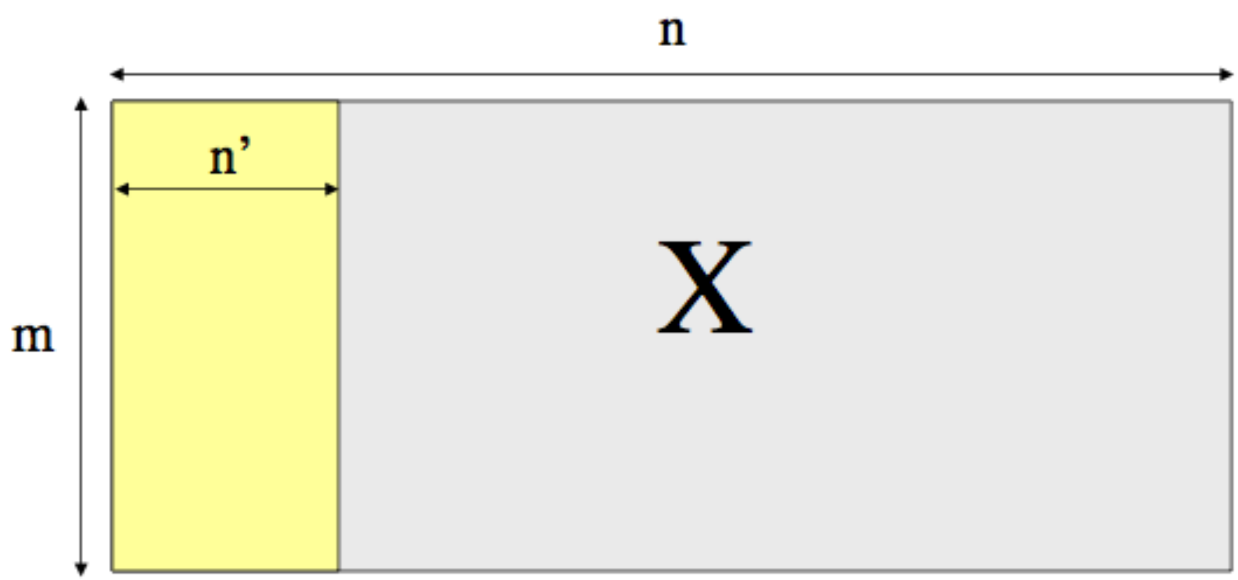
Feature Selection

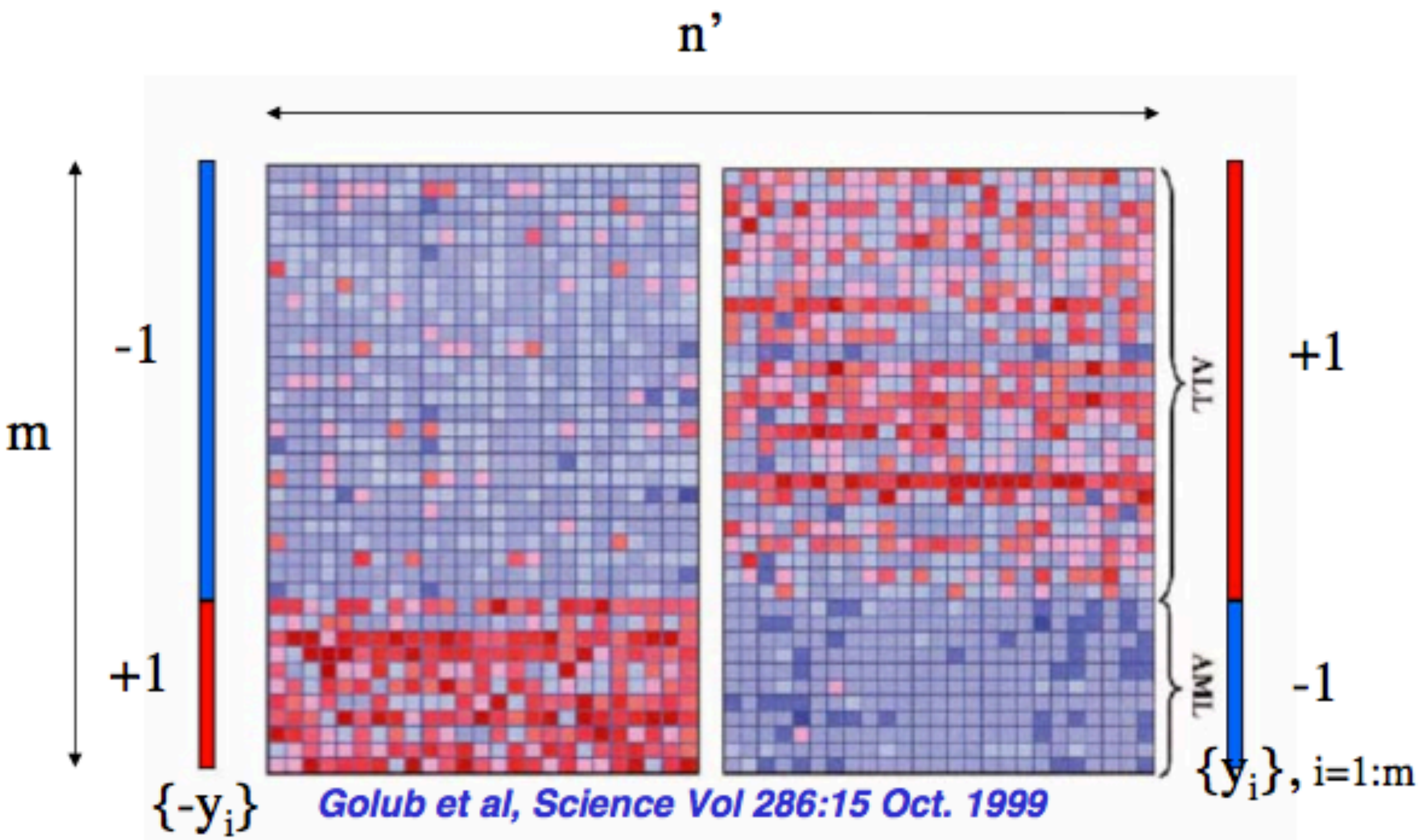
Modified based on "Feature selection and causal discovery – fundamentals and applications" by I. Guyon

Feature Selection

- **Thousands to millions of low level features: select the most relevant ones to build better, faster, and easier to understand learning machines.**



