Boğaziçi University EE Department

ee58J | 2012 – 2013 spring

Data Mining for Visual Media

Lecture IX

Feature Selection*

Ceyhun Burak Akgül, PhD in EE

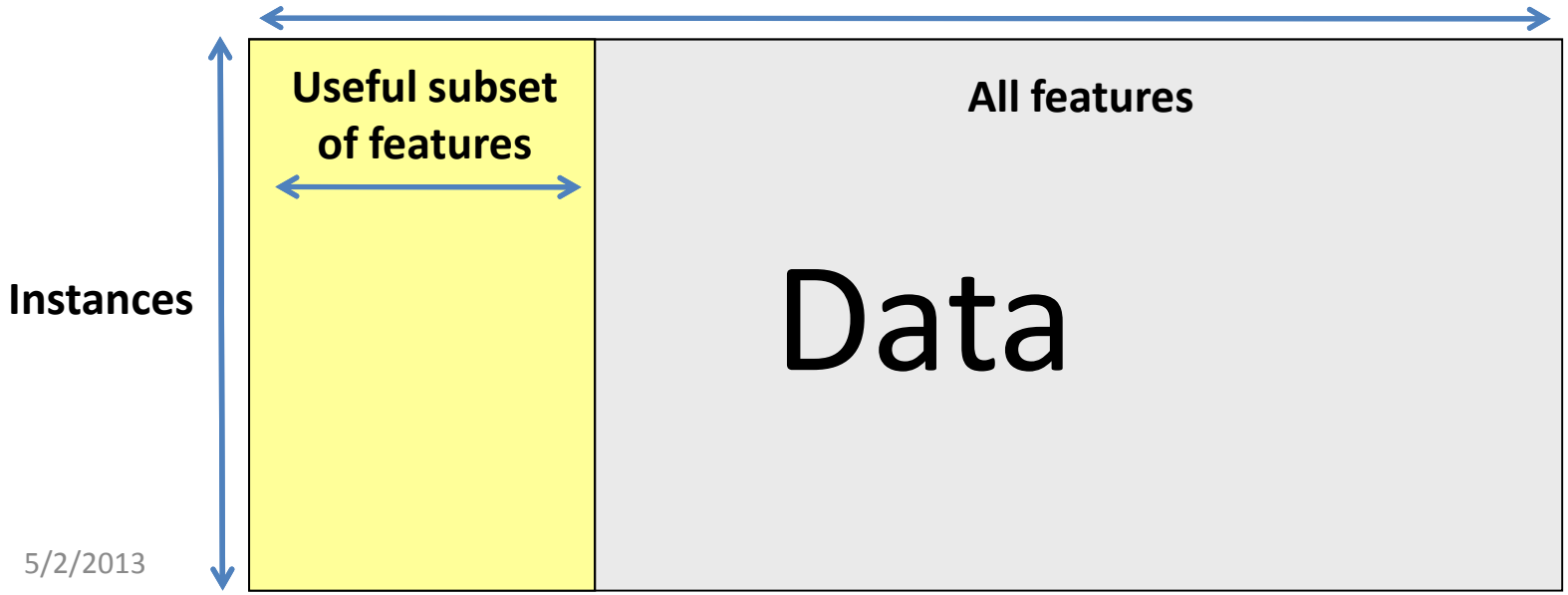
www.cba-research.com

* Based on Isabelle Guyon's slides at http://clopinet.com/

Feature Selection: Why?

For any data mining problem:

- There are usually a lot of features that can be computed: thousands to millions. Examples?
- One usually doesn't know in advance which ones are good for the present problem
- Add to this time and space constraints!



