Feature Selection

Goals

- What is Feature Selection for classification?
- Why feature selection is important?
- What is the filter and what is the wrapper approach to feature selection?
- Examples

What is Feature Selection for classification?

- Given: a set of predictors ("features") *V* and a target variable *T*
- Find: minimum set *F* that achieves maximum classification performance of *T (*for a given set of classifiers and classification performance metrics)

Why feature selection is important?

- May Improve performance of classification algorithm
- Classification algorithm may not scale up to the size of the full feature set either in sample or time
- Allows us to better understand the domain
- Cheaper to collect a reduced set of predictors
- Safer to collect a reduced set of predictors

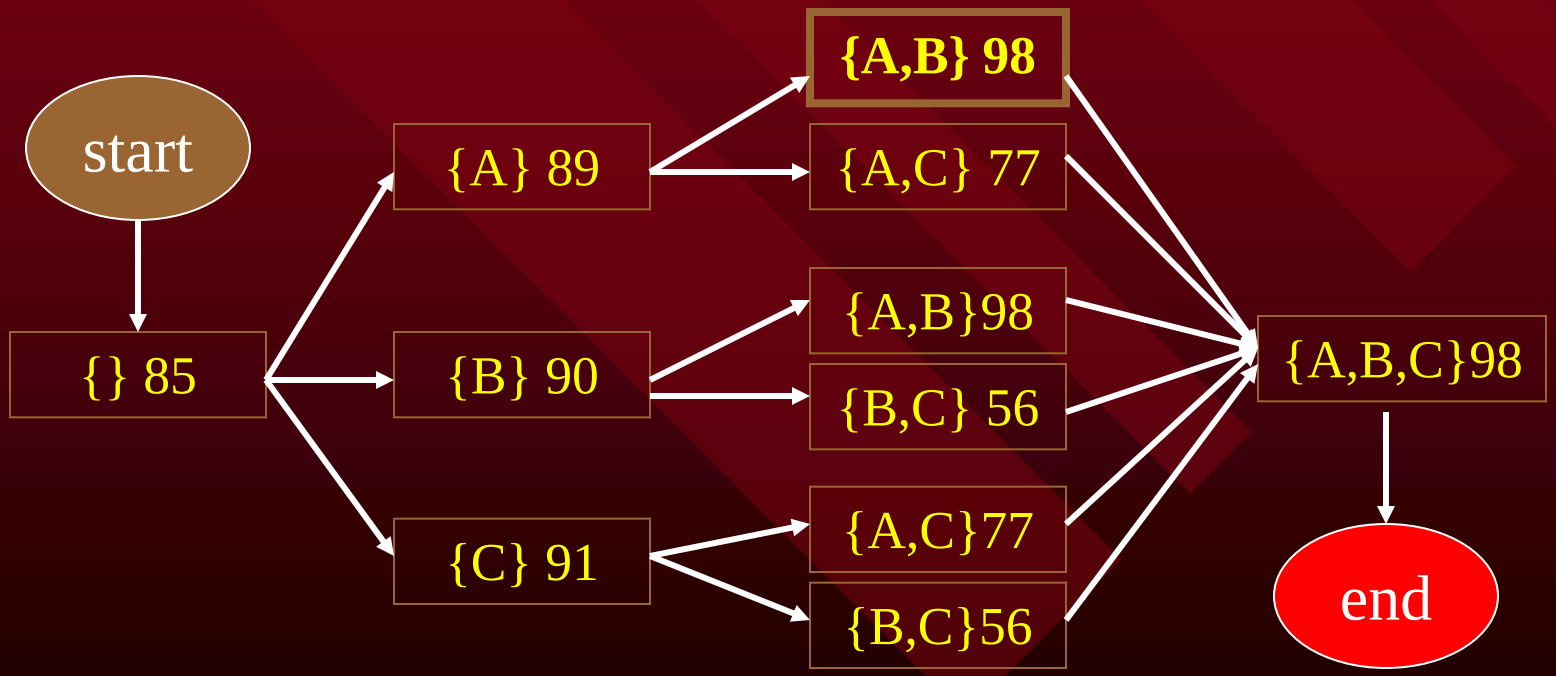
Filters vs Wrappers: Wrappers

Say we have predictors A, B, C and classifier *M*. We want to predict T given the smallest possible subset of {A,B,C}, while achieving maximal performance (accuracy)