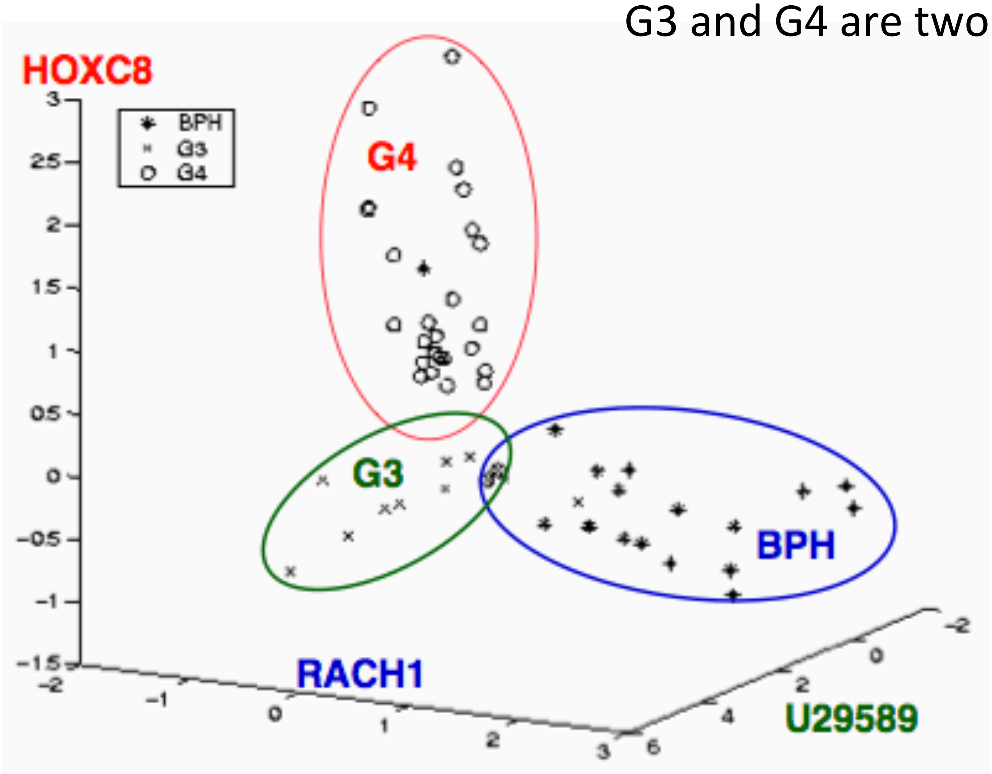
Leukemia Diagnosis



ALL – positive class, Acute lymphoblastic leukemia
AML -- Acute myeloid leukemia

Prostate Cancer Genes

BPH: benign prostate
G3 and G4 are two grades of prostate cancer

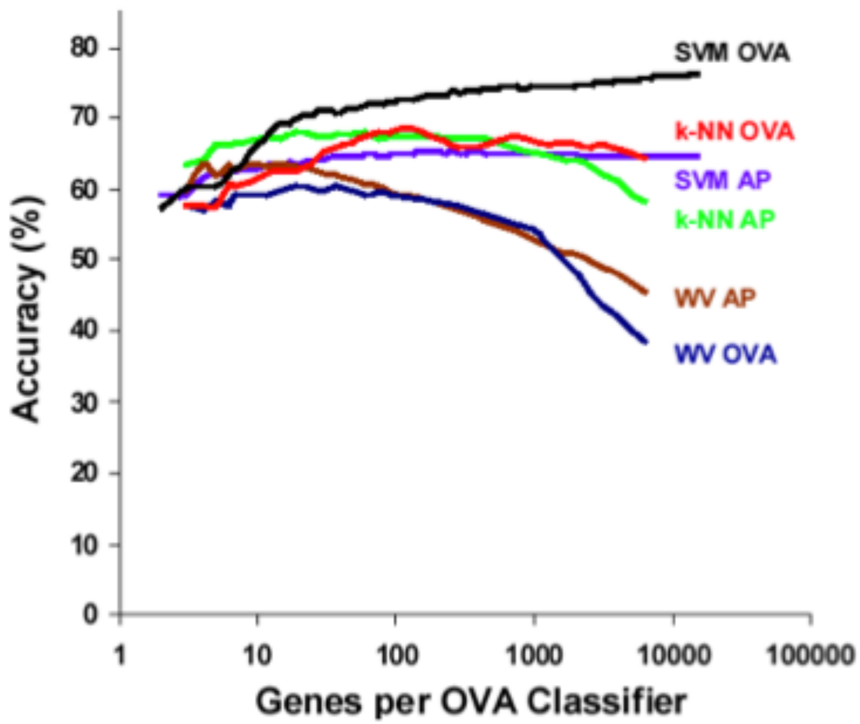


RFE SVM, *Guyon, Weston, et al. 2000.* **US patent 7,117,188**

Application to prostate cancer. *Elisseeff-Weston, 2001*

RFE SVM for Cancer Diagnosis

Differentiation of 14 tumors (Ramaswamy et al., PNAS 2001)



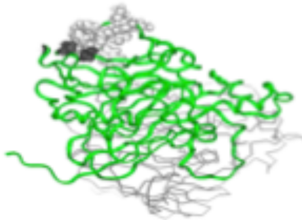
RFE – recursive feature elimination
OVA – One vs. All strategy
AP- advanced P-tree with microarray data
WV – weighted voting
KNN – K nearest neighbor
SVM – support vector machine

Curse of Dimensionality?

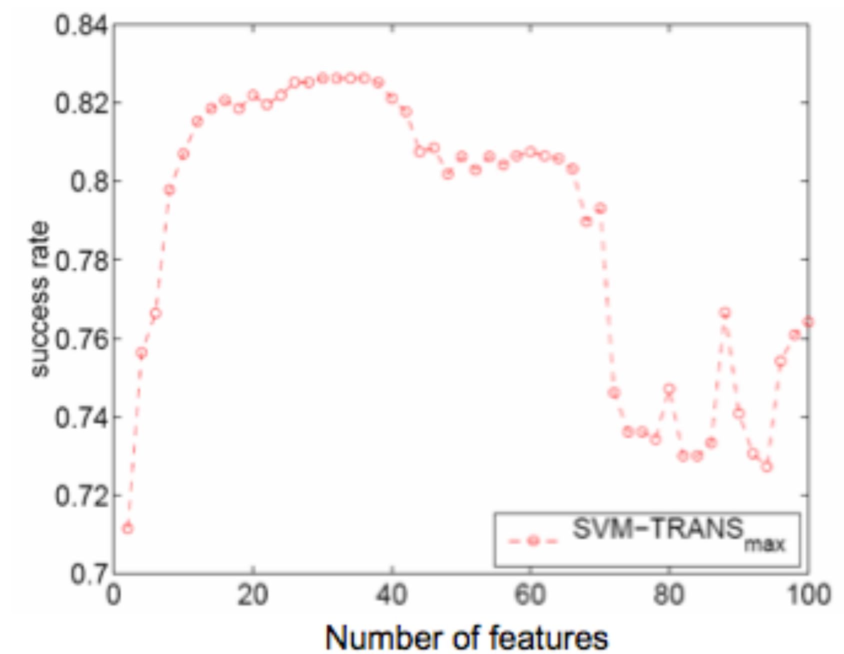
Curse of Dimensionality

- Phenomena arises when analyzing and organizing data in high dimensional spaces that do not occur in low dimensional settings
  - When dimensionality increases, the volume of the space increases so fast that the available data becomes sparse
    - Difficult to have enough data to obtain and support results that are statistically sound and reliable
    - All objects appear to be sparse and dissimilar making it difficult to detect areas where objects form groups of similar properties
  - In machine learning, Hughes phenomenon

Drug Screening

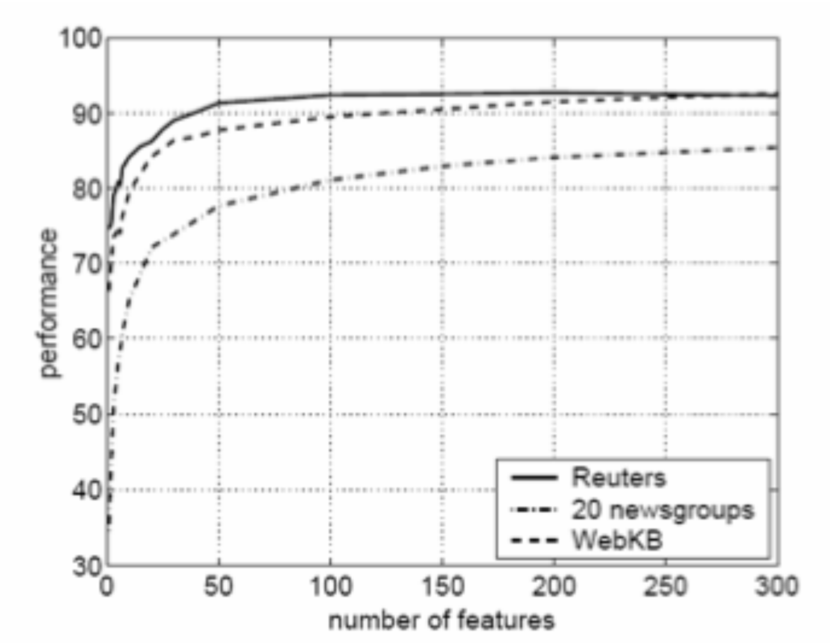


- Binding to Thrombin (DuPont Pharmaceuticals)
  - 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.
  - 139, 351 binary features, which describe three-dimensional properties of the molecule.



Text Filtering

- Reuters: 21578 news wires, 114 semantic categories
- 20 newsgroups: 19997 articles, 20 categories
- WebKB: 8282 web pages, 7 categories
- Bag-of-words: > 100000 features

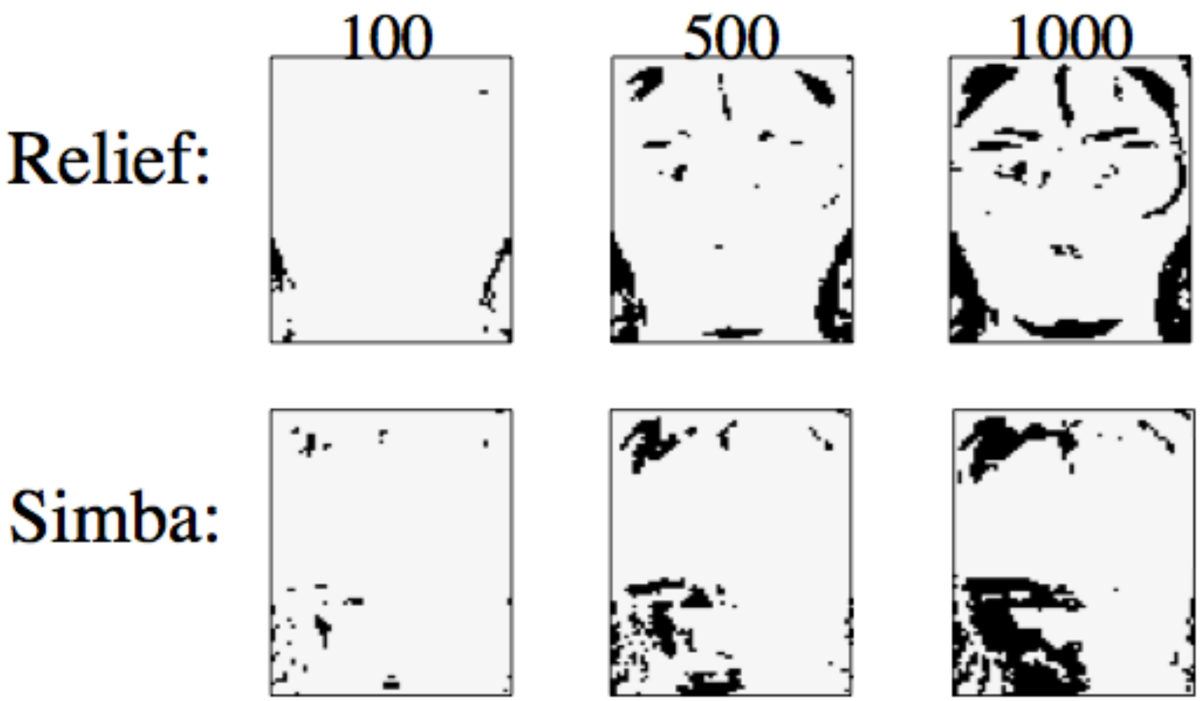


- Top 3 words of some categories:
  - **Alt.atheism: atheism, atheists, morality**
  - **Comp.graphics: image, jpeg, graphics**
  - **Sci.space: space, nasa, orbit**
  - **Soc.religion.christian: god, church, sin**
  - **Talk.politics.mideast: israel, armenian, turkish**
  - **Talk.religion.misc: jesus, god, jehovah**

Bekkerman et al, JMLR, 2003

Face Recognition

- Male/female classification
- 1450 images (1000 train, 450 test), 5100 features (images 60x85 pixels)