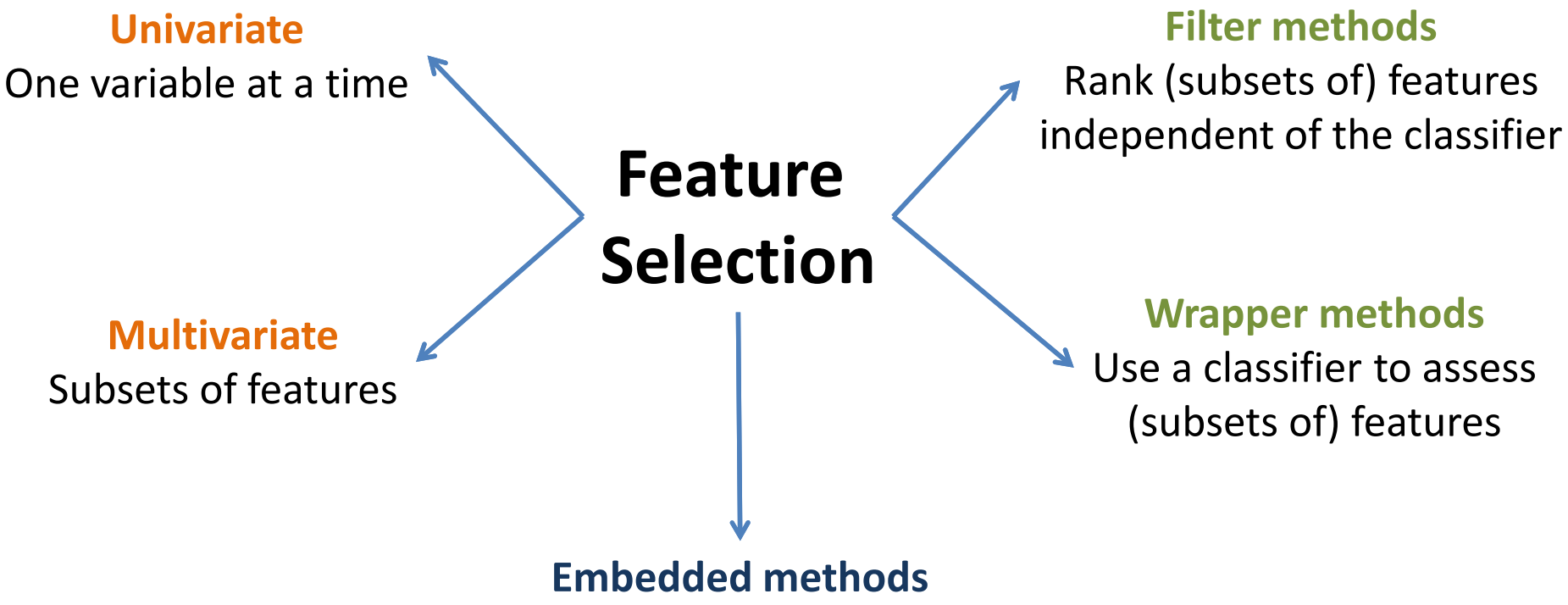
Example Application

Gender Classification

- Images are registered -

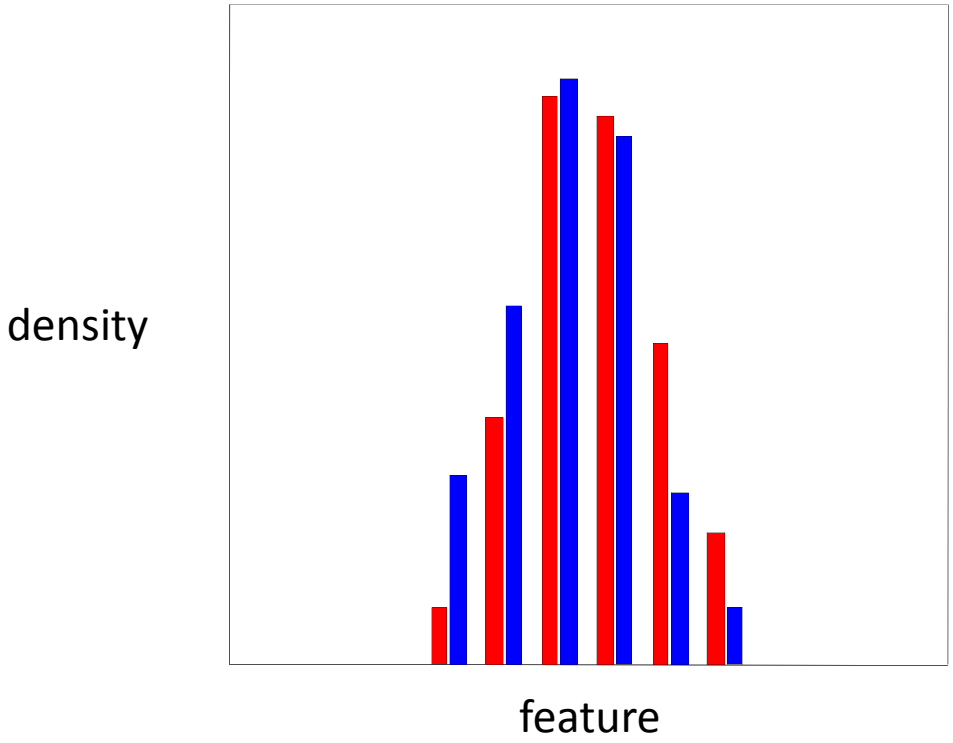
Which subset the features tells more about the gender?

