Review of some concepts in predictive modeling

Brigham and Women's Hospital

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support

A disjoint list of topics?

- Naïve Bayes
- Bayesian networks
- Logistic Regression
- Rough and Fuzzy Sets
- CART
- Neural Networks
- Support Vector Machines
- Hierarchical clustering, K-means
- Survival Analysis
- Evaluation
- Optimization
- Essentials of time series

Implied knowledge of

- Linear regression
- K-nearest neighbors

What predictive models do

Predict this

Case 1	0.7	-0.2	0.8
Case 2	0.6	0.5	-0.4
	-0.6	0.1	0.2
	0	-0.9	0.3
	-0.4	0.4	0.2
	-0.8	0.6	0.3
	0.5	-0.7	-0.4

Using these

and evaluate performance on new cases

0.6	-0.1	?
0.4	0.6	?
-0.1	0.2	?
0	-0.5	?
-0.3	0.4	?
-0.8	0.7	?
0.3	-0.7	?

Predictive Model Considerations

- Select a model
  - Linear, Nonlinear
  - Parametric, non-parametric
  - Data separability
  - Continuous versus discrete (categorical) outcome
  - Continuous versus discrete variables
  - One class, multiple classes
- Estimate the parameters (i.e., “learn from data”)
- Evaluate

Predictive Modeling Tenets

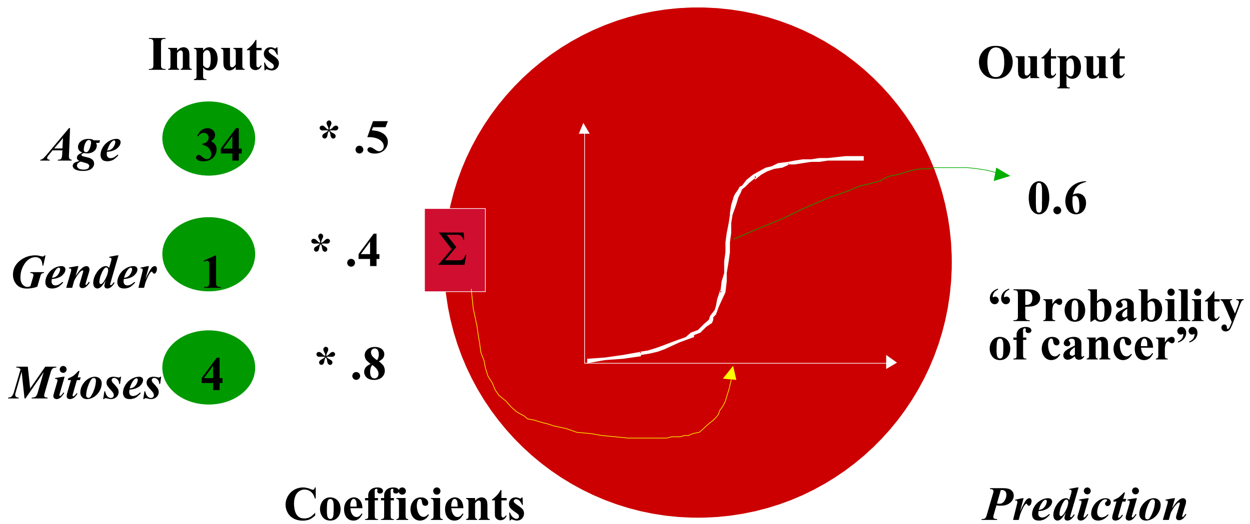
- Evaluate performance on a set of new cases
- Test set should not be used in any step of building the predictive modeling (model selection, parameter estimation)
- Avoid overfitting
  - "Rule of thumb": 2-10 times more cases than attributes
  - Use a portion of the training set for model selection or parameter tuning
- Start with simpler models as benchmarks

Desirable properties of models

- Good predictive performance (even for non-linearly separable data)
- Robustness (outliers are ignored)
- Ability to be interpreted
  - Indicate which variables contribute more for the predictions
  - Indicate the nature of variable interactions
  - Allow visualization
- Be easily applied, be generalizable to other measurement instruments, and easily communicated

Logistic Regression

- Good for interpretation
- Works well only if data are linearly separable
- Interactions need to be entered manually
- Not likely to overfit if # variables is low