Navot-Bachrach-Tishby, ICML 2004

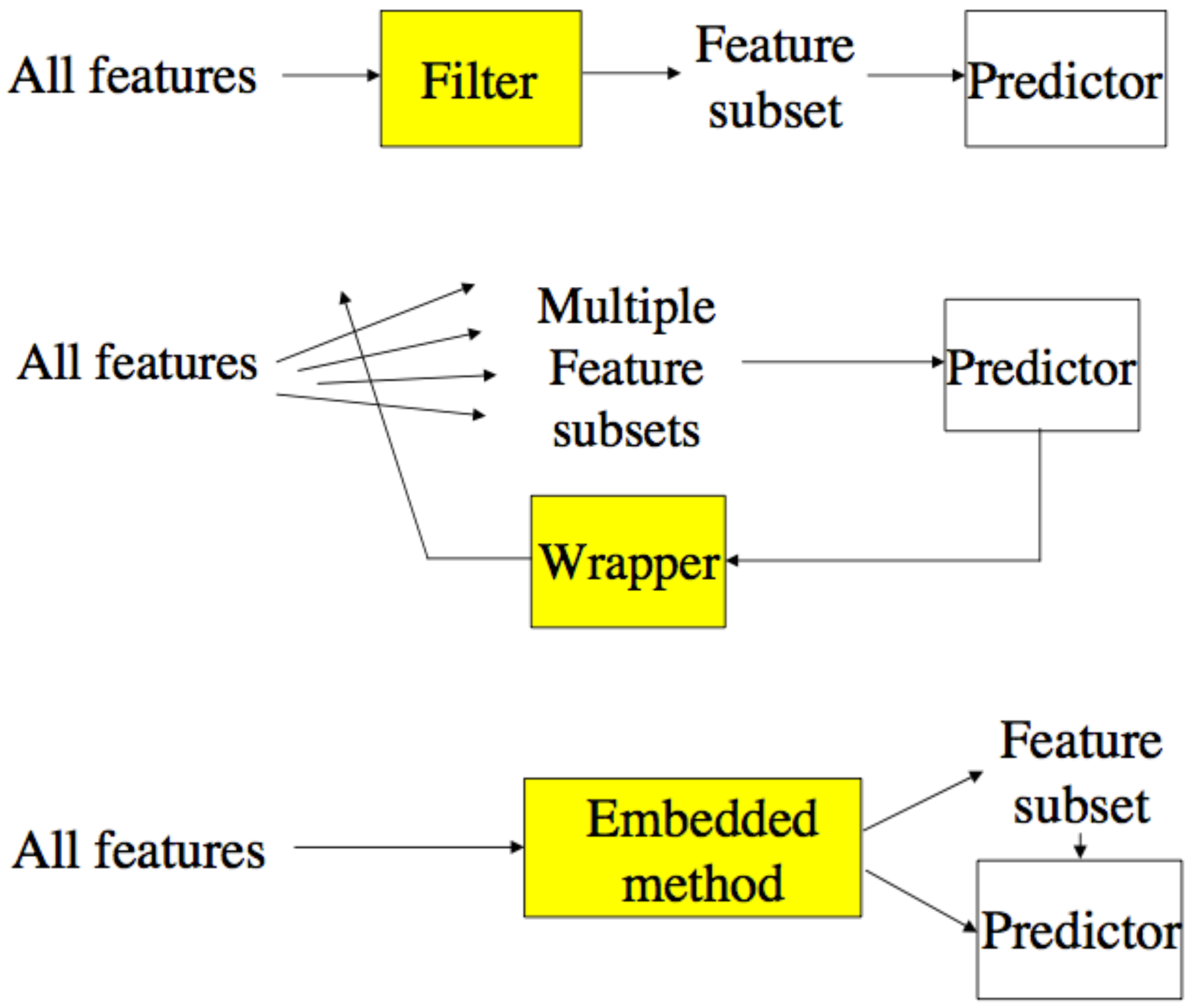
Feature Selection Methods

- **Univariate method**: considers one variable (feature) at a time.
  - Filter methods
- **Multivariate method**: considers subsets of variables (features) together.
  - Filter methods
  - Wrapper method
  - Embedded method

Filters, Wrappers and Embedded methods

- **Filter method: ranks features or feature subsets independently of the predictor (classifier).**
- **Wrapper method: uses a classifier to assess features or feature subsets.**
  - **Validation data set is used to rank features**
  - **More computationally expensive**
  - **Embedded method:** similar to wrapper method except an intrinsic model building metric is used during learning

Filters, Wrappers and Embedded methods



Univariate Filter Methods

- Variance Threshold
- Correlation
- Feature Relevance
- T-test
- Chi-squared test
- Mutual Information

Variance Threshold

- Assuming features with larger variance contain higher information
- Compute the variance of all the features, and eliminate features having variance lower than a predefined threshold value, or
- Keep only the top k features having the largest variance values

Pearson Correlation

- Standardize/normalize data $$x_i = \frac{x_i - \mu}{\sigma}$$

  (μ and σ are the mean and standard deviation of the attribute)

- Compute Pearson correlation between X and Y

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

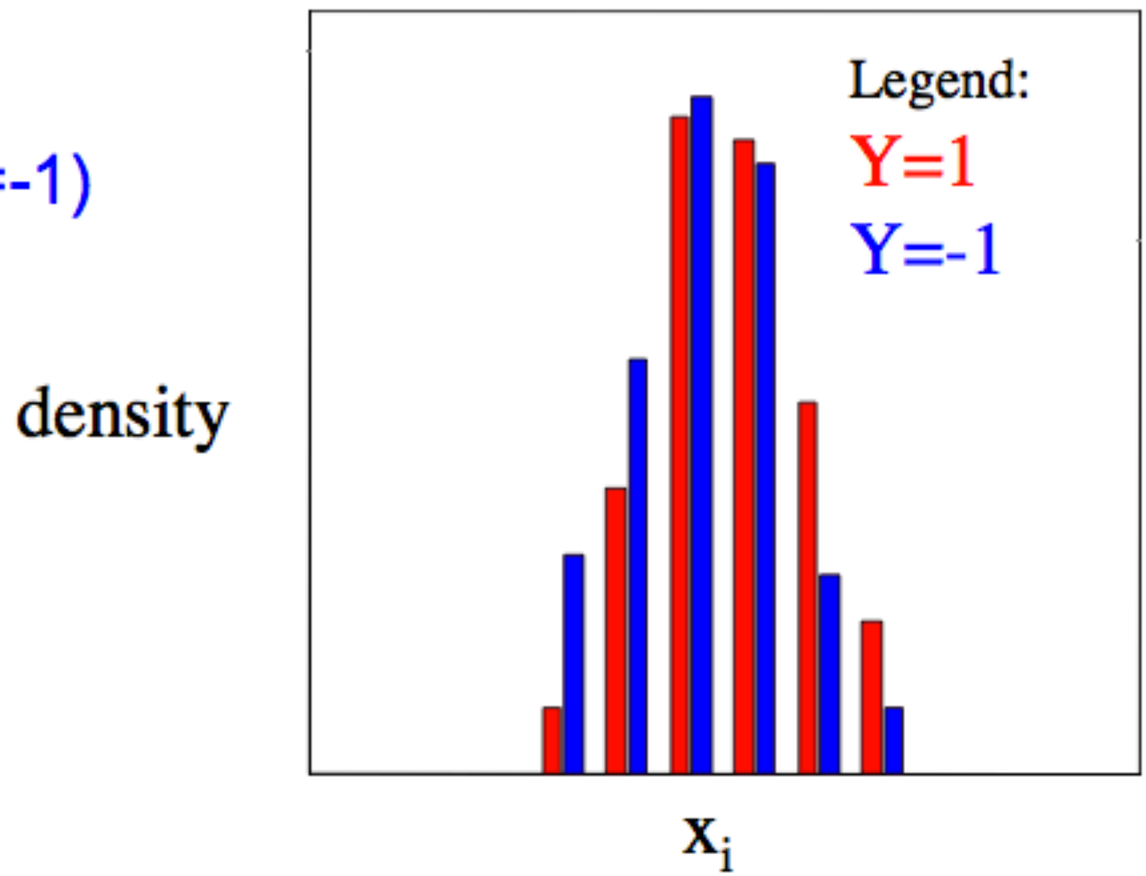
- low correlation to target variable Y → Less predictive features

