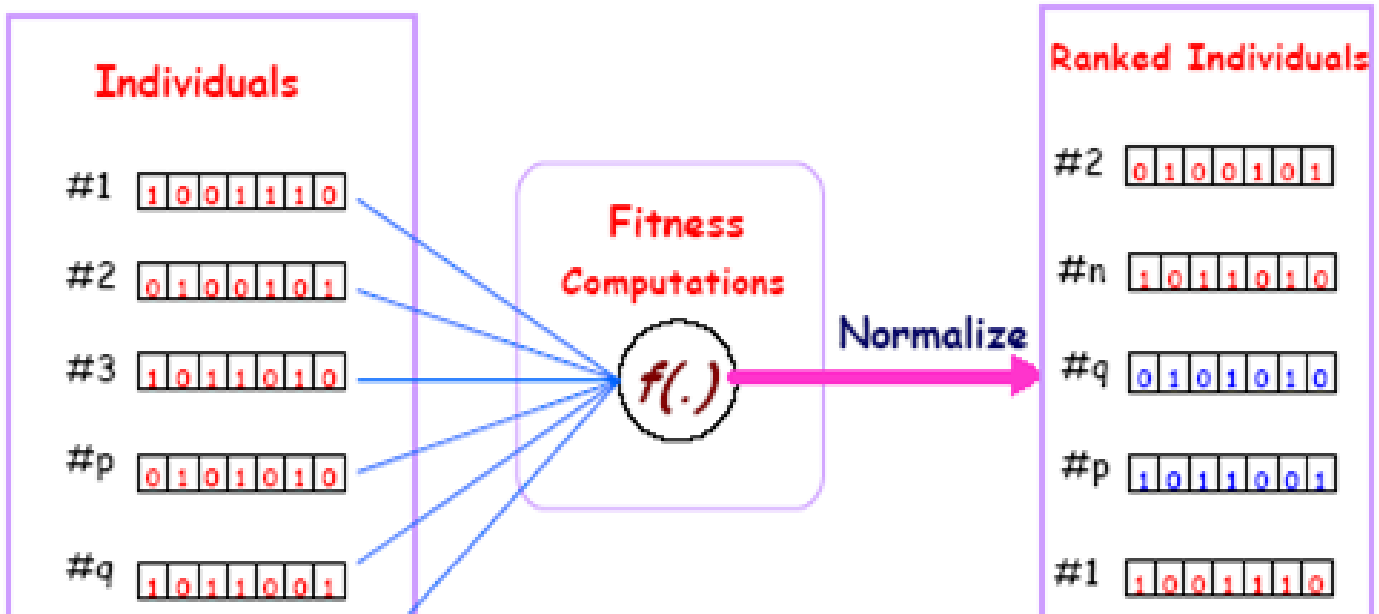
11000101 01011000 01101010

GA: Fitness evaluation

- Fitness of an "individual": Prediction performance when that feature subset is used by the classifier. eg.,