Taxonomy of FS Methods



Univariate Filter Methods

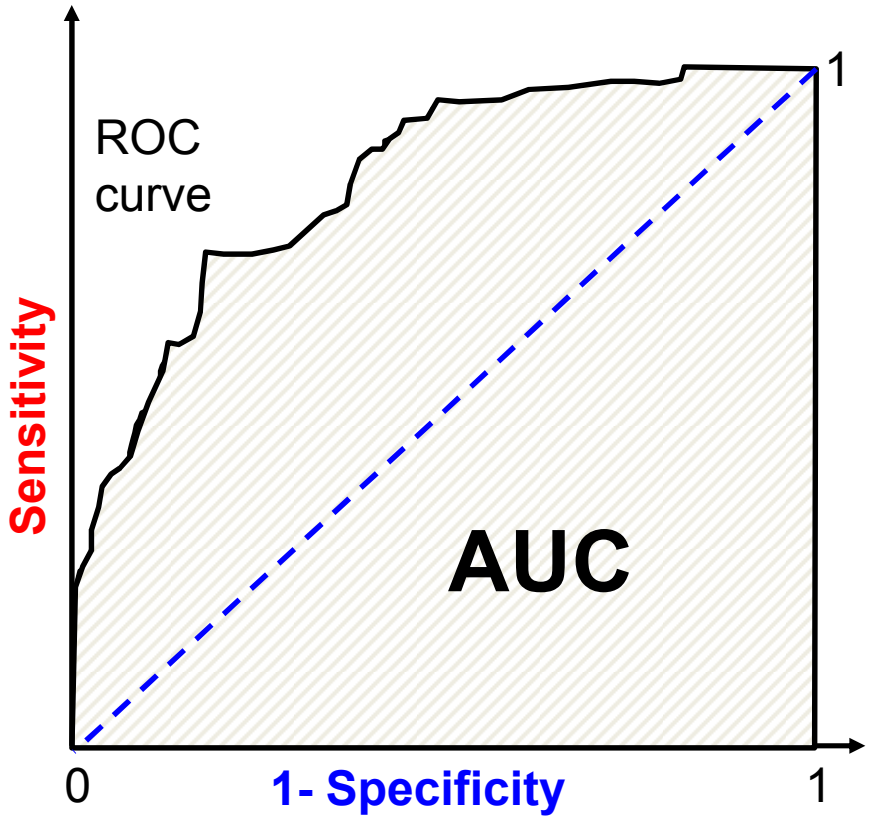
Individual Feature Irrelevance

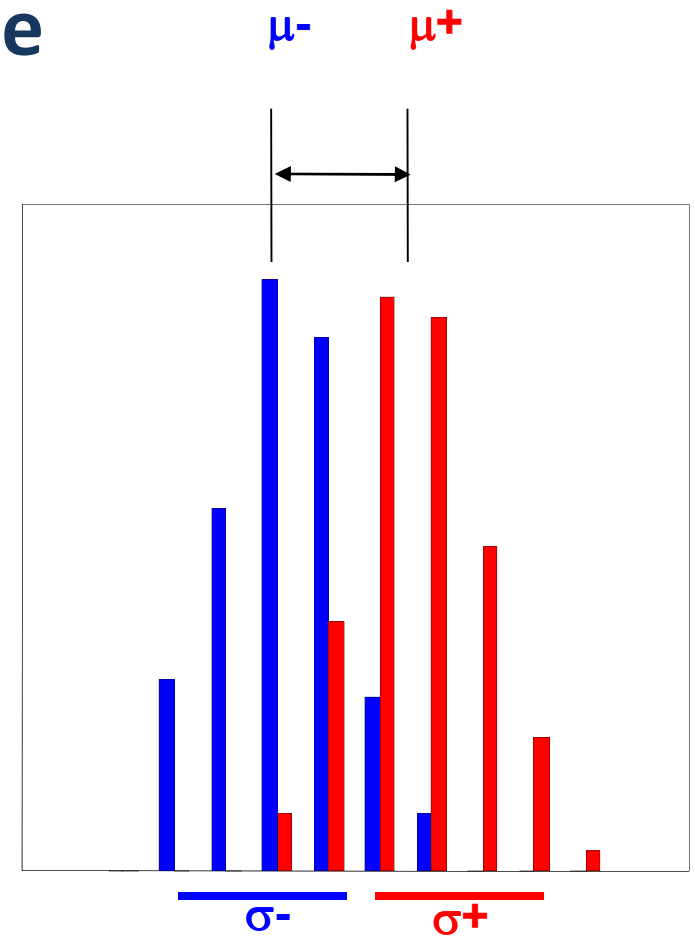


Does this feature tell something about the class variable (red or blue)?

Univariate Filter Methods

Individual Feature Irrelevance





What now?

Univariate Filter Methods

How do we measure dependence?

- Independence: $P(X,Y) = P(X)P(Y)$
- Measures of dependence
  - Mutual Information $MI(X,Y) = \int P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dxdy$

$$= KL\big(P(x,y) \,\|\, P(x)P(y)\big)$$

  - Correlation $R(X,Y) = E\{XY\} = \int xyp(x,y)dxdy$
  - Pearson's correlation

$$NR(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Univariate Filter Methods