Individual Feature Irrelevance

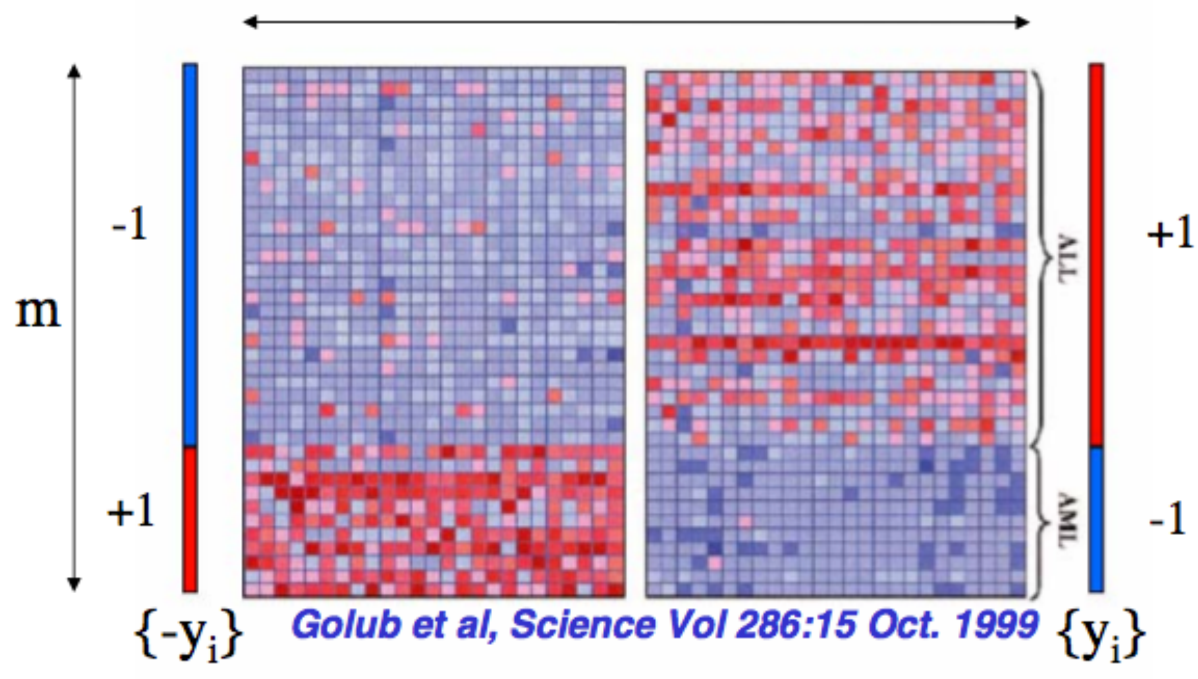
$P(X_i, Y) = P(X_i)\, P(Y)$

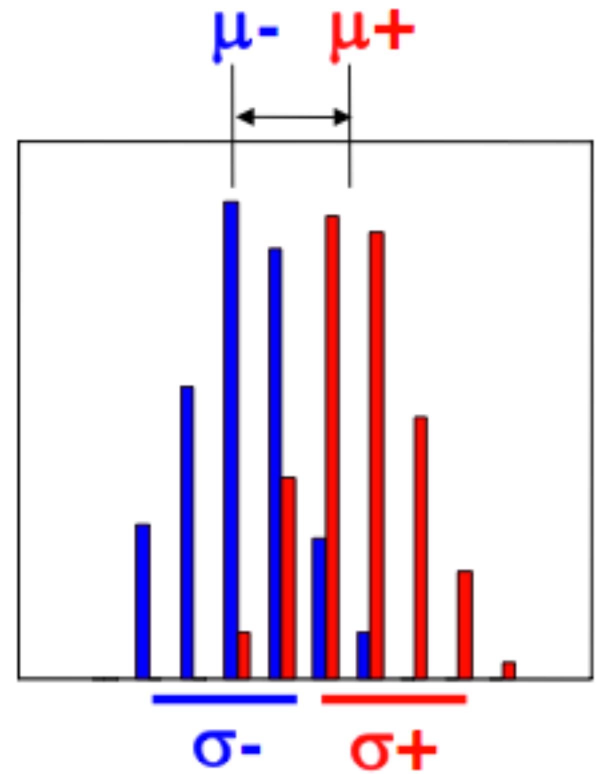
$P(X_i | Y) = P(X_i)$

$P(X_i | Y=1) = P(X_i | Y=-1)$



Signal to Noise Ratio (S_2N)





$$S_2N = \frac{|\mu^+ - \mu^-|}{\sigma^+ + \sigma^-}$$

Higher S_2N → more predictive the feature is to the target variable Y

Practice Question

Obj	A_1	A_2	A_3	…	class
1	30	28	.		Pos
2	24	16	.		Pos
3	20	15	.		Neg
4	28	17	.		Pos
5	10	19	.		Neg
6	20	20	.		Neg
7	16	16	.		Neg
8	34	15	.		Pos
.	.	.	.		.
.	.	.	.		.

Which attribute is better for predicting the class label?

A_1 or A_2 ?

Practice Question (cont.)

Obj	A_1	A_2	Class
1	30	28	positive
2	24	16	
4	28	17	
8	34	15	

$\mu_1 = 29$ $\mu_2 = 19$
$\sigma_1 = 4.16$ $\sigma_2 = 6.05$

Obj	A_1	A_2	Class
3	20	15	negative
5	10	19	
6	20	20	
7	16	16	

$\mu_1 = 16.5$ $\mu_2 = 17.5$
$\sigma_1 = 4.72$ $\sigma_2 = 2.38$

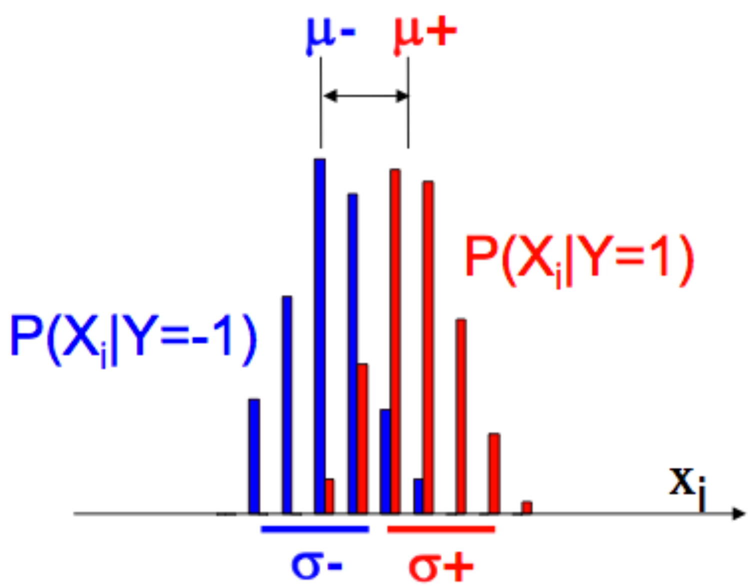
T-test (two tailed)

- Normally distributed classes, equal variance σ^2 unknown; estimated from data as σ^2_{within}.
- Null hypothesis H_0: $\mu^+ = \mu^-$
- T statistic: If H_0 is true,

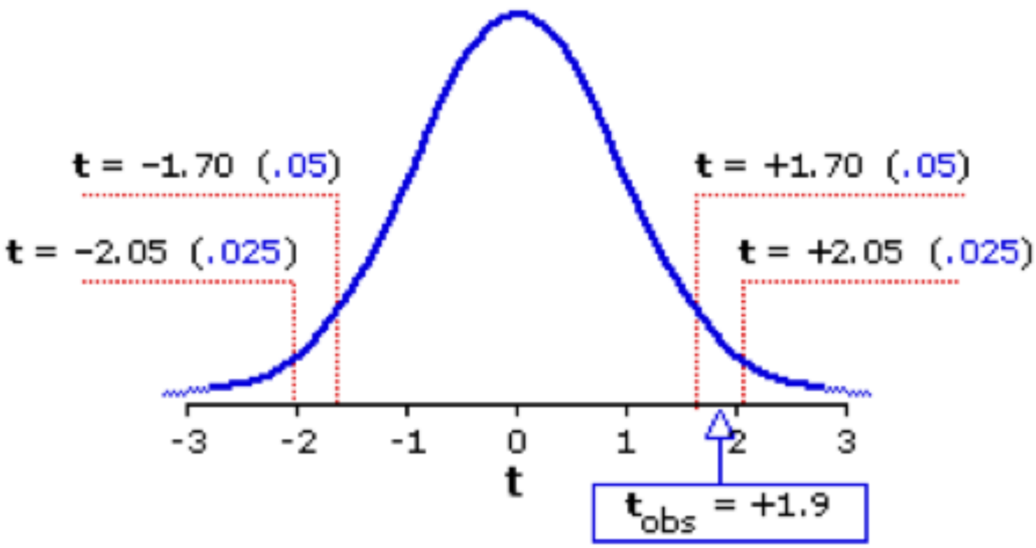
$$t = \frac{(\mu^+ - \mu^-)}{\sigma_{within}\sqrt{\frac{1}{m^+} + \frac{1}{m^-}}},$$

m^+ and m^- are the numbers of rows in Y = 1 and Y = -1 classes

- Degree of freedom: $m^+ + m^- - 2$
- Confidence, e.g., 95%



T-test



	Level of Significance for a Directional Test				
	.05	.025	.01	.005	.0005
	Level of Significance for a Non-Directional Test				
	---	.05	.02	.01	.001
df = 28	1.70	2.05	2.47	2.76	3.67

Procedure

Two t-values:

- t-value calculated from the observations
- Critical t-value

If $t_{obs} > t_{critical}$,
reject the hypothesis
Otherwise,
accept the hypothesis

Null hypothesis H_0: $\mu^+ = \mu^-$

Practice Question

Obj	A_1	A_2	A_3	…	class
1	30	28	.		Pos
2	24	16	.		Pos
3	20	15	.		Neg
4	28	17	.		Pos
5	10	19	.		Neg
6	20	20	.		Neg
7	16	16	.		Neg
8	34	15	.		Pos
.	.	.	.		.
.	.	.	.		.

Are attributes A_1 or A_2 having the same distribution in terms of predicting the class label?