$$p = \frac{1}{1 + e^{-(\Sigma+\alpha)}}$$

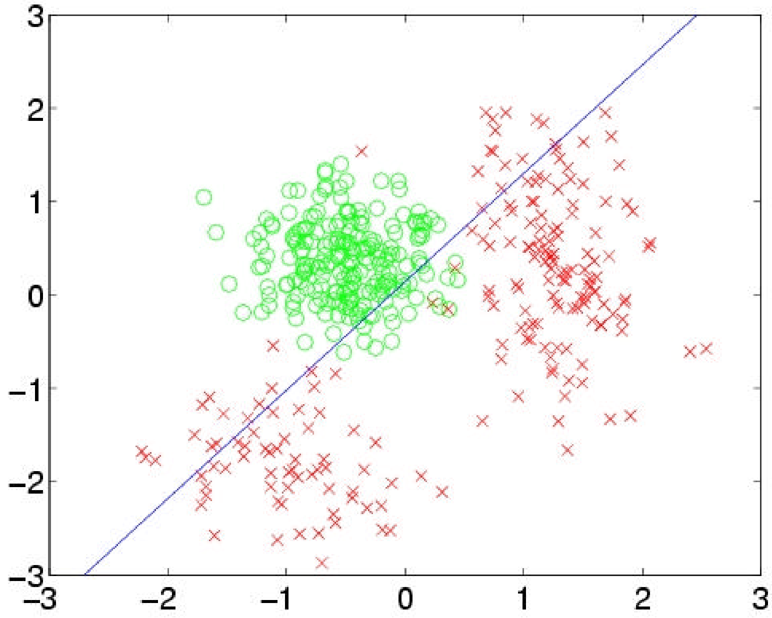
Logistic Regression and non-linearly-separable problems

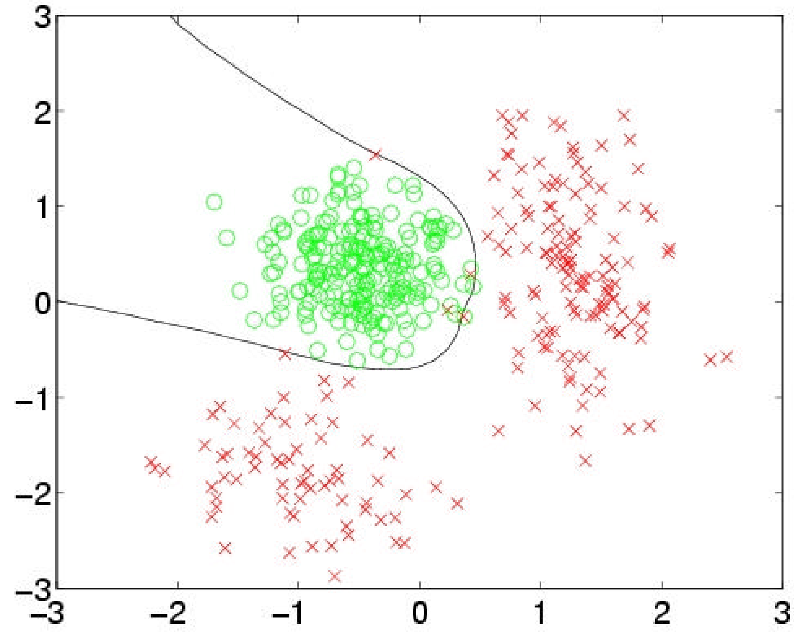
- Simple form cannot deal with it
- Y = 1/(1+exp-(ax1+bx2)

- Adding interaction terms transforms the space such that problem may become linearly separable
- Y = 1/(1+exp-(ax1 + bx2 + cx1x2))

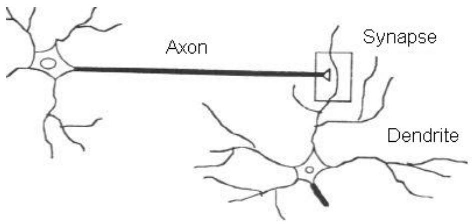
From perceptrons to multilayer perceptrons

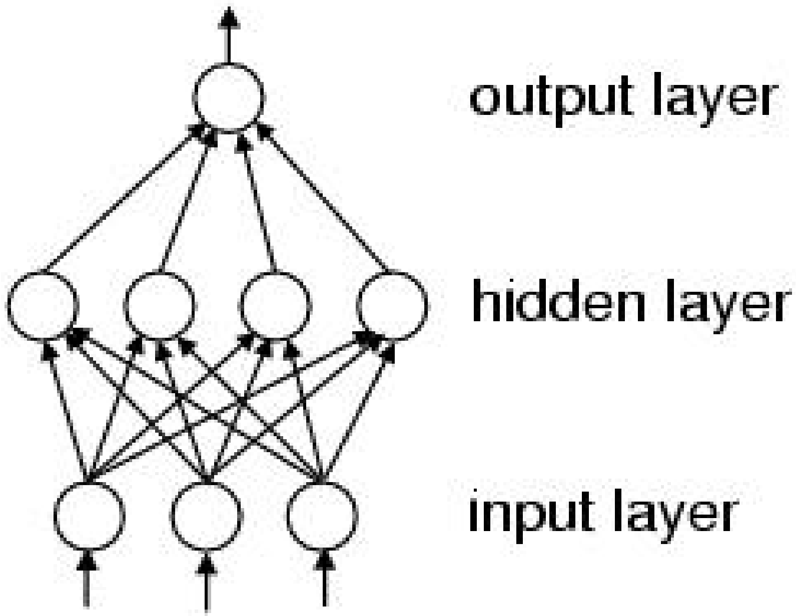
Why?