FEATURE SET	CLASSIFIER	PERFORMANCE
{A,B,C}	*M*	**98%**
{A,B}	*M*	**98%**
{A,C}	*M*	77%
{B,C}	*M*	56%
{A}	*M*	89%
{B}	*M*	90%
{C}	*M*	91%
{.}	*M*	85%

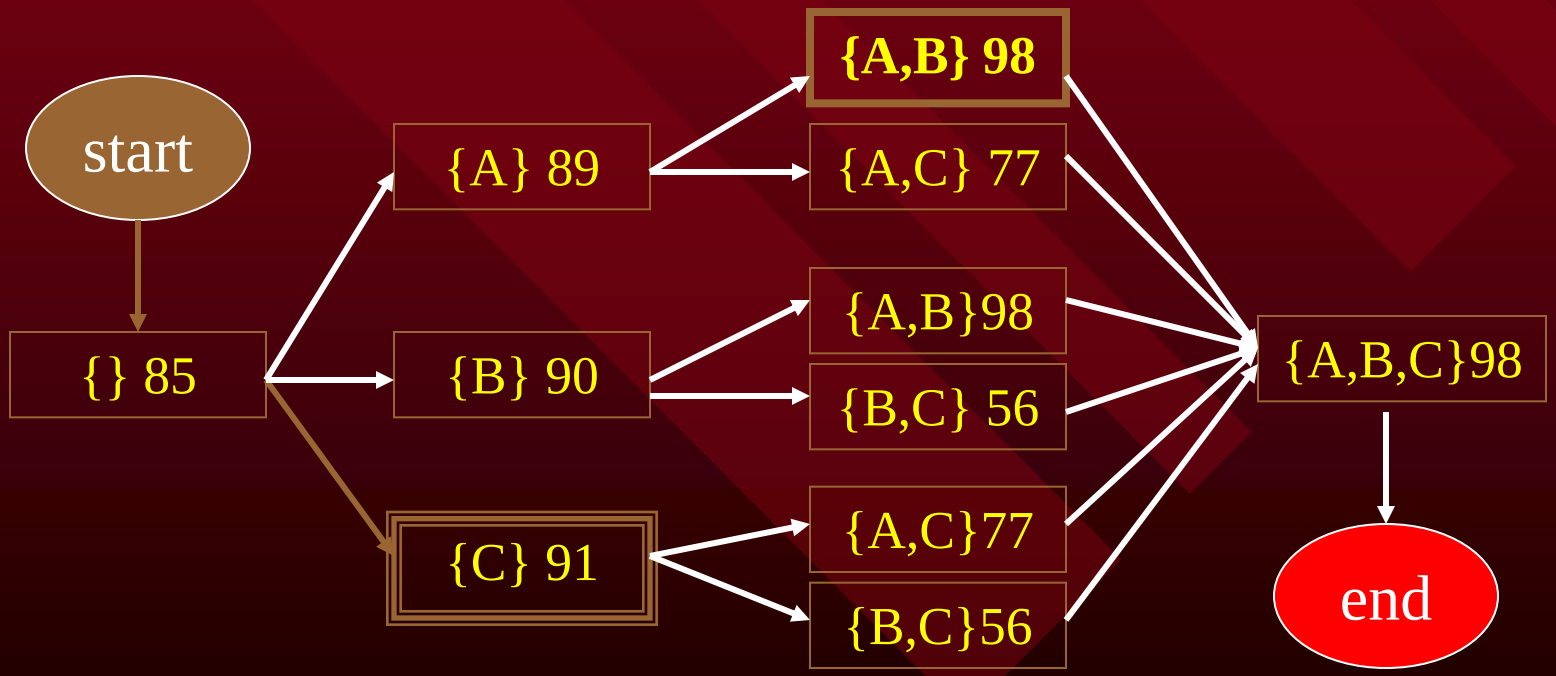
Filters vs Wrappers: Wrappers

The set of all subsets is the power set and its size is $2^{|V|}$.
Hence for large V we cannot do this procedure exhaustively; instead we rely on *heuristic search* of the space of all possible feature subsets.



Filters vs Wrappers: Wrappers

A common example of heuristic search is hill climbing: keep adding features one at a time until no further improvement can be achieved.