Chi-Squared test for feature selection

- Assume: the samples are a good random sample of the population it represents
- Is "Gender" what you can use to predict an undergrad's preference of his/her footwear?
- Null hypothesis "Gender and Footwear Preference have no relationship"

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

Chi-Squared test for feature selection

Male/Sandals: ((19 X 50)/100) = 9.5
Male/Sneakers: ((22 X 50)/100) = 11
Male/Leather Shoes: ((20 X 50)/100) = 10
Male/Boots: ((25 X 50)/100) = 12.5
Male/Other: ((14 X 50)/100) = 7

Female/Sandals: ((19 X 50)/100) = 9.5
Female/Sneakers: ((22 X 50)/100) = 11
Female/Leather Shoes: ((20 X 50)/100) = 10
Female/Boots: ((25 X 50)/100) = 12.5
Female/Other: ((14 X 50)/100) = 7

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male Observed	6	17	13	9	5	50
Male Expected	9.5	11	10	12.5	7	
Female Observed	13	5	7	16	9	50
Female Expected	9.5	11	10	12.5	7	
Total	19	22	20	25	14	100

Chi-Squared test for feature selection

$$\sum_{i=1}^{rowsize}\sum_{j=1}^{colsize}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Total = 14.026

Male/Sandals: $((6 - 9.5)^2/9.5) = 1.289$

Male/Leather Shoes: $((13 - 10)^2/10) = 0.900$

Male/Other: $((5 - 7)^2/7) = 0.571$

Female/Sneakers: $((5 - 11)^2/11) = 3.273$

Female/Boots: $((16 - 12.5)^2/12.5) = 0.980$

Male/Sneakers: $((17 - 11)^2/11) = 3.273$

Male/Boots: $((9-12.5)^2/12.5) = 0.980$

Female/Sandals: $((13-9.5)^2/9.5) = 1.289$

Female/Leather Shoes: $((7-10)^2/10) = 0.900$

Female/Other: $((9 - 7)^2/7) = 0.571$

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male Observed	6	17	13	9	5	50
Male Expected	9.5	11	10	12.5	7	
Female Observed	13	5	7	16	9	50
Female Expected	9.5	11	10	12.5	7	
Total	19	22	20	25	14	100

Chi-Squared test for feature selection

- What odds are we willing to accept that we are wrong in generalizing from the results in our sample to the population it represents? ➔ confidence 5%
- Degree of Freedom of this problem

  = (# of rows - 1)(# of cols - 1) = (2-1)(5-1)=4
- From Chi Square table of statistics book, with p=0.05, r=4, critical value is 9.49,
  - if Chi square value is less than 9.49, accept the null hypothesis that there is no statistically significant relationship between gender and shoe preference
- In this case, Chi square value is 14.026 > 9.49, so we can reject the null hypothesis and conclude: male and female undergraduates of the Univ. differ in their footwear preferences.

Mutual Information

- Consider feature X and target variable Y:

$$P(x,y) \; = \; \text{joint probability of } (x,y), \quad x \in \{1,\ldots,r\} \text{ and } y \in \{1,\ldots,s\}$$

$$P(x) = \sum_{\mathrm{y}} P(x,y) \; = \; \text{marginal probability of } x$$

$$P(y) = \sum_{\mathrm{x}} P(x,y) \; = \; \text{marginal probability of } y$$

- (In)Dependence often measured by MI

$$0 \leq MI(X,Y) = \sum_{xy} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

  - Also known as *cross-entropy* or *information gain*
  - Selection of relevant variables for the task at hand

Mutual Information

- For feature X and target variable Y, information gain of feature X is

$$IG(Y;X) = I(Y;X) = Entropy(Y) - Entropy(Y \mid X)$$

$$= H(Y) - H(Y \mid X)$$

$$= \sum_{y} -P(y)\log_2 P(y) - \sum_{x} p(x)\Big(\sum_{y} -P(y \mid x)\log_2 P(y \mid x)\Big)$$

$$(algebraic\ \ manipulations)$$

$$= \sum_{xy} P(x,y)\log_2 \frac{P(x,y)}{P(x)p(y)}$$

Mutual Information

- Consider feature X and target variable Y:

$$P(x,y) \;=\; \text{joint probability of } (x,y), \quad x \in \{1,\ldots,r\} \text{ and } y \in \{1,\ldots,s\}$$

$$P(x) = \int_{\mathrm{y}} P(x,y)dy \;=\; \text{marginal probability of } x$$

$$P(y) = \int_{x} P(x,y)dx \;=\; \text{marginal probability of } y$$

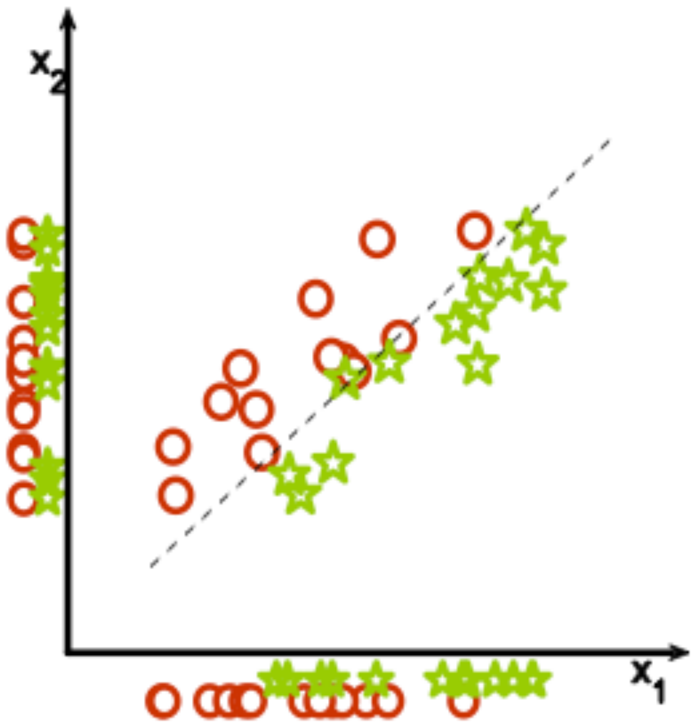
- (In)Dependence often measured by MI

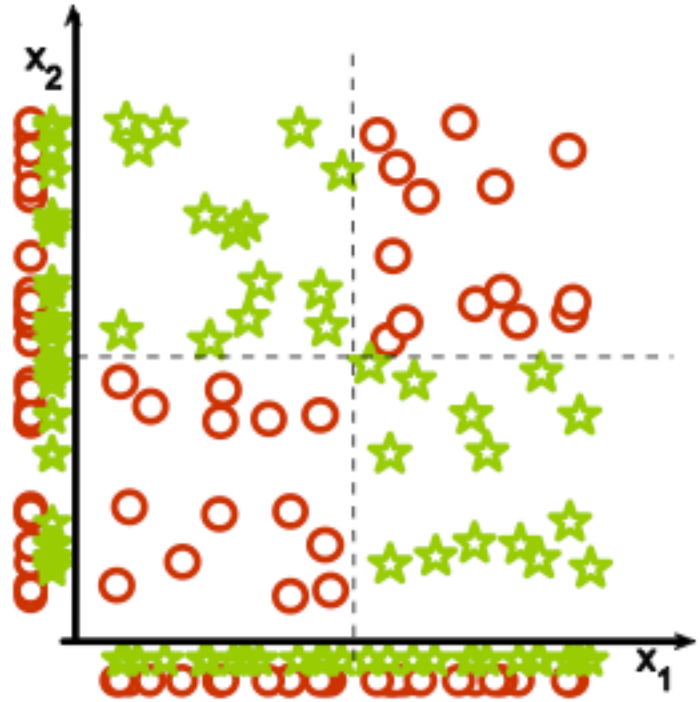
$$0 \le MI(X;Y) = \iint_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}dxdy$$

Multivariate Methods

- Filters
  - Mutual Information
  - Relief
- Wrapper
- Embedded methods

Univariate may fail





Feature Selection with Mutual Information

- Basic idea:
  - Given an initial set F with *n* features, find a subset S of F with *k* features that maximizes the Mutual Information I(C; S), i.e., minimizes H(C|S).
  - Exhaustive search for S is computationally prohibitive.
  - Approximation methods are used instead

MIFS Algorithm

- Basic Idea:
  - Given a set of already selected features, the algorithm chooses the next feature as the one that maximizes the information about the class corrected by subtracting a quantity proportional to the average MI with the selected features.
  - In order to be selected, a feature must be informative about the class without being predictable from the current (chosen) set of features.
    - if two features ***f* and *f'* are highly dependent, *I* (*f* ; *f'*)will be large and, after the better one is picked, the selection of the second one is penalized.**

MIFS Algorithm

Step 1: Initialization: F={initial set of n features}; S={}

Step 2: Computation of the MI with the output class

for each feature *f* compute I(C; *f*).

Step 3: Choice of the first feature

find the feature f that maximizes I(C, *f*): F = F – {*f*}; S = S + {*f*}

Step 4: Greedy Search:

Repeat until |S|=k:

(a) Computation of MI between features

for all pairs of features (*f*, s) where *f* is in F, and s is in S,

compute I(*f*; s)

(b) Selection of the next feature

choose feature f as the one that maximizes $I(C; f) - \beta \sum_{s \in S} I(f; s)$