Neural Networks



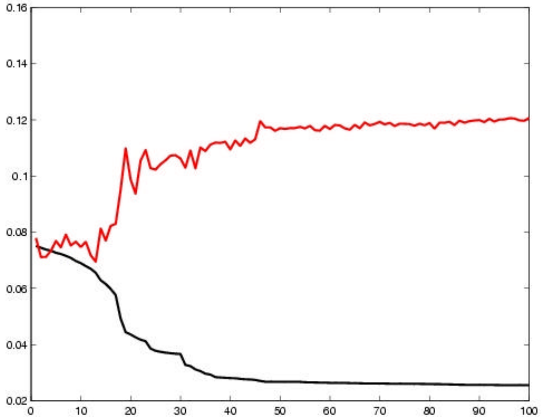


Work well even with non-linearly separable data

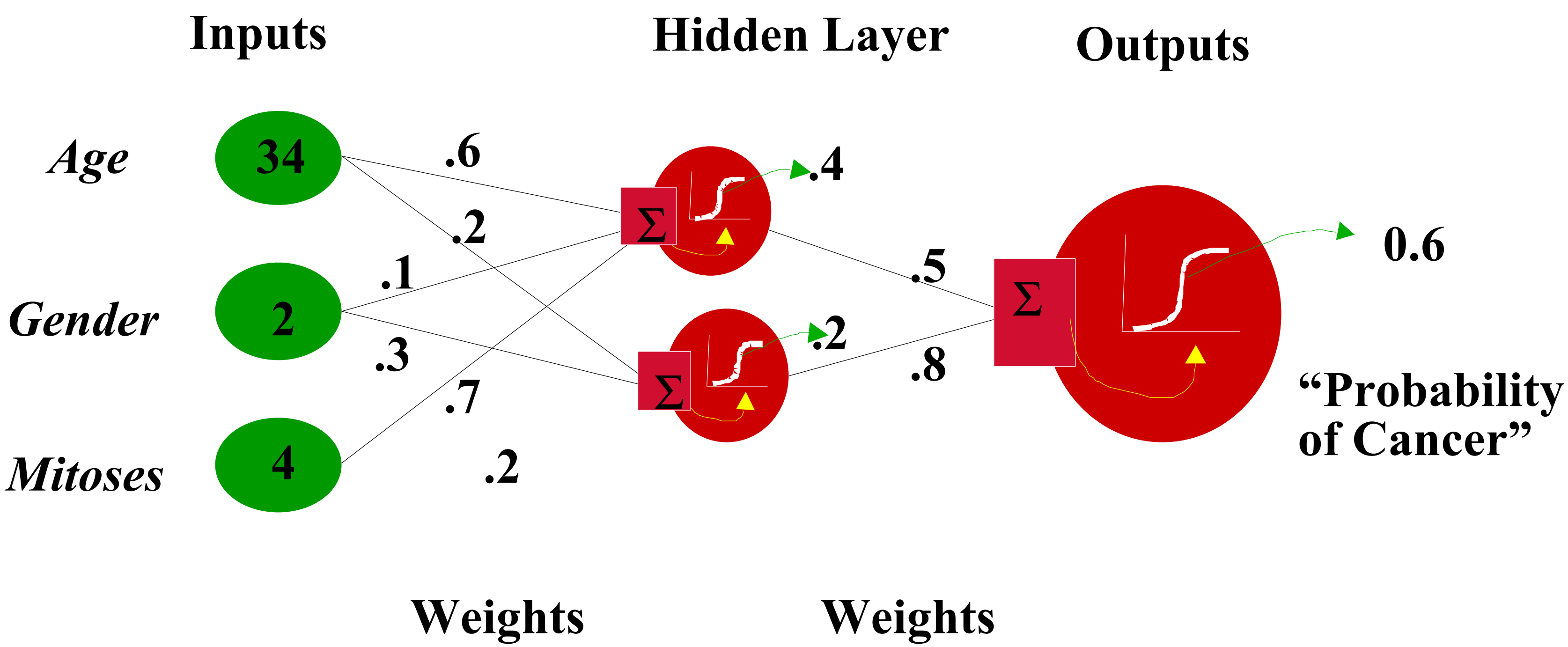
Overfitting control:

•Few weights

•Little training

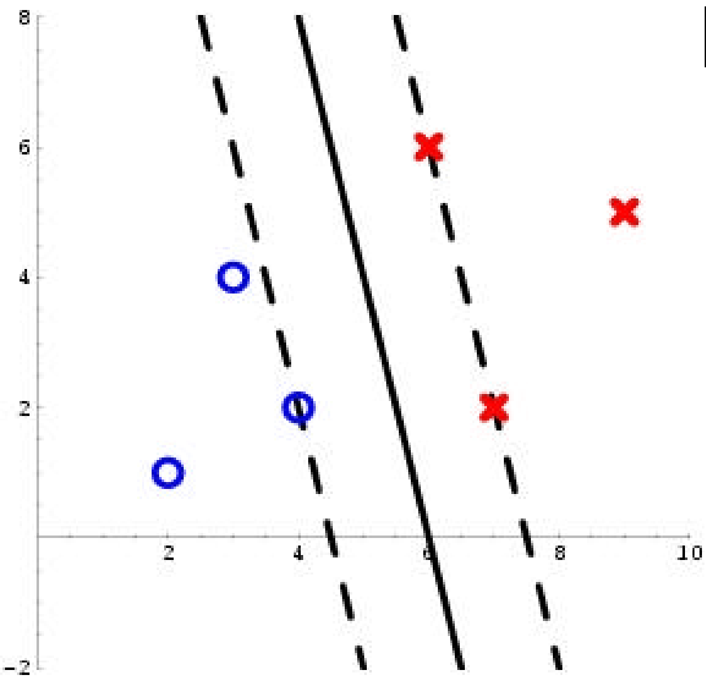


Neural Networks



Classification Trees

Support Vector Machines

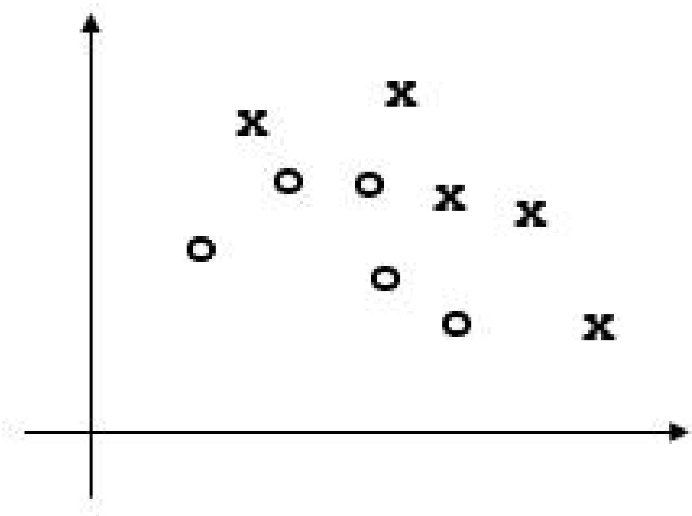


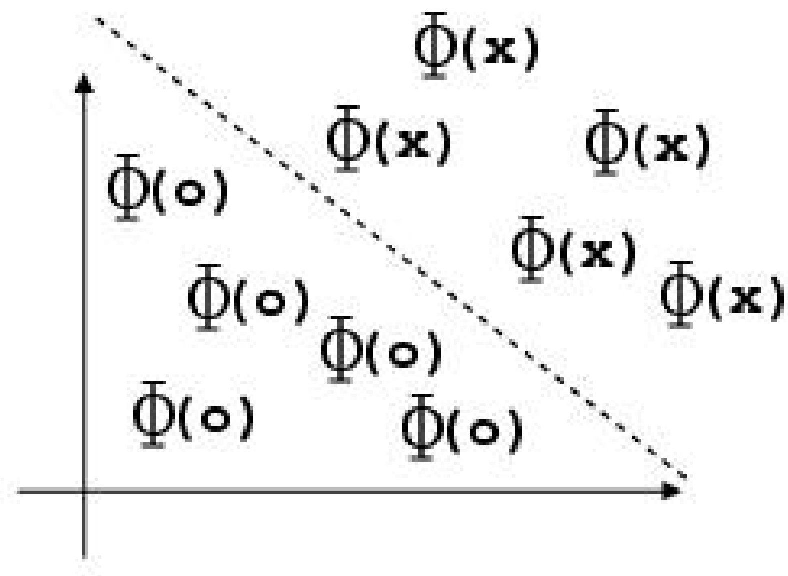
Points on dashed lines satisfy $g(x) = +1$ resp. $g(o) = -1$

- Support Vectors are at the margin
- If there are outliers in the margin, then the vectors may be "wrong"

Nonlinear SVM

- Idea: Nonlinearly project data into higher dimensional space with $\Phi: R^m \to H$
- Apply optimal hyperplane algorithm in *H*





“LARGE” data sets

- In predictive modeling, large data sets have several cases (with few attributes or variables for each case)
- In some domains, “large” data sets with several attributes and few cases are subject to analysis (predictive modeling)
- The main tenets of predictive modeling should be always used

"Large *m* small *n*" problem

- *m* variables, *n* cases
- Underdetermined systems
- Simple memorization even with simple models
- Poor generalization to new data
- Overfitting

Reducing Columns

Some approaches:

•Principal Components Analysis

(a component is a linear combination of variables with specific coefficients)

•Variable selection

0.7	~~-0.2~~	0.8
0.6	~~0.5~~	-0.4
-0.6	~~0.1~~	0.2
0	~~-0.9~~	0.3
-0.4	~~0.4~~	0.2
-0.8	~~0.6~~	0.3
0.5	~~-0.7~~	-0.4

Principal Component Analysis

- Identify direction with greatest variation (combination of variables with different weights)
- Identify next direction conditioned on the first one, and so on until the variance accounted for is acceptable

PCA disadvantage

- No class information used in PCA
- Projected coordinates may be bad for classification

Variable Selection

- Ideal: consider all variable combinations
  - Not feasible: 2^n
  - Greedy Backward: may not work if more variables than cases
- Greedy Forward:
  - Select most important variable as the “first component”
  - Select other variables conditioned on the previous ones
  - Stepwise: consider backtracking
- Other search methods: genetic algorithms that optimize classification performance and # variables

Simple Forward Variable Selection

- Conditional ranking of most important variables is possible
- Easy interpretation of resulting LR model
  - No artificial axis that is a combination of variables as in PCA
- No need to deal with too many columns
- Selection based on outcome variable
  - uses classification problem at hand