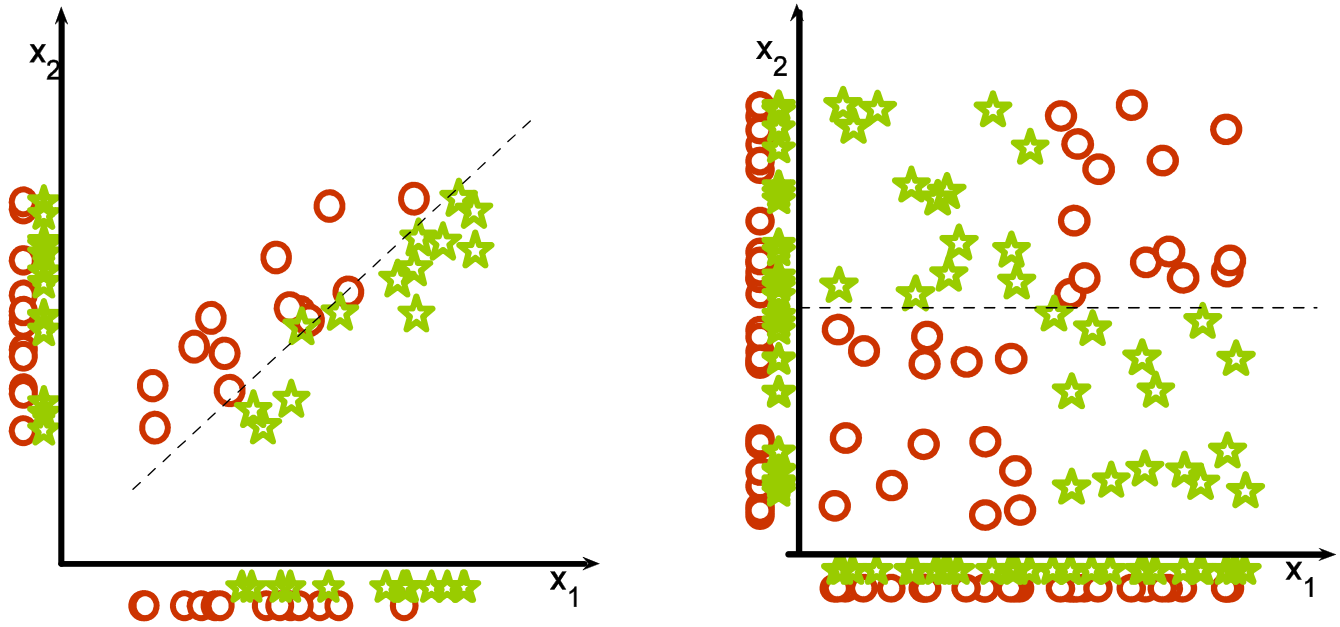
How do we measure dependence?

A lot!

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.

Why Multivariate Methods?