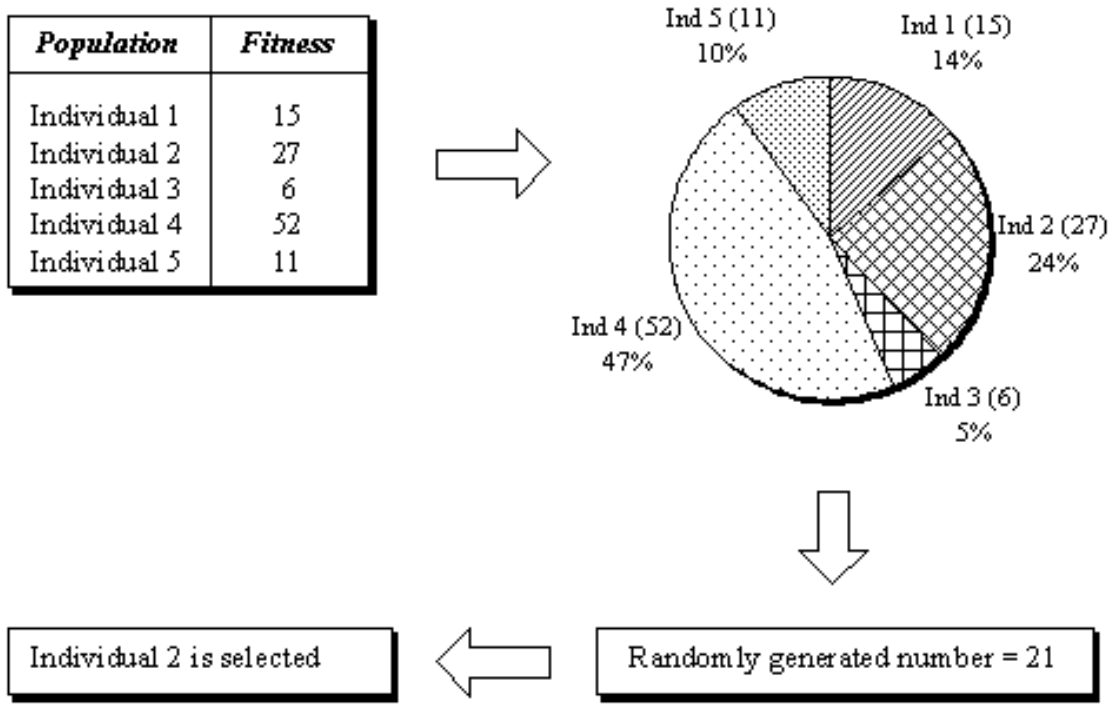
$$fitness(F) = w_1 * Accuracy + w_2 * NumZeros$$

$$w_1 \gg w_2$$



GA: Selection

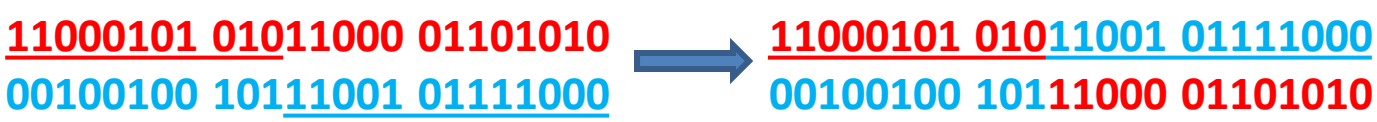
- Best-performing individuals survive
  - Perform selection probabilistically, e.g. using "Roulette Wheel Selection"



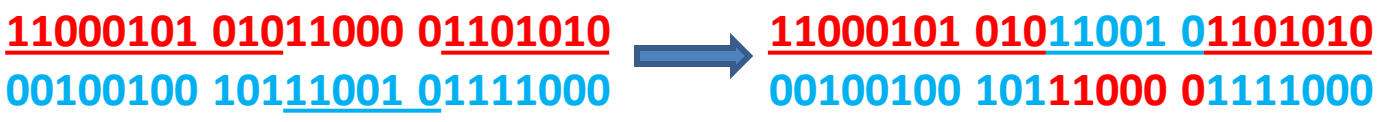
http://www.softchitech.com/ec_intro_html

GA: Cross-over

- Cross-over: hope to combine the good parts of the two solutions to obtain a better solution.
- Perform cross-over with probability p_c (typically [0.5..0.8])
  - Single-Point:



  - Two-Point:



GA: Mutation

- Mutation adds diversity to the population of solutions.
- Perform mutation with probability p_m (typically [0.001..0.01])

GA: Termination

- Termination Criteria:
  - A solution is found that satisfies minimum criteria
  - Fixed number of generations reached
  - Allocated budget (computation time/money) reached
  - The highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
  - Manual inspection
  - Tired of waiting
  - Any Combinations of the above