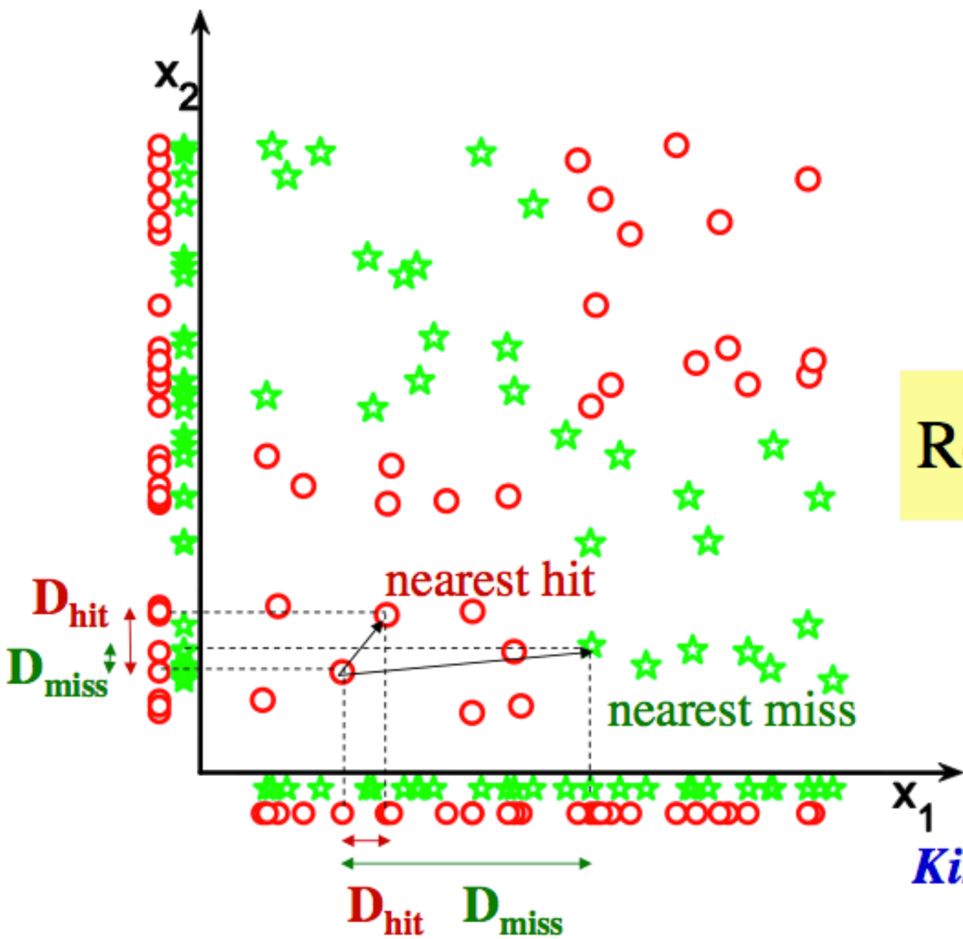
F = F – {*f*}; S = S + {*f*}

Step 5: Output the set S containing the selected features

Relief - Context based feature weighting

- Basic idea:
  - Determine the importance of features in classification based on how feature value changes affect the change of class label
  - A change in feature value accompanied by a change in class → the feature value change could be responsible for the class change → increase weights for this feature
  - A change in feature value leads to no change in class label → this feature has no effect on class → decrease weights for this feature

Relief



$$\text{Relief} = \langle D_{miss} / D_{hit} \rangle$$

Kira and Rendell, 1992

```
Relief(S, m, τ)
    Separate S into S+ = {positive instances} and
        S- = {negative instances}
    W = (0, 0, . . . , 0)
    For i = 1 to m
        Pick at random an instance X ∈ S
        Pick at random one of the positive instances
            closest to X, Z+ ∈ S+
        Pick at random one of the negative instances
            closest to X, Z- ∈ S-
        if (X is a positive instance)
            then    Near-hit = Z+;  Near-miss = Z-
            else    Near-hit = Z-;  Near-miss = Z+
        update-weight(W, X, Near-hit, Near-miss)
    Relevance = (1/m)W
    For i = 1 to p
        if (relevance_i ≥ t)
            then    f_i is a relevant feature
            else    f_i is an irrelevant feature

update-weight(W, X, Near-hit, Near-miss)
    For i = 1 to p
        W_i = W_i - diff(x_i, near-hit_i)^2 + diff(x_i, near-miss_i)^2
```

Relief

- Weight update:

When x_k and y_k are nominal,

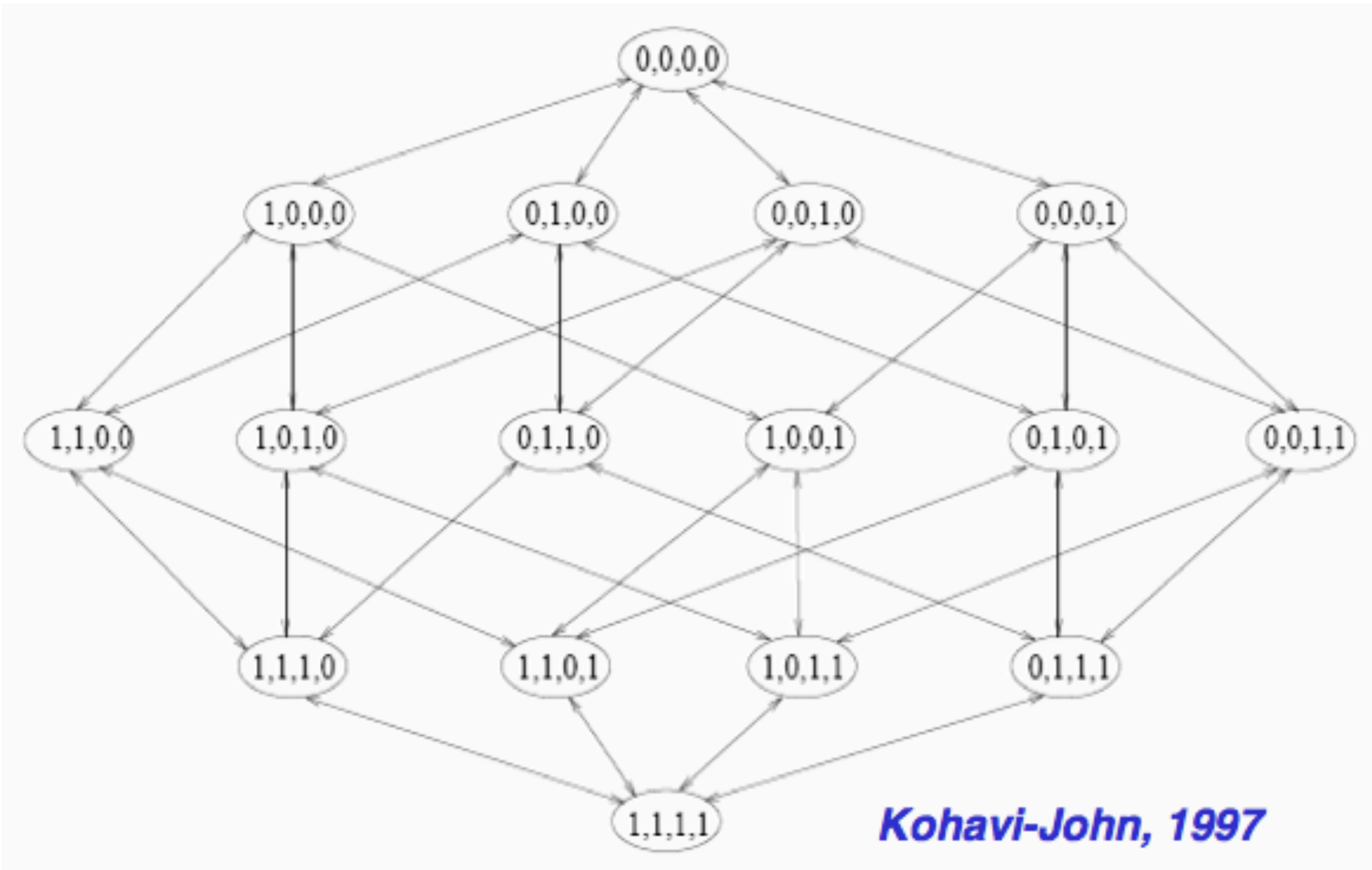
$$\mathrm{diff}(x_k, y_k) = \begin{cases} 0 & \text{<if } x_k \text{ and } y_k \text{ are the same>} \\ 1 & \text{<if } x_k \text{ and } y_k \text{ are different>} \end{cases}$$

When x_k and y_k are numerical,

$$\mathrm{diff}(x_k, y_k) = (x_k - y_k)/nu_k$$

where nu_k is a normalization unit to normalize the values of diff into the interval [0, 1]

Wrapper for feature selection

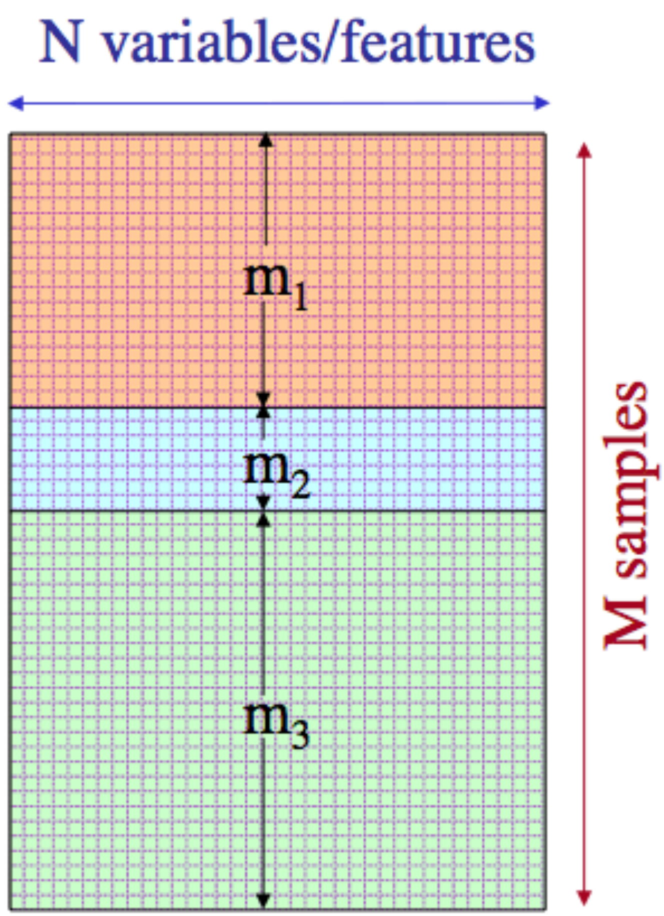


N features, 2^N possible feature subsets!

