David R. Thompson
Jet Propulsion Laboratory, California Institute of Technology

Feature Selection

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Objectives

1. Know techniques for combinatorial feature selection
2. Know the difference between wrapper and filter methods
3. Use both forward and backward feature selection

Forward greedy selection

Wikipedia.es

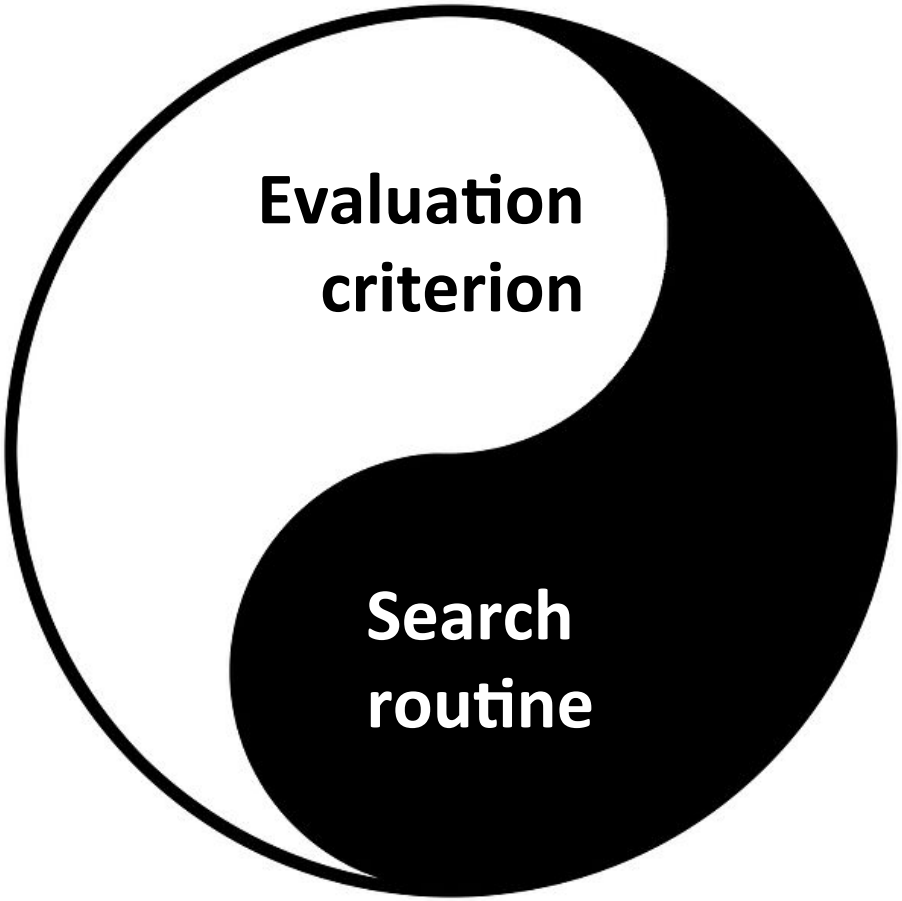
Motivation

- Reduce the number of dimensions for pattern recognition and statistical modeling
- Reveal key relationships in the data
- Preserve information