Filters vs Wrappers: Wrappers

A common example of heuristic search is hill climbing: keep adding features one at a time until no further improvement can be achieved ("forward greedy wrapping")

Alternatively we can start with the full set of predictors and keep removing features one at a time until no further improvement can be achieved ("backward greedy wrapping")

A third alternative is to interleave the two phases (adding and removing) either in forward or backward wrapping ("forward-backward wrapping").

Of course other forms of search can be used; most notably:

-Exhaustive search

-Genetic Algorithms

-Branch-and-Bound (e.g., cost=# of features, goal is to reach performance *th* or better)

Filters vs Wrappers: Filters

In the filter approach we do not rely on running a particular classifier and searching in the space of feature subsets; instead we select features on the basis of statistical properties. A classic example is univariate associations:

FEATURE	ASSOCIATION WITH TARGET	
{A}	91%	Threshold gives suboptimal solution
{B}	90%	Threshold gives optimal solution
{C}	89%	Threshold gives suboptimal solution

Example Feature Selection Methods in Biomedicine: Univariate Association Filtering

- Order all predictors according to strength of association with target
- Choose the first *k* predictors and feed them to the classifier
- Various measures of association may be used: X^2, G^2, Pearson r, Fisher Criterion Scoring, etc.
- How to choose *k*?
- What if we have too many variables?

Example Feature Selection Methods in Biomedicine: Recursive Feature Elimination

– Filter algorithm where feature selection is done as follows:

1. build linear Support Vector Machine classifiers using *V* features
2. compute weights of all features and choose the best *V/2*
3. repeat until 1 feature is left
4. choose the feature subset that gives the best performance (using cross-validation)

Example Feature Selection Methods in Bioinformatics: GA/KNN

Wrapper approach whereby:

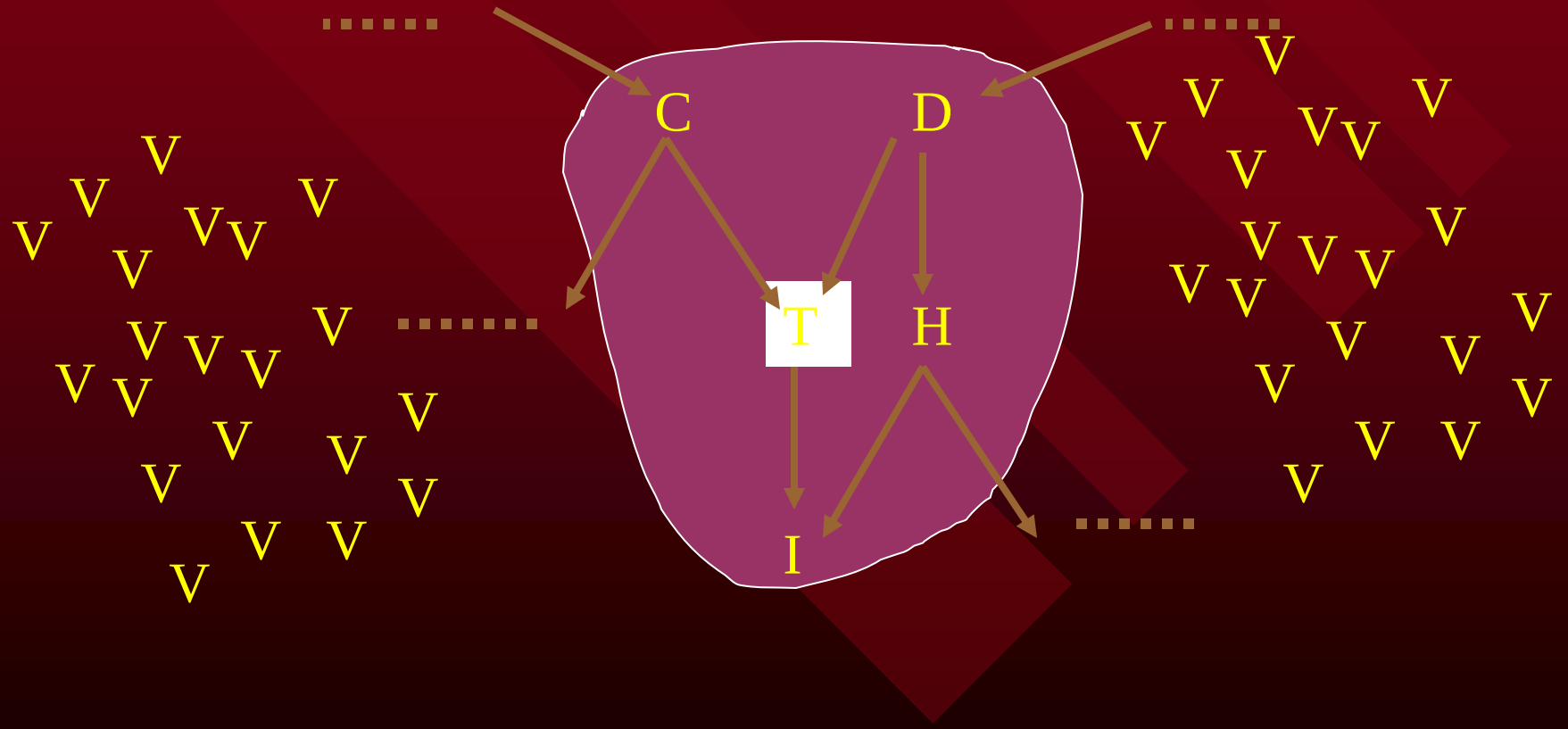
- heuristic search=Genetic Algorithm, and
- classifier=KNN