Optimization

- Not only for variable or case selection! But it helps here
- Family of problems in which there is a function to maximize/minimize, and constraints
- For example, in classification problems you may be minimizing squared errors, cross-entropy
- Knowing the complexity of the problem helps select algorithm (hill-climbing, etc.)
- Used in classification and decision making

Visualization

- Capabilities of predictive models in this area are limited
- Clustering is often good for visualization, but it is generally not very useful to separate data into pre-defined categories
  - Hierarchical trees
  - 2-D or 3-D multidimensional scaling plots
  - Self-organizing maps

Visualizing the classification potential of selected inputs

- Clustering visualization that uses classification information may help display the separation of the cases in a limited number of dimensions
- Clustering without selection of dimensions important for classification is less expected to display this separation

See Khan et al. Nature Medicine, 7(6): 673 - 679.

Generalizing a model

- Assume:
  - You have a good classification model
  - You know which variables are important
- Wouldn't it also be nice to be able to extract simple rules such as:
  - "If X_1 is high and X_2 is high, then Y is high"
  - "If X_3 is high or X_2 is low, then Y is low"
- Maybe there is a role for using fuzzy systems that "compute with words"

Fuzzy Sets

- Zadeh (1965) introduced "Fuzzy Sets" where he replaced the characteristic function with membership
- ψ_S: U $\rightarrow$ {0,1} is replaced by
  m_S : U $\rightarrow$ [0,1]
- Membership is a generalization of characteristic function and gives a "degree of membership"
- Successful applications in control theoretic settings (appliances, gearbox)

Fuzzy Sets

- Vague concepts can be represented
- Example: Let S be the set of people of normal height
- Normality is clearly not a crisp concept

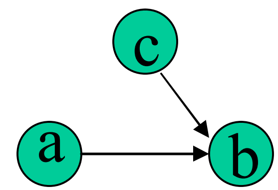
- A fuzzy classification system can use simple English rules from inputs to estimate some output

Common steps in a simple Fuzzy System

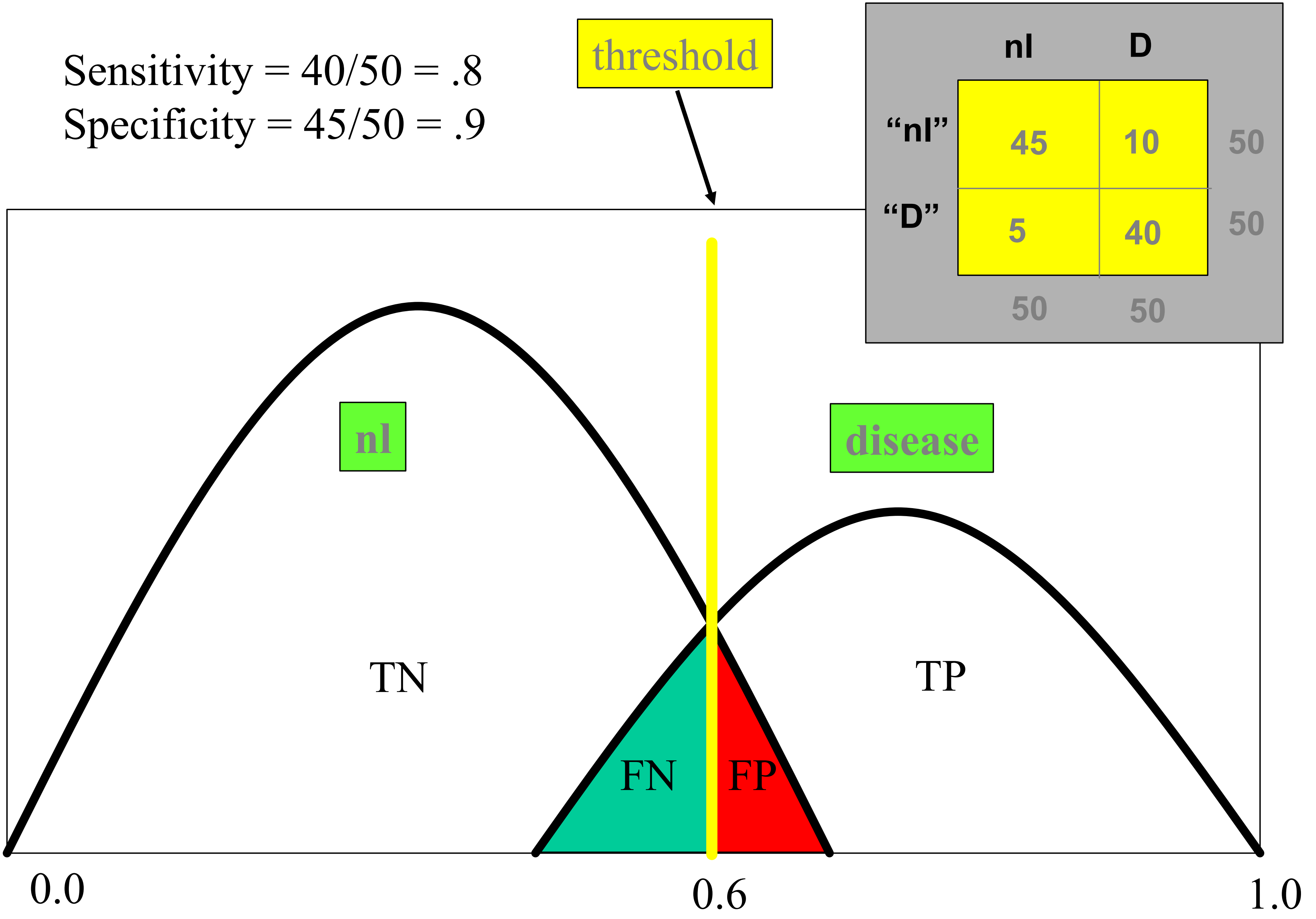
- Fuzzification of variables (determine the parameters of the membership functions)
- Fuzzy inference (apply fuzzy operations on the sets)
- Deffuzification of output (translate membership of outcome variable back to numbers if necessary)