Search Strategies

- **Exhaustive search**
- **Simulated annealing**
- **Genetic algorithms**
- **Beam search: keep k best path at each step**
- **Greedy search: forward selection or backward elimination.**

Feature subset assessment

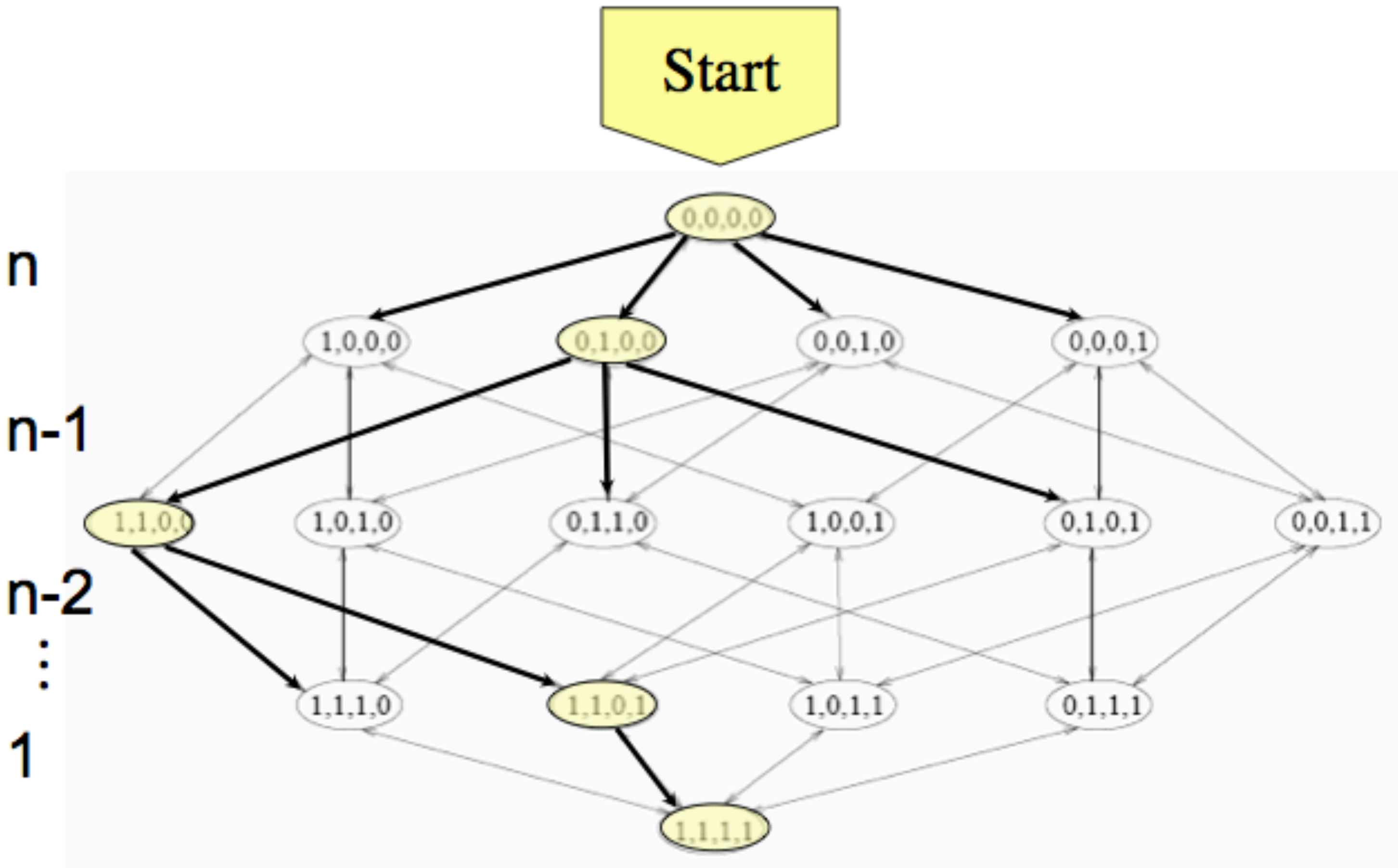


Split data into 3 sets:

training, validation, and test set.

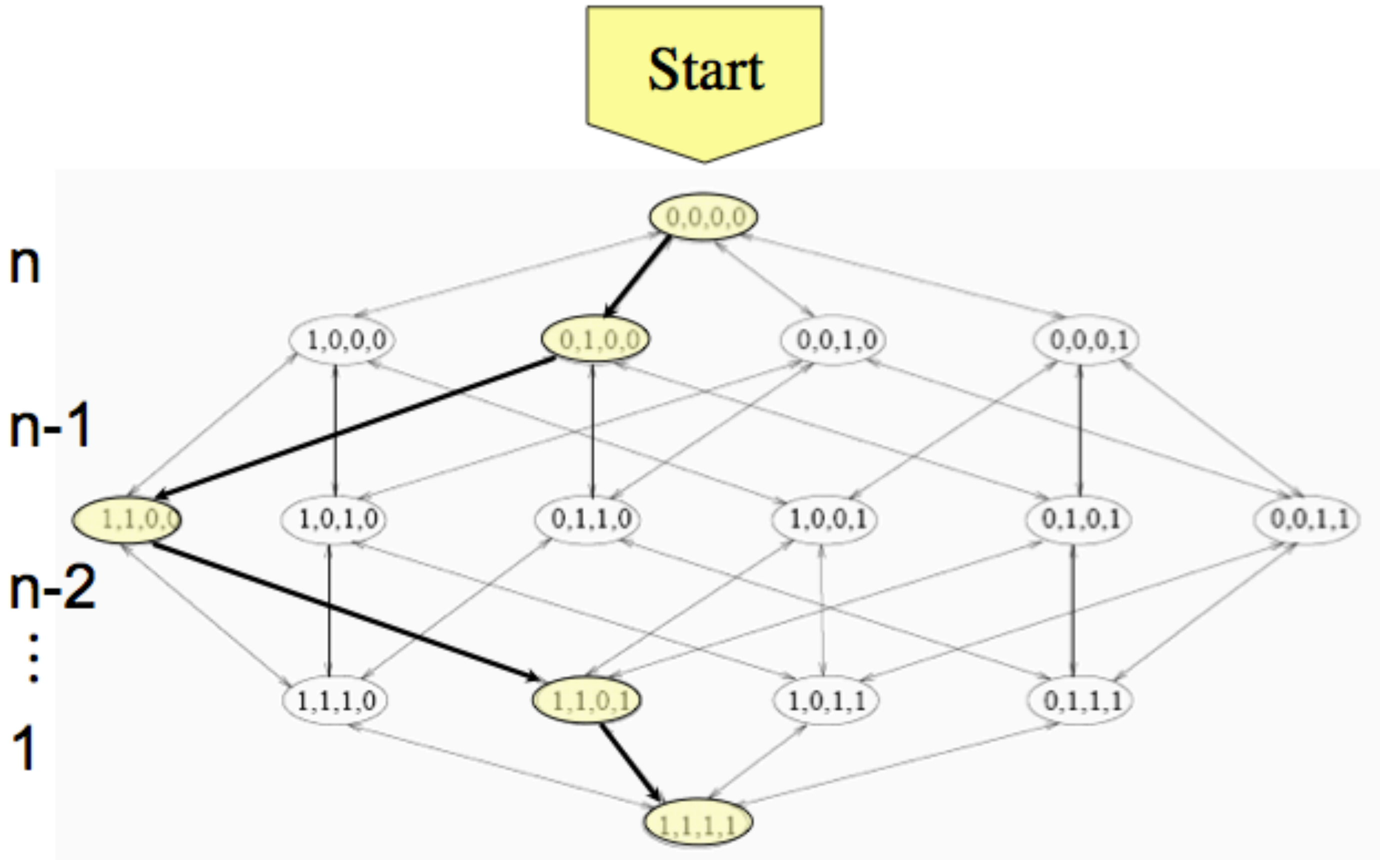
1) For each feature subset, train predictor on training data.
2) Select the feature subset, which performs best on validation data.
   – Repeat and average if you want to reduce variance (cross-validation).
3) Test on test data.

Forward Selection (Wrapper)



Also referred to as SFS: Sequential Forward Selection

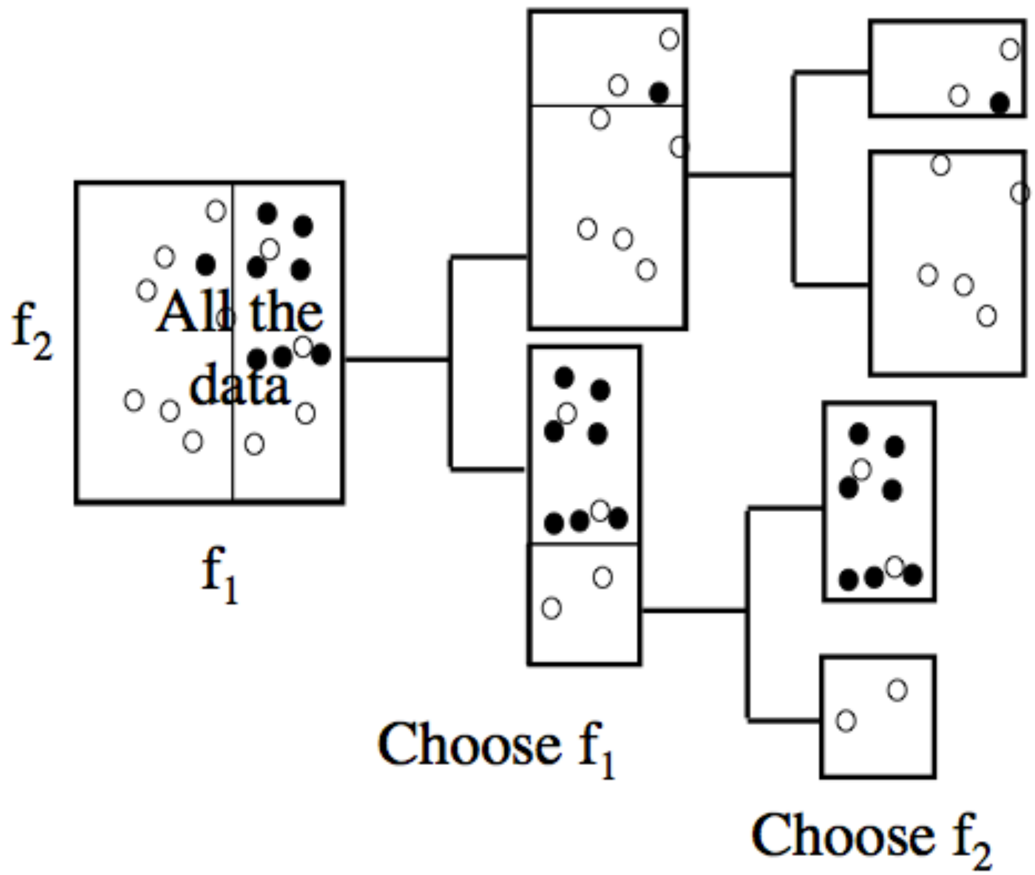
Forward Selection (Embedded)