How do we approach the feature selection problem in our research?

- Find the Markov Blanket
- Why?

A fundamental property of the Markov Blanket

- MB(T) is the minimal set of predictor variables needed for classification (diagnosis, prognosis, etc.) of the target variable **T** **(given a powerful enough classifier and calibrated classification)**



HITON: An algorithm for feature selection that combines MB induction with wrapping

C.F. Aliferis M.D., Ph.D., I. Tsamardinos Ph.D., A. Statnikov M.S.

Department of Biomedical Informatics, Vanderbilt University

AMIA Fall Conference, November 2003

HITON: An algorithm for feature selection that combines MB induction with wrapping

ALGORITHM	SOUND	SCALABLE	SAMPLE EXPONENTIAL TO |MB|	COMMENTS
Cheng and Greiner	YES	NO	NO	Post-processing on learning BN
Cooper et al.	NO	NO	NO	Uses full BN learning
Margaritis and Thrun	YES	YES	YES	Intended to facilitate BN learning
Koller and Sahami	NO	NO	NO	Most widely-cited MB induction algorithm
Tsamardinos and Aiferis	YES	YES	YES	Some use BN learning as sub-routine
HITON	YES	YES	NO	

HITON: An algorithm for feature selection that combines MB induction with wrapping

- Step #1: Find the parents and children of *T;* call this set **PC(*T*)**
- Step #2: Find the **PC(.)** set of each member of **PC(*T*);** take the union of all these sets to be **PCunion**
- Step #3: Run a special test to filter out from **PCunion** the non-members of MB(*T*) that can be identified as such (not all can); call the resultant set TMB (tentative MB)
- Step #4: Apply heuristic search with a desired classifier/loss function and cross-validation to identify variables that can be dropped from TMB without loss of accuracy

HITON (Data *D*; Target *T*; Classifier *A*)
"returns a minimal set of variables required for optimal classification of *T* using algorithm *A*"
MB(T) = HITON-MB(D, T) // Identify Markov Blanket
Vars = Wrapper(*MB(T), T, A*) // Use wrapping to remove unnecessary variables
Return Vars

HITON-MB(Data D, Target T)
"returns the Markov Blanket of *T*"
PC = parents and children of T returned by HITON-PC(D, T)
PCPC = parents and children of the parents and children or T
CurrentMB = $PC \cup PCPC$
// Retain only parents of common children and remove false positives
$\forall$ potential spouse X in *CurrentMB* and $\forall$ *Y in* PC*:*
if not$\exists$ S in $\{Y\} \cup \boldsymbol{V} - \{T, X\}$ so that $\perp (T ; X \mid S)$
then retain *X* in *CurrentMB*
else remove it
Return CurrentMB

HITON-PC(Data D, Target T)
"returns parents and children of *T*"

Wrapper(Vars, T, A)
"returns a minimal set among variables *Vars* for predicting *T* using algorithm *A* and a wrapping approach"
Select and remove a variable.
If internally cross-validated performance of *A* remains the same permanently remove the variable.
Continue until all variables are considered.