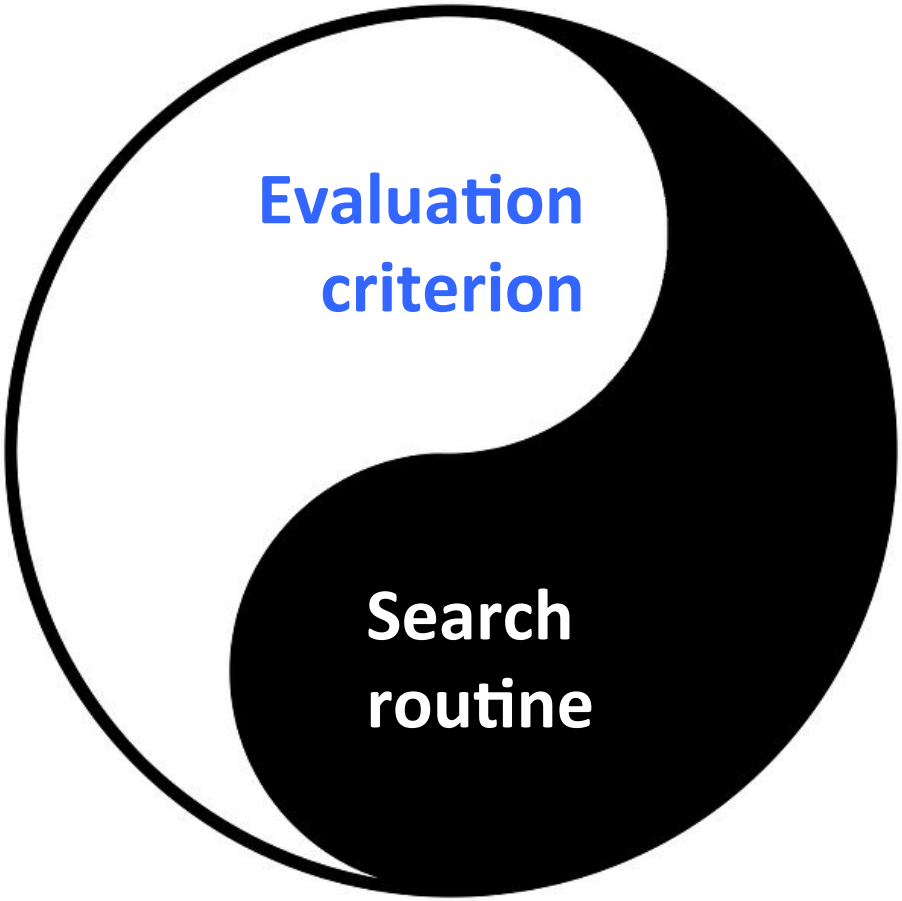
Forward greedy selection

Wikipedia.es

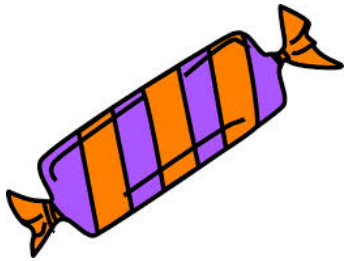
Two halves of feature selection



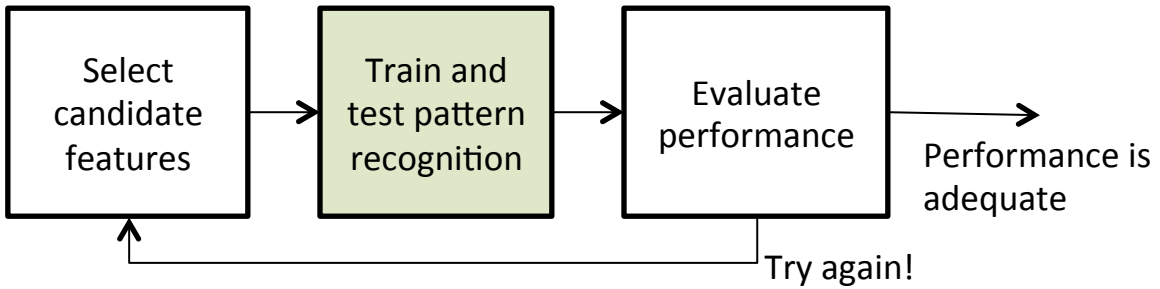
Two halves of feature selection



Evaluation: Wrappers vs. Filters



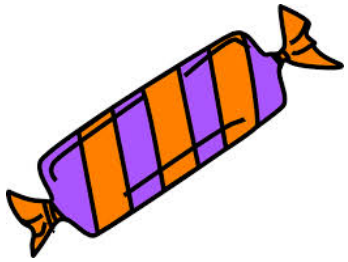
Wrappers evaluate features using pattern recognition performance



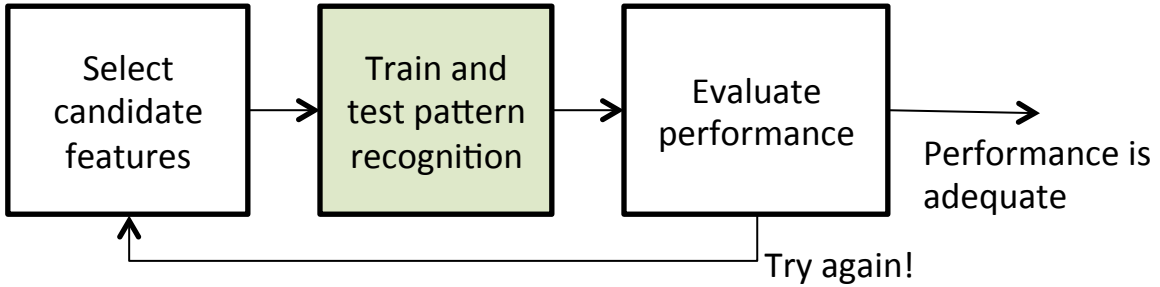
[Kohavi and John, 1997; Das, 2001]