Guided search: we do not consider alternative paths.

Forward Selection w. Trees

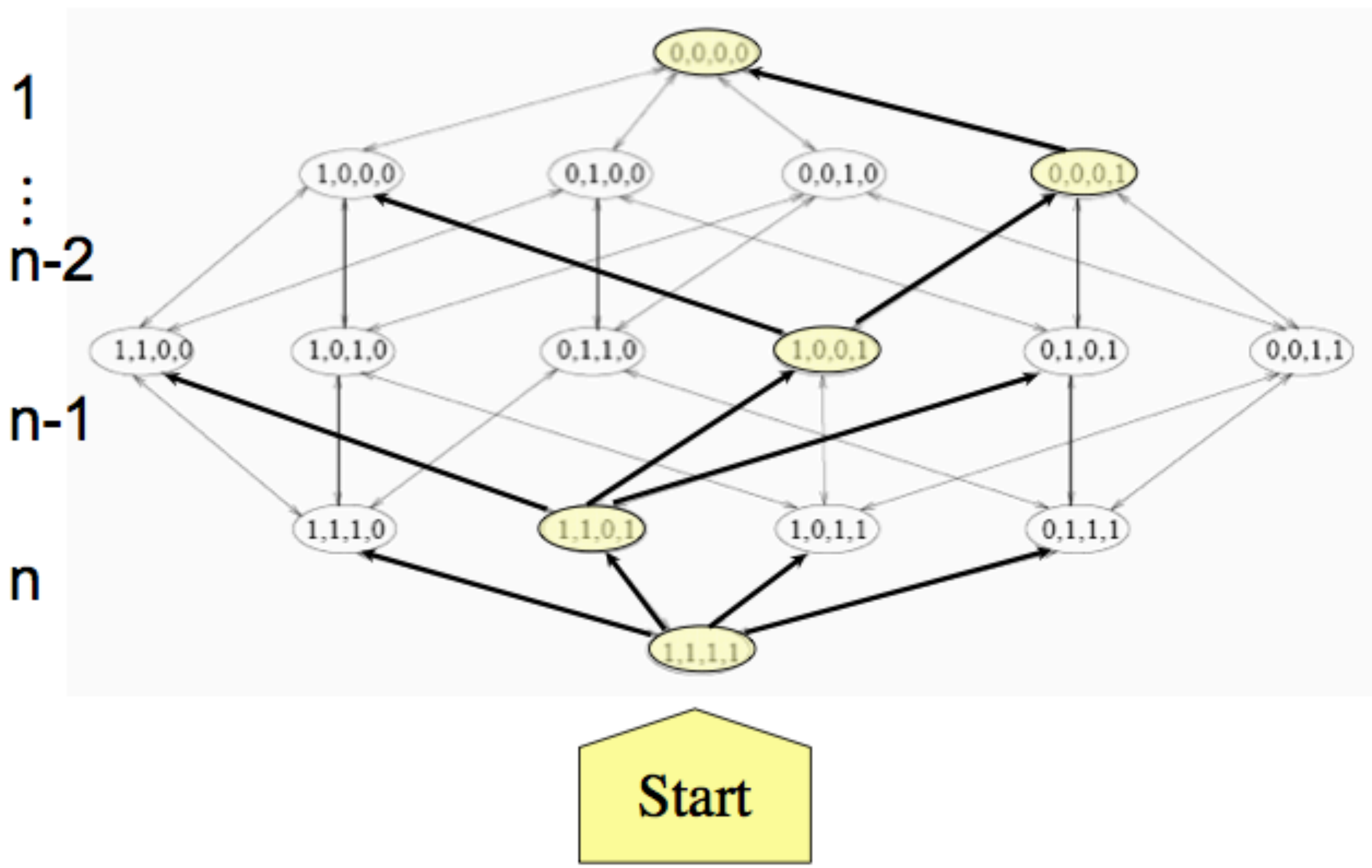
- Tree classifiers,
  like *CART (Breiman, 1984)* or *C4.5 (Quinlan, 1993)*



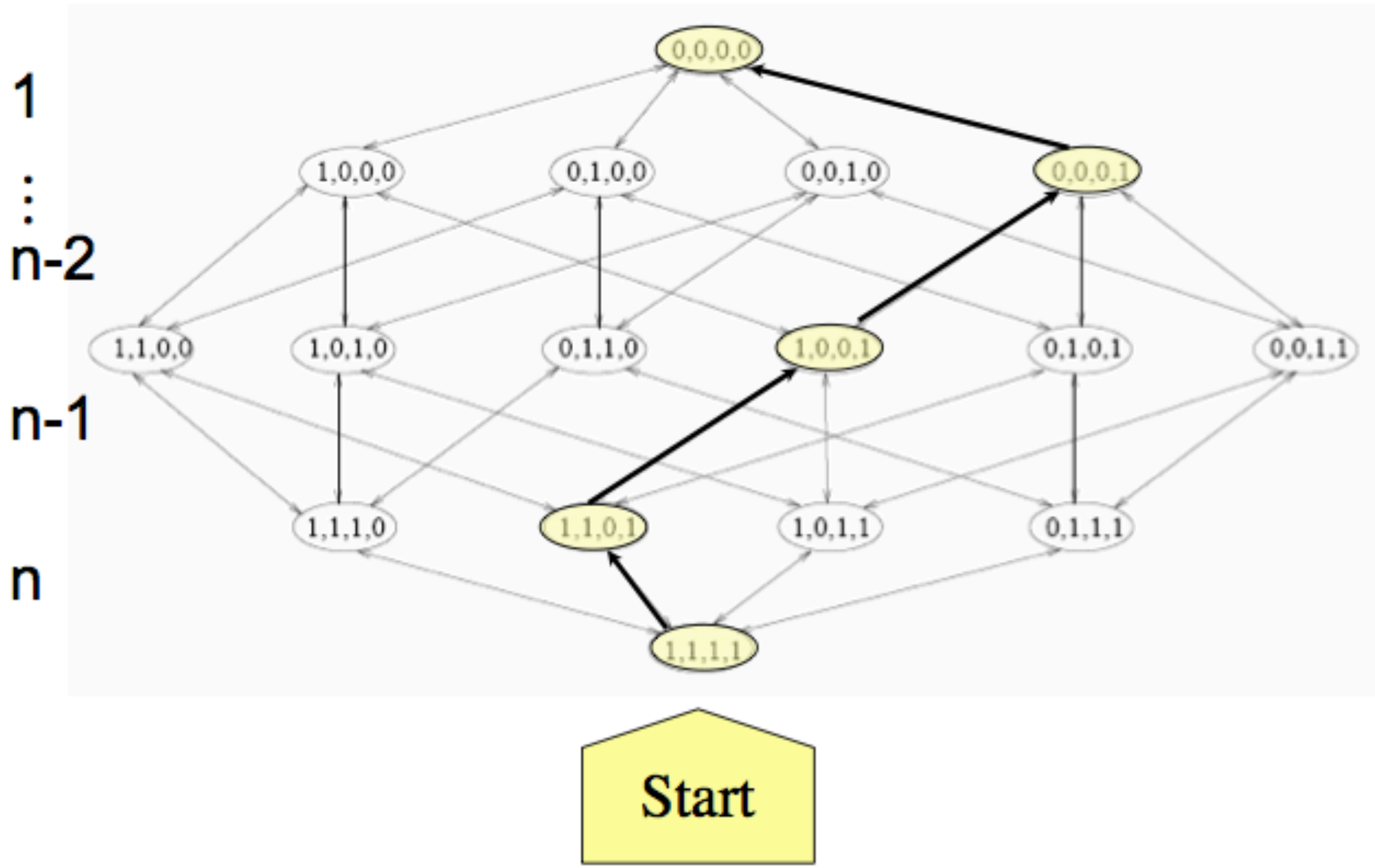
At each step, choose the feature that "reduces entropy" most. Work towards "node purity".

Backward Elimination (Wrapper)

Also referred to as SBS: Sequential Backward Selection



Backward Elimination (embedded)



Conclusions

- Feature selection focuses on uncovering subsets of variables X_1, X_2, ... predictive of the target Y.
- Multivariate feature selection is in principle more powerful than univariate feature selection, but not always in practice.
- Taking a closer look at the type of dependencies in terms of causal relationships may help refining the notion of variable relevance.