Incorporating Knowledge

- Incorporating prior beliefs: Bayesian framework
- Conditional probabilities
- Bayesian networks
- Structures and probabilities can be obtained from data or experts
- Structure tells which variables are conditionally independent (d-separation)

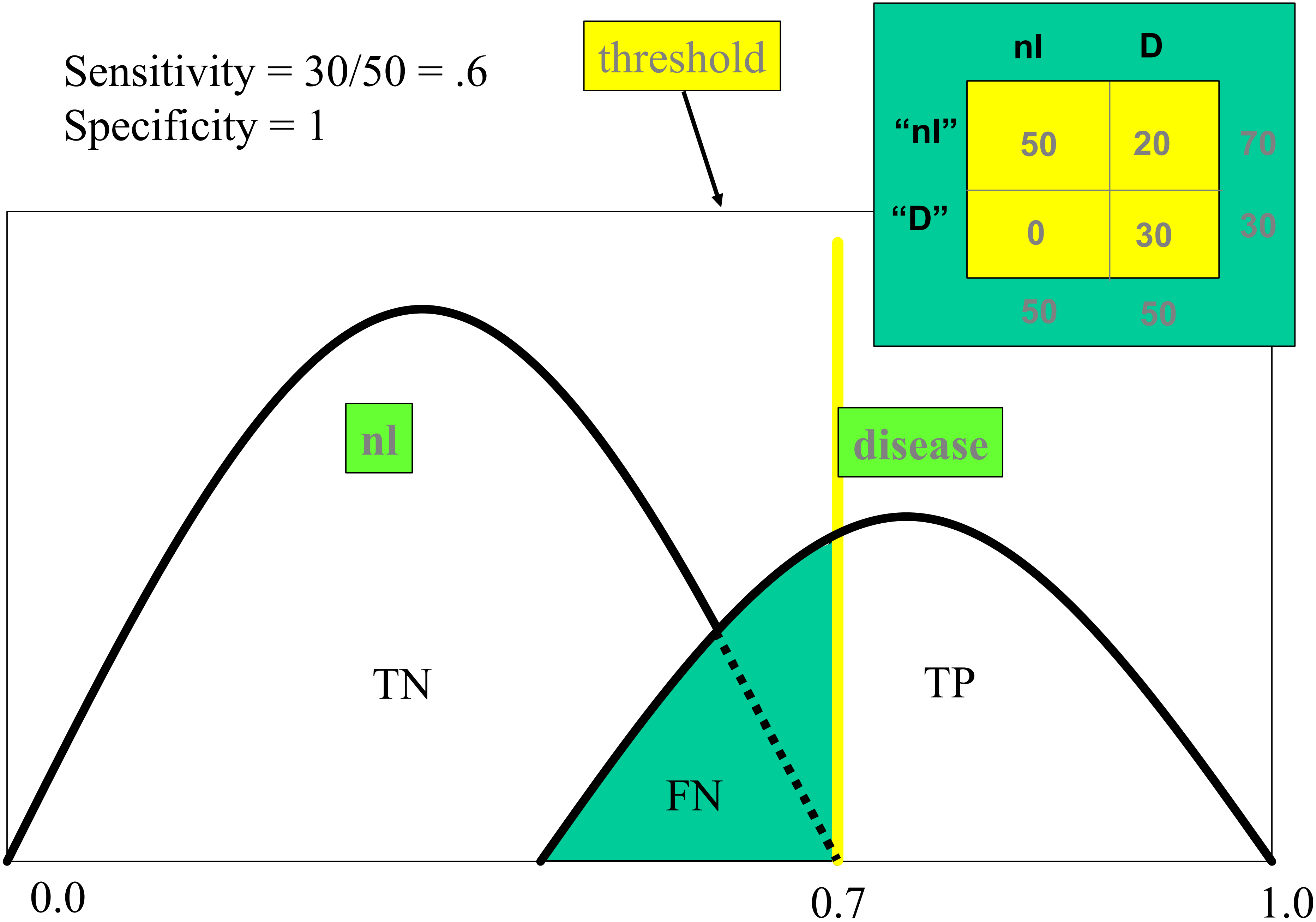


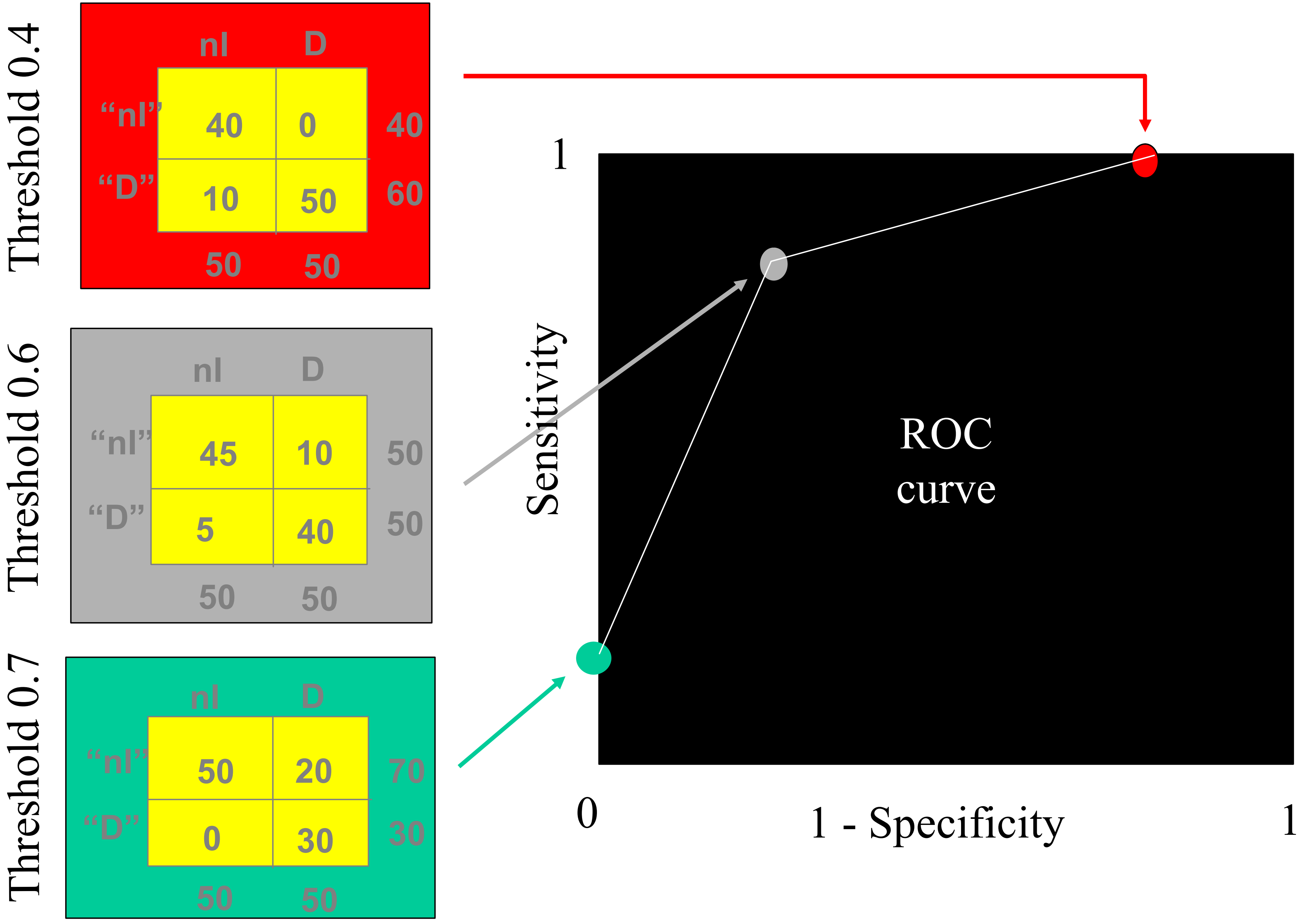
Sensitivity = 40/50 = .8
Specificity = 45/50 = .9



Sensitivity = 30/50 = .6

Specificity = 1





Threshold 0.4
nl
D
"nl"
40
0
40
"D"
10
50
60
50
50
Threshold 0.6
nl
D
"nl"
45
10
50
"D"
5
40
50
50
50
Threshold 0.7
nl
D
"nl"
50
20
70
"D"
0
30
30
50
50
1
Sensitivity
ROC
curve
0
1 - Specificity
1

Data set I

- **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks**
  Javed Khan[1, 2, 7], Jun S. Wei[1, 7], Markus Ringnér[1, 3, 7], Lao H. Saal[1], Marc Ladanyi[4], Frank Westermann[5], Frank Berthold[6], Manfred Schwab[5], Cristina R. Antonescu[4], Carsten Peterson[3] & Paul S. Meltzer[1]
- **Nature Medicine, June 2001 Volume 7 Number 6 pp 673 - 679**

- **4 small, round blue cell tumors (SRBCTs) of childhood: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS).**
- **6567 genes, 63 training samples (40 cell lines, 23 tumor samples), 25 test samples**