Pixbay, wikipedia

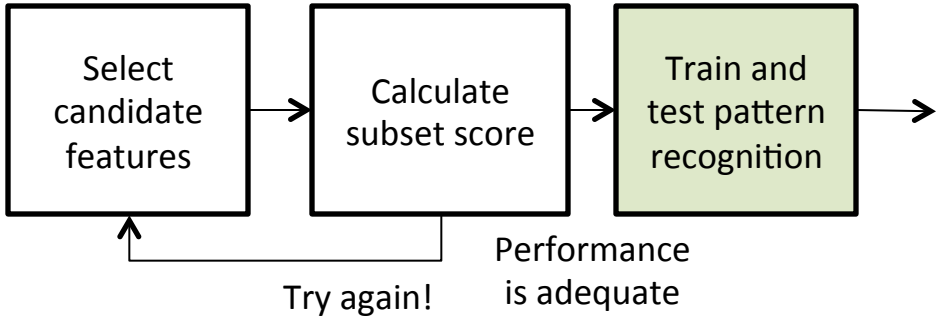
Evaluation: Wrappers vs. Filters



Wrappers evaluate features using pattern recognition performance



Filters evaluate features using intrinsic properties of the data

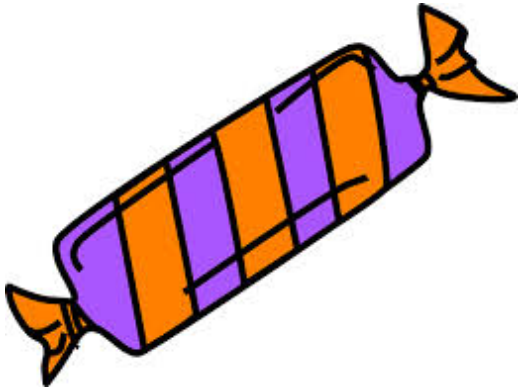


[Kohavi and John, 1997; Das, 2001]

Pixbay, wikipedia

"Wrapper" criterion

- Error on held out data



"Wrapper" advantages

- Accurate indicator of performance

"Wrapper" disadvantages

- Computational / operational complexity
- Specificity

Pixbay, wikipedia

Example "Filter" criteria

- Kolmogorov-Smirnov test,
- Pearson correlation [Miyahara and Pazzani, 2000]
- Mutual information [Torkkola, 2003]
- Fisher Score [Furey et al., 2000]

"Filter" advantages

- Good when core pattern recognition cannot be performed on the full dataset

"Filter" disadvantages

- Implies a new (redundant) model?

Pixbay, wikipedia

Another filter: Conditional Mutual Information

Conditional entropy

$$I(Y; a_i | A) = H(Y | A) - H(Y | a_i, A)$$

Conditional Mutual information

Current set

Candidate feature

[wikipedia, Fleuret JMLR 2004]