HITON-PC(Data D, Target T)

"returns parents and children of *T*"

CurrentPC = {}

Repeat

Find variable V_i *not* in *CurrentPC* that maximizes *association(Vi, T)* and admit V_i into *CurrentPC*

If there is a variable *X* and a subset S of CurrentPC s.t. $\perp(X : T \mid S)$

remove V_i *from CurrentPC;*

mark V_i *and do not consider it again*

Until *no more variables are left to consider*

Return *CurrentPC*

Dataset	Thrombin	Arrythmia	Ohsumed	Lung Cancer	Prostate Cancer
Problem Type	Drug Discovery	Clinical Diagnosis	Text Categorization	Gene Expression Diagnosis	Mass-Spec Diagnosis
Variable #	139,351	279	14,373	12,600	779
Variable Types	binary	nominal/ordinal /continuous	binary and continuous	continuous	continuous
Target	binary	nominal	binary	binary	binary
Sample	2,543	417	2000	160	326
Vars-to-Sample	54.8	0.67	7.2	60	2.4
Evaluation metric	ROC AUC	Accuracy	ROC AUC	ROC AUC	ROC AUC
Design	1-fold c.v.	10-fold c.v.	1-fold c.v.	5-fold c.v.	10-fold c.v.

Figure 2: Dataset Characteristics

1. Drug Discovery (Thrombin)				
	UAF*	RFE	HITON	ALL
SVM	96.12%	93.29%	93.23%	93.69%
KNN	87.25%	89.71%	92.23%	88.21%
NN	*NA*	92.04%	92.65%	*NA*
Average	91.69%	91.68%	**92.7%**	90.95%
# of variables	34837	8709	**32**	139351
2. Clinical Diagnosis (Arrythmia)				
	UAF*	B/F*	HITON*	ALL*
DTI	73.94%	72.85%	71.87%	73.94%
KNN	63.22%	63.45%	65.30%	63.22%
NN	58.29%	60.90%	60.38%	58.29%
Average	65.15%	65.73%	**65.85%**	65.15%
# of variables	279	96	**63**	279
3. Text Categorization (OHSUMED)				
	IG	X^2	HITON	ALL*
SVM	82.43%	85.91%	82.85%	90.50%
SBCtc	84.18%	86.23%	85.10%	84.25%
KNN	75.55%	81.76%	80.25%	77.56%
NN	82.47%	85.27%	83.97%	*NA*
Average	81.16%	**84.79%**	83.04%	84.10%
# of variables	224	112	**34**	14373

4. Gene Expression Diagnosis (Lung Cancer)				
	UAF*	RFE*	HITON*	ALL*
SVM	99.32%	98.57%	97.83%	99.07%
NN	99.63%	98.70%	98.92%	N/A
KNN	95.57%	91.49%	96.06%	97.59%
Average	98.17%	96.25%	97.60%	**98.33%**
# of variables	330	19	**16**	12,600
5. Mass-Spectrometry Diagnosis (Prostate Cancer)				
	UAF*	RFE*	HITON*	ALL*
SVM	98.50%	98.95%	99.10%	99.40%
NN	98.62%	98.78%	97.95%	99.27%
KNN	77.52%	86.53%	91.36%	76.94%
Average	91.55%	94.75%	**96.14%**	91.87%
# of variables	706	87	**16**	779

Averages Over All Tasks			
	Av. Over BaselineAlgorithms	HITON	ALL
Av. Perf. over classifiers	86.1%	**87.1%**	86.1%
Av. variable #	4540	**32.3**	33,476
Av. reduction	x 8	x **1124**	x 1

Figure 3: Task-specific and average model reduction performance (in bold, best performance per row; asterisks indicate that the corresponding algorithm yield the best model or a non-statistically significantly worse model than the best one).

Filters vs Wrappers: Which Is Best?

- None over all possible classification tasks!
- We can only prove that a *specific* filter (or wrapper) algorithm for a *specific* classifier (or class of classifiers), and a *specific* class of distributions yields optimal or sub-optimal solutions. Unless we provide such proofs we are operating on faith and hope…

A final note: What is the biological significance of selected features?

- In MB-based feature selection and CPN-faithful distributions: causal neighborhood of target (i.e., direct causes, direct effects, direct causes of the direct effects of target).
- In other methods: ???