SRBCT

	Log-R	ANN
# genes	8	96
Correct	18	18
Unclassified	5	7
Incorrect	2	0

Hierarchical Clustering on Khan Data using 8 genes

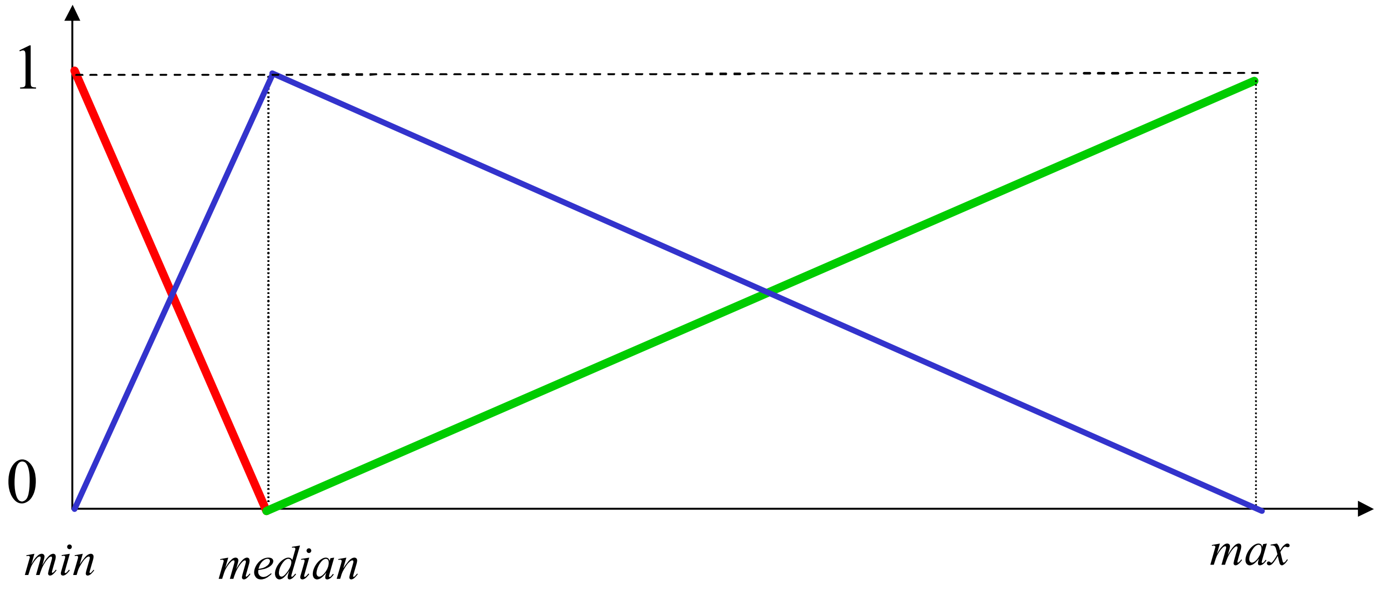
Gene	Category
H62098	BL
T54213 *	RMS
AA705225 *+	RMS
R36960	RMS
H06992 *+	NB
W49619 +	NB
AA459208 +	EWS
AA937895 +	EWS

"+" : gene in the list of 96 from Khan model

"*" : strong support from literature

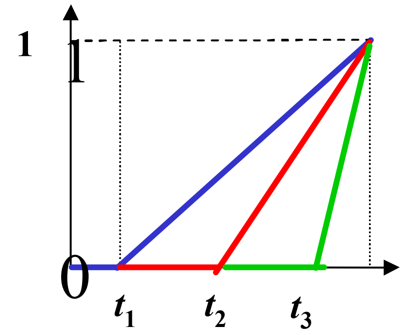
A Simple Fuzzy System

- Membership in *Low, Medium, High*
- Triangular function based on *min, median, max*



Class memberships

- Establish a threshold for membership for each class
  - t_1: max of negative cases (low)
  - t_3 : min of positive cases (high)
  - t_2 : medium point between t_1 and t_3



- Select class with highest membership
- Reject if more than more than one class with same max membership

Rule Discovery

- Ad-hoc human inspection of the training set
- A rule could have the variables from the LR model, would use “HIGH” if coefficient positive and “LOW” otherwise
  - IF H62098 is HIGH then BL
  - If any two of (T54123, AA705225, R36960) are HIGH, then RMS
  - If (H06992 is HIGH and W49619) is HIGH then NB
  - If (AA459208 is HIGH and AA937895 is HIGH) then EWS

Gene	Category
H62098	BL
T54213 *	RMS
AA705225 *+	RMS
R36960	RMS
H06992 *+	NB
W49619 +	NB
AA459208 +	EWS
AA937895 +	EWS

Some reminders

- Simple models may perform at the same level of complex ones for certain data sets
- A benchmark can be established with these models, which can be easily accessed
- Simple rules may have a role in generalizing results to other platforms
- No model can be proved to be best, need to try all

Project guidelines

- Justify choices for the data and classification model
- Research what has already been done (benchmarks)
- Establish your own benchmarks in simple models
- Evaluate appropriately
- Acknowledge limitations
- Indicate what else could have been done (and why you did not do it)