Another filter: Conditional Mutual Information

Conditional entropy

$$I(Y; a_i | A) = H(Y|A) - H(Y|a_i, A)$$

Conditional Mutual information

Current set

Candidate feature

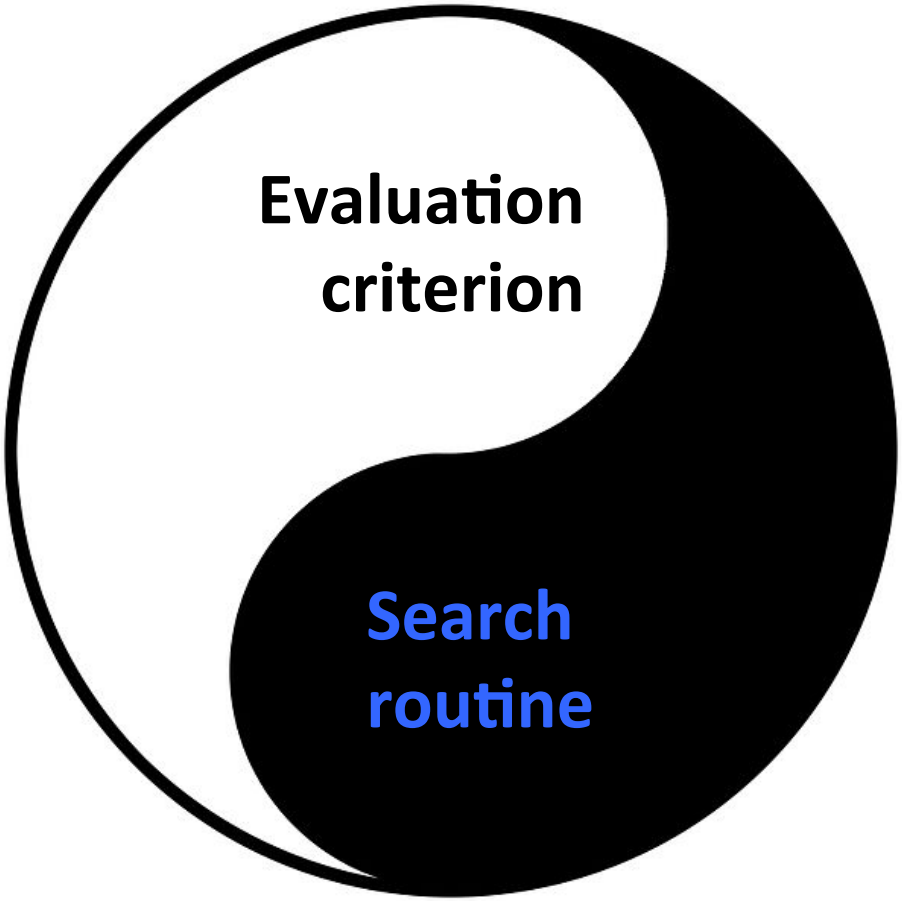
Where H(Y|A) is the *Conditional Entropy:*

$$H(Y|X) \equiv \sum_{x \in \mathcal{X}} p(x)\, H(Y|X=x)$$

$$= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x,y)}.$$

[wikipedia, Fleuret JMLR 2004]

Two halves of feature selection



Greedy forward search

Start with empty feature set

$$A = \varnothing$$

While performance improves:

- For each candidate feature a_i:
  - Add a_i to the set $A' = A \cup \{a_i\}$
  - Evaluate A' using selection criterion
- Add the best feature a^*

$$A = A \cup \{a^*\}$$

Forward greedy selection

Greedy backward elimination

Start with complete feature set

$$A = \{a_1, a_2 \ldots a_n\}$$

While performance improves:

For each candidate feature $a_i \in \mathcal{A}$

Remove a_i from the set $A' = A \setminus a_i$

Evaluate A' using selection criterion

Remove optimal feature a^*

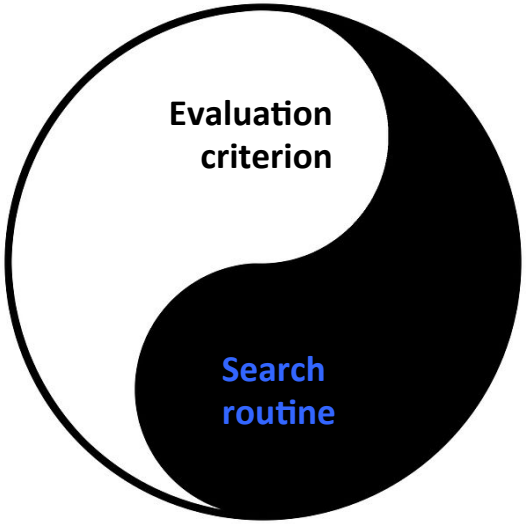
$$A = A \setminus a^*$$

Eventually performance suffers

Wikipedia

Non-greedy search

- Simulated annealing
- Branch and bound
- Genetic algorithms?



Summary

Every feature selection system has:

- An evaluation criterion
  - "Wrappers"
  - "Filters"
- Search strategy
  - Forward selection
  - Backward elimination