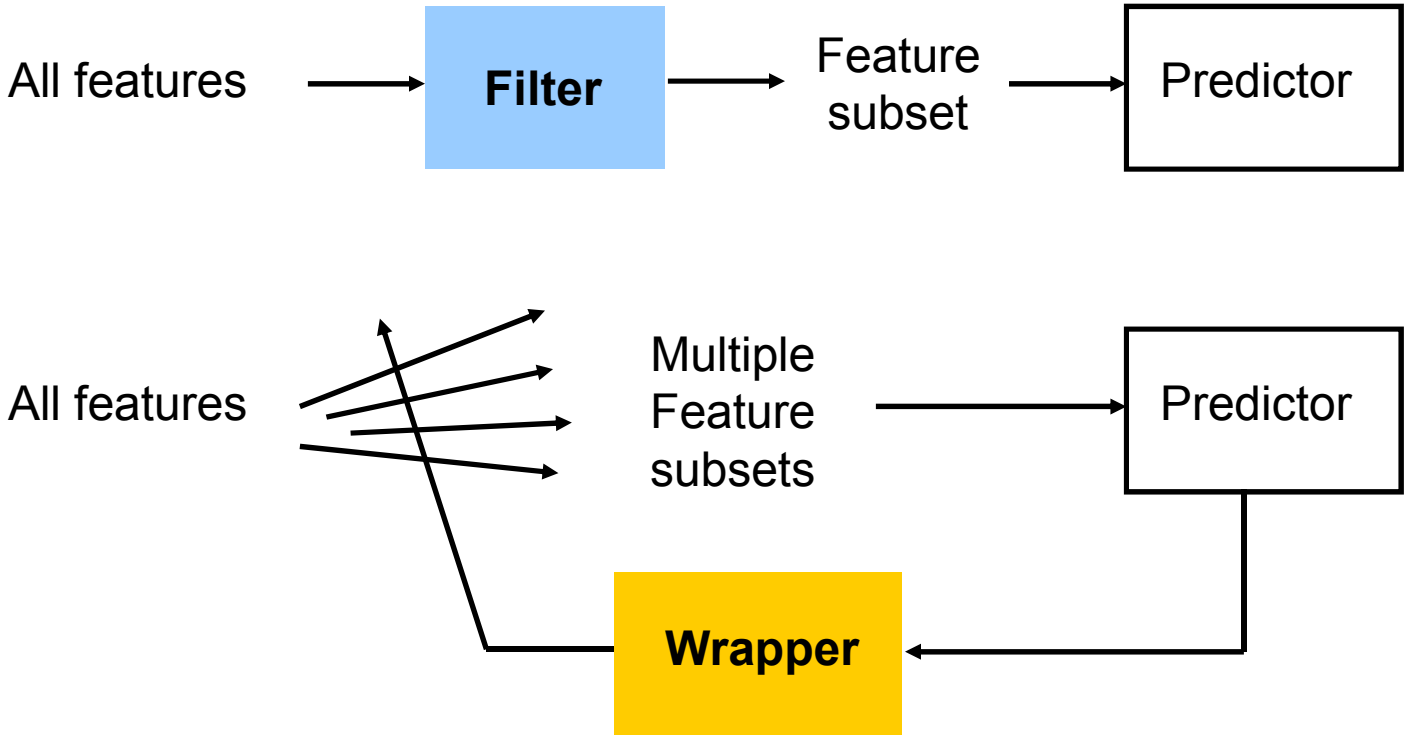
Univariate methods have limitations



Guyon-Elisseeff, JMLR 2004; Springer 2006

Filters vs Wrappers

MAIN GOAL of FS: Rank subsets of useful features



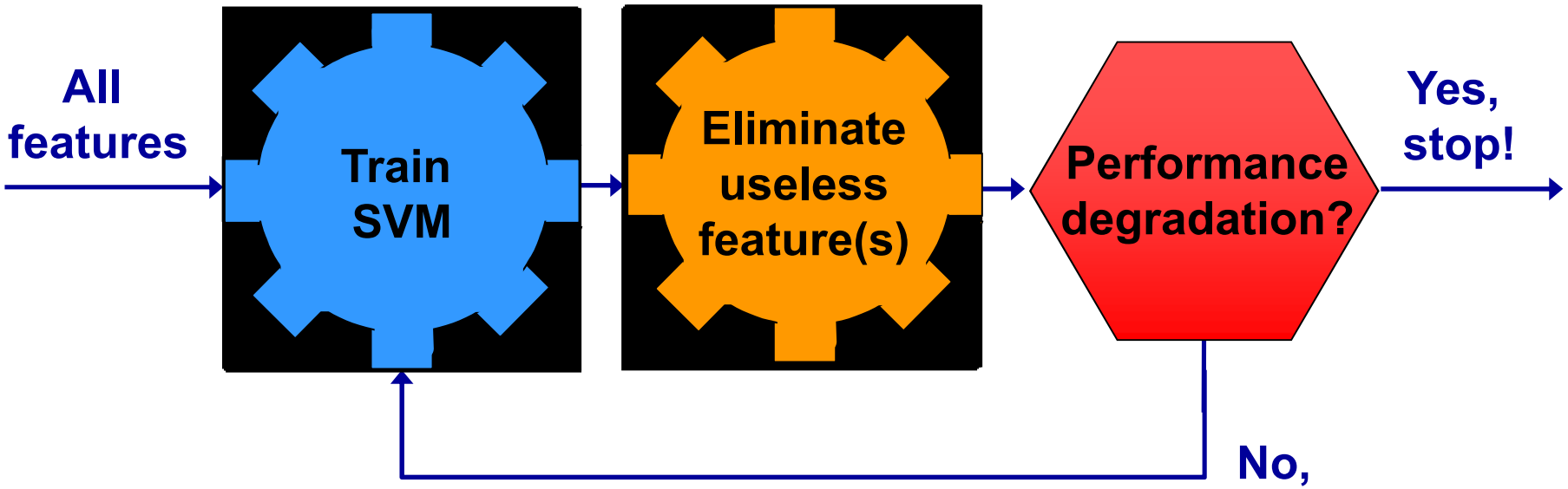
BEWARE! Risk of overfitting with intensive search

Multivariate FS

- Exhaustive search?
  - 2^N subsets with N features = NP-hard problem
- Greedy search alternatives
  - **Forward search** or **backward elimination**
  - **Beam search:** Keep k best path at each step
  - **Floating search (SFFS and SBFS):** One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far. Any time, if a better subset of the same size was already found, switch abruptly.

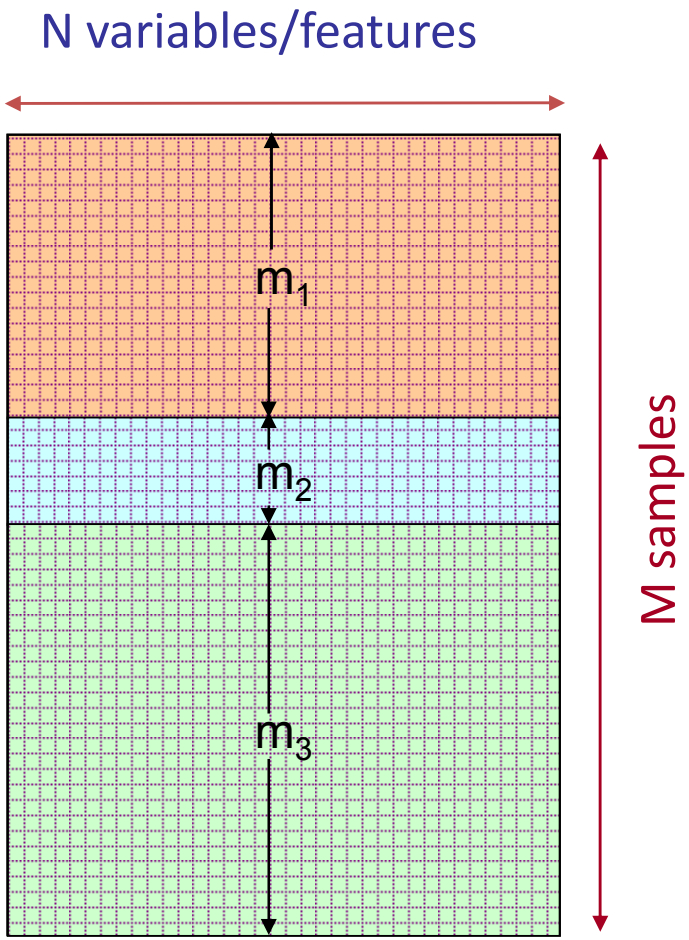
Embedded Methods

- **AdaBoost** – last week!
- SVM-based – **Recursive Feature Elimination**



Recursive Feature Elimination (RFE) SVM.
Guyon-Weston, 2000. US patent 7,117,188

How do we assess feature subsets?



Split data into 3 sets:

training, **validation**, and **test set**.

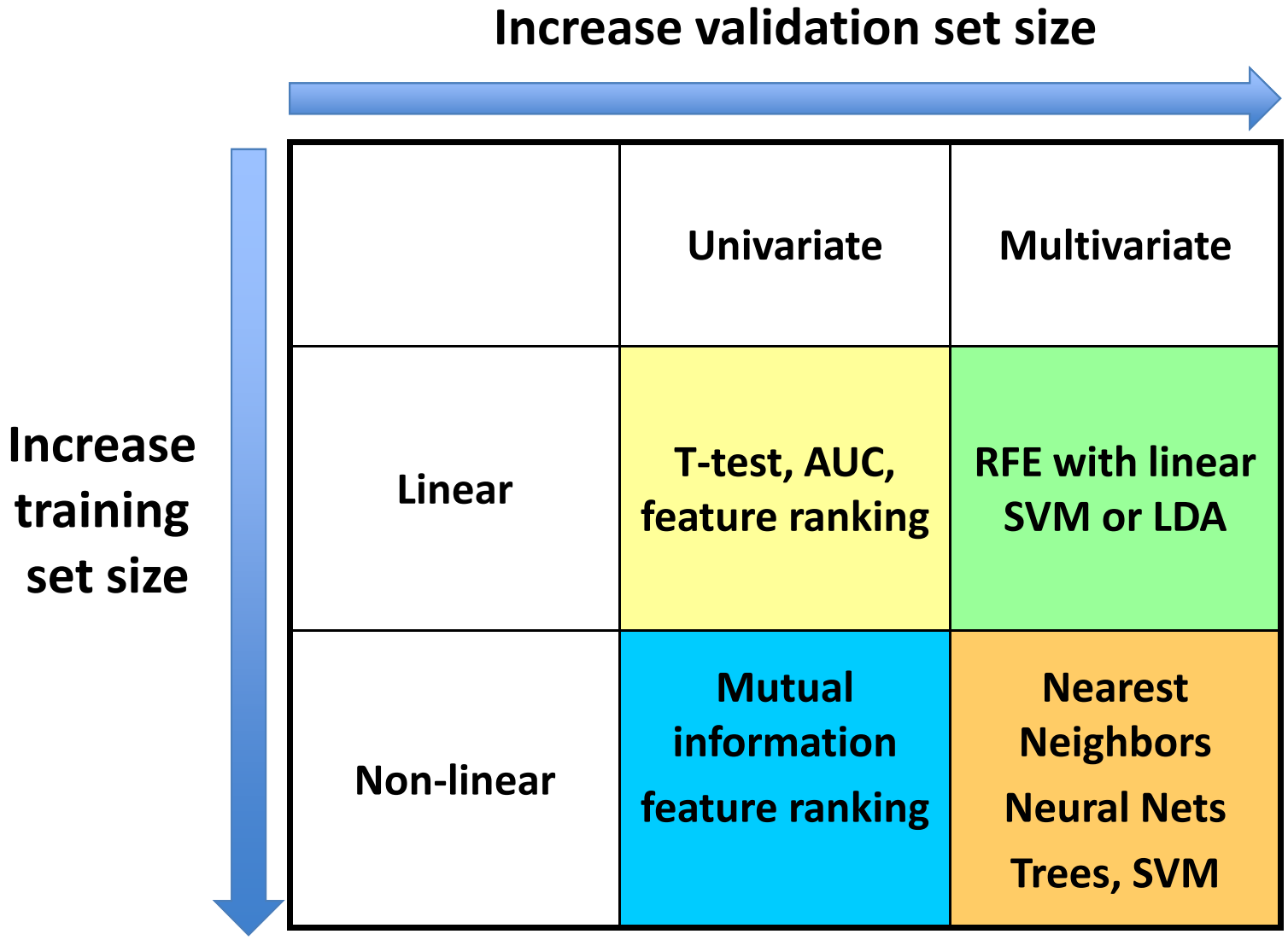
(1) For each feature subset, train predictor on **training data**.

(2) Select the feature subset, which performs best on **validation data**.

- Repeat and average if you want to reduce variance (cross-validation).

(3) Test on **test data**.

Examples of FS Methods



Feature Selection Summary

- **No method is universally better:**
  - wide variety of types of variables, data distributions, learning machines, and objectives.
- **Match the method complexity to the ratio M/N:**
  - univariate feature selection may work better than multivariate feature selection; non-linear classifiers are not always better.
- **Feature selection is not always necessary to achieve good performance.**

The Ten Commandments in FS

(1) Use domain knowledge to construct features

(2) Have your features normalized

(3) Exploit interdependence of features (e.g. augment the subset by product features or via kernels)

(4) Prune your features if necessary

(5) Use individual feature ranking

(6) Eliminate dirty (outlier) instances

(7) Try the simplest approach first: variable ranking then a linear classifier for instance

(8) Increase the gear with more complex methods if necessary

(9) Validate your results

(10) Cross your fingers :P