Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

Review of some concepts in predictive modeling

Decision Systems Group

Lucila Ohno-Machado,

Brigham and Women's Hospital

Topics

- Decision trees
- Linear regression
- Logistic regression
- Evaluation
- Classification trees
- Ensembles
- PCA
- Clustering
- MDS
- Neural nets

2 x 2 table
(contingency table)

	PPD+	PPD-	
TB	**8**	**2**	**10**
no TB	**3**	**87**	**90**
	11	**89**	**100**

Probability of TB given PPD- = 2/89

Bayes rule

- Definition of conditional probability:
- $P(A|B) = P(AB)/P(B)$

$$P(B|A) = P(BA)/P(A)$$

$$P(AB) = P(BA)$$

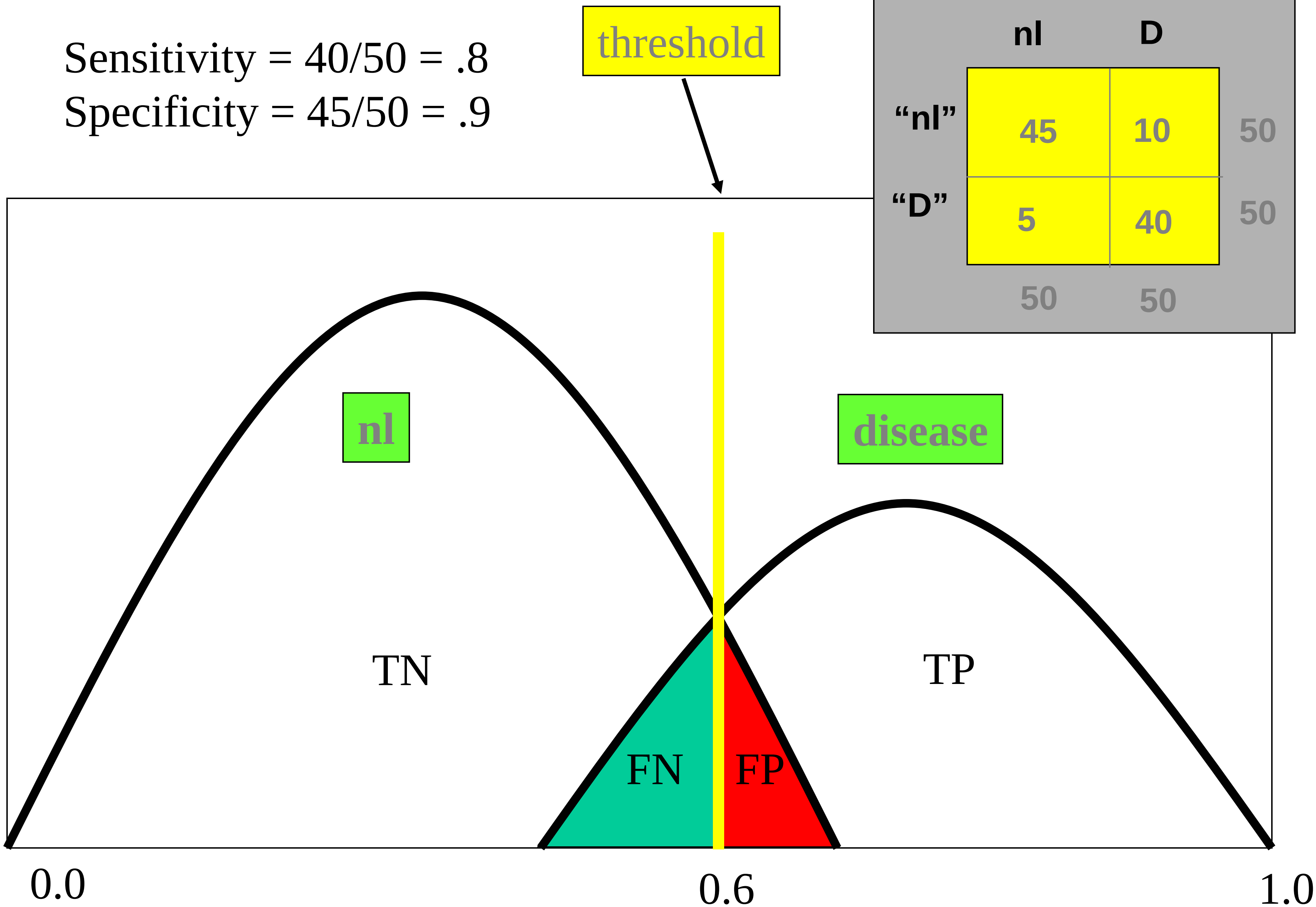
$$P(A|B)P(B) = P(B|A)P(A)$$

$$\mathbf{P(A|B) = P(B|A)P(A)/P(B)}$$

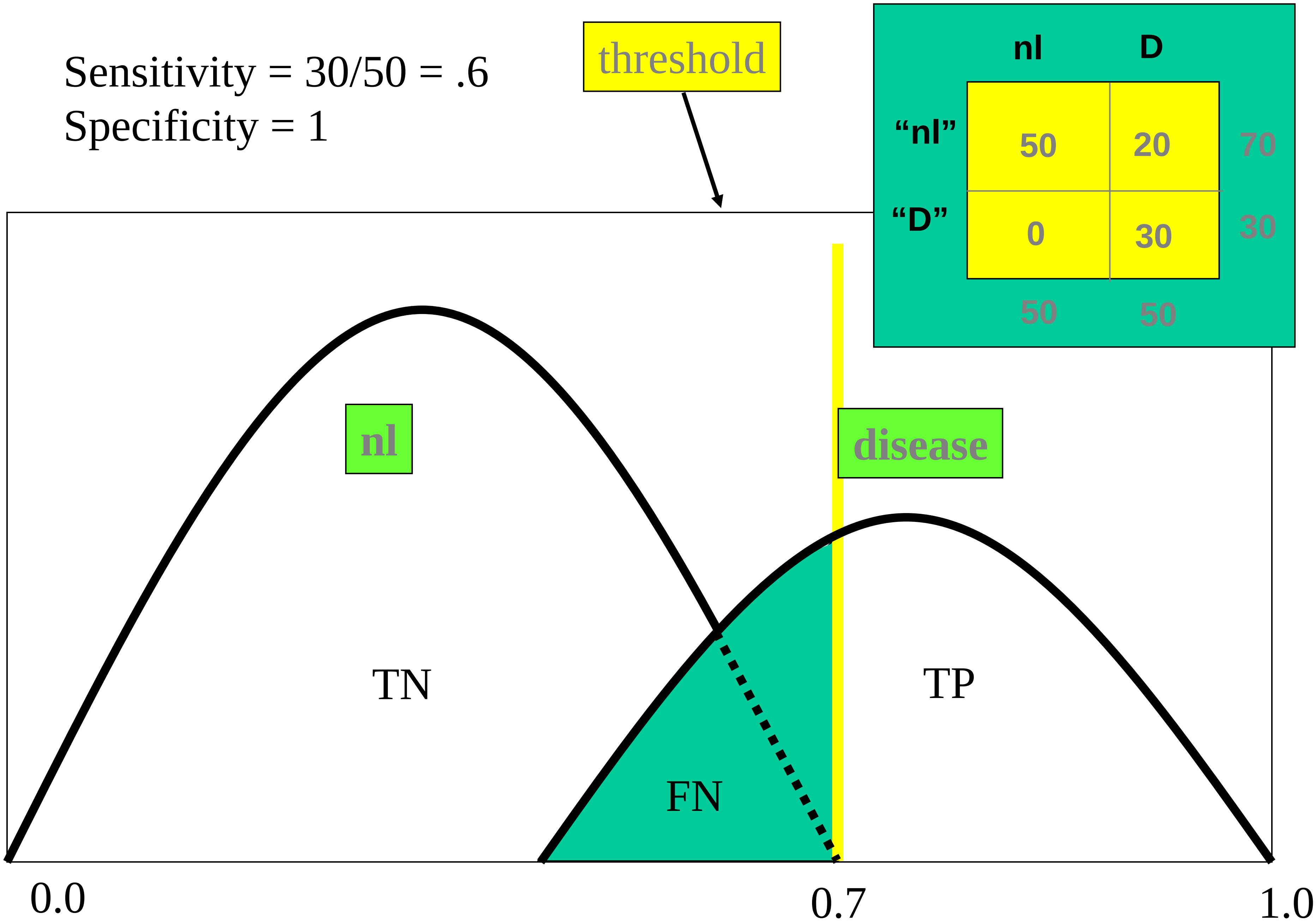
Sensitivity = 40/50 = .8

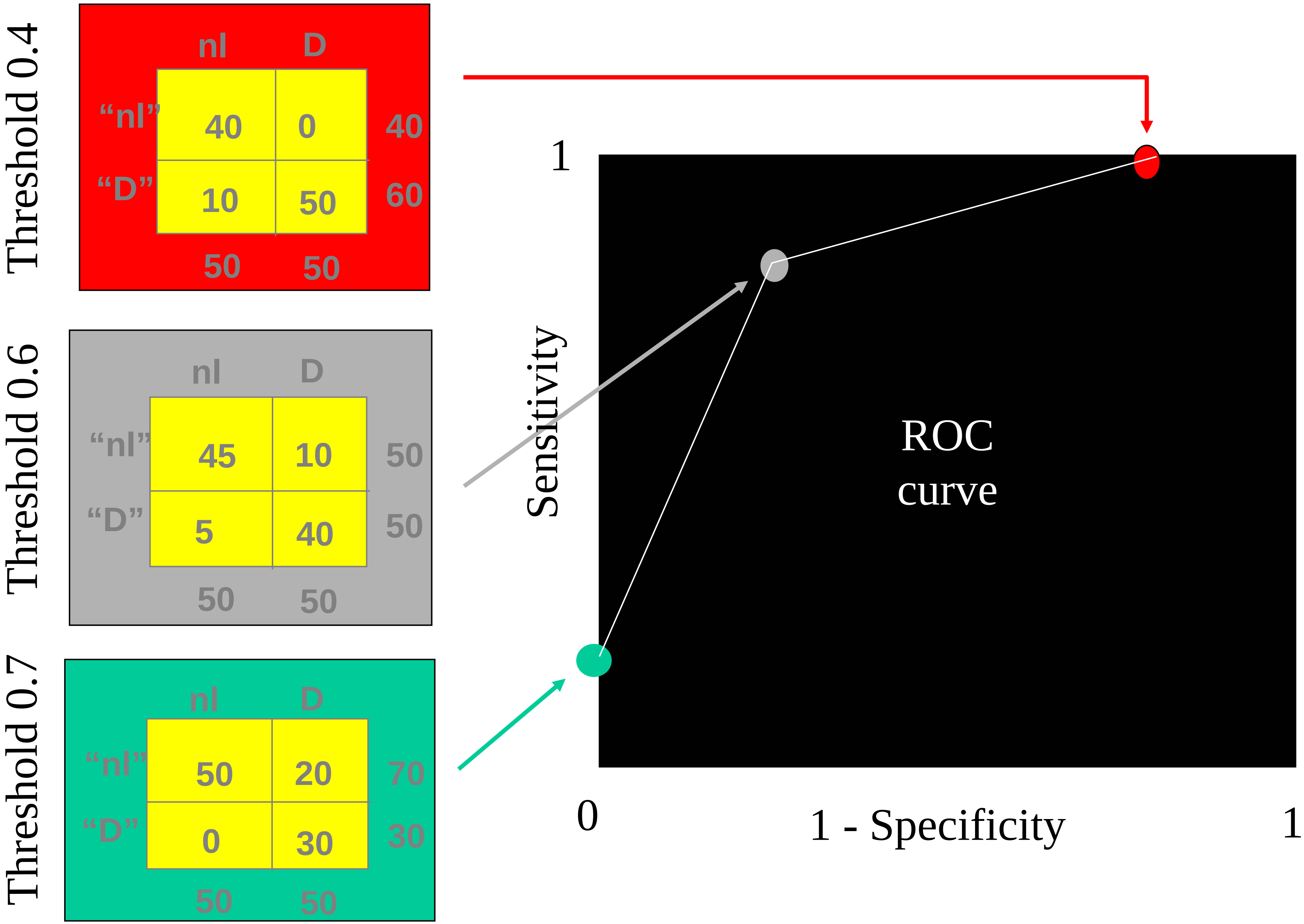
Specificity = 45/50 = .9

	nl	D	
"nl"	45	10	50
"D"	5	40	50
	50	50	



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

All possible pairs 0-1

- Healthy

0.3 <

0.2

0.5

0.1

0.7

- Sick

0.8

0.2

0.5

0.7

0.9

concordant

discordant

concordant

concordant

concordant

All possible pairs 0-1

Systems' estimates for

- Healthy

0.3

0.2

0.5

0.1

0.7

- Sick

0.8

0.2

0.5

0.7

0.9

concordant

tie

concordant

concordant

concordant

C - index

- Concordant

18

- Discordant

4

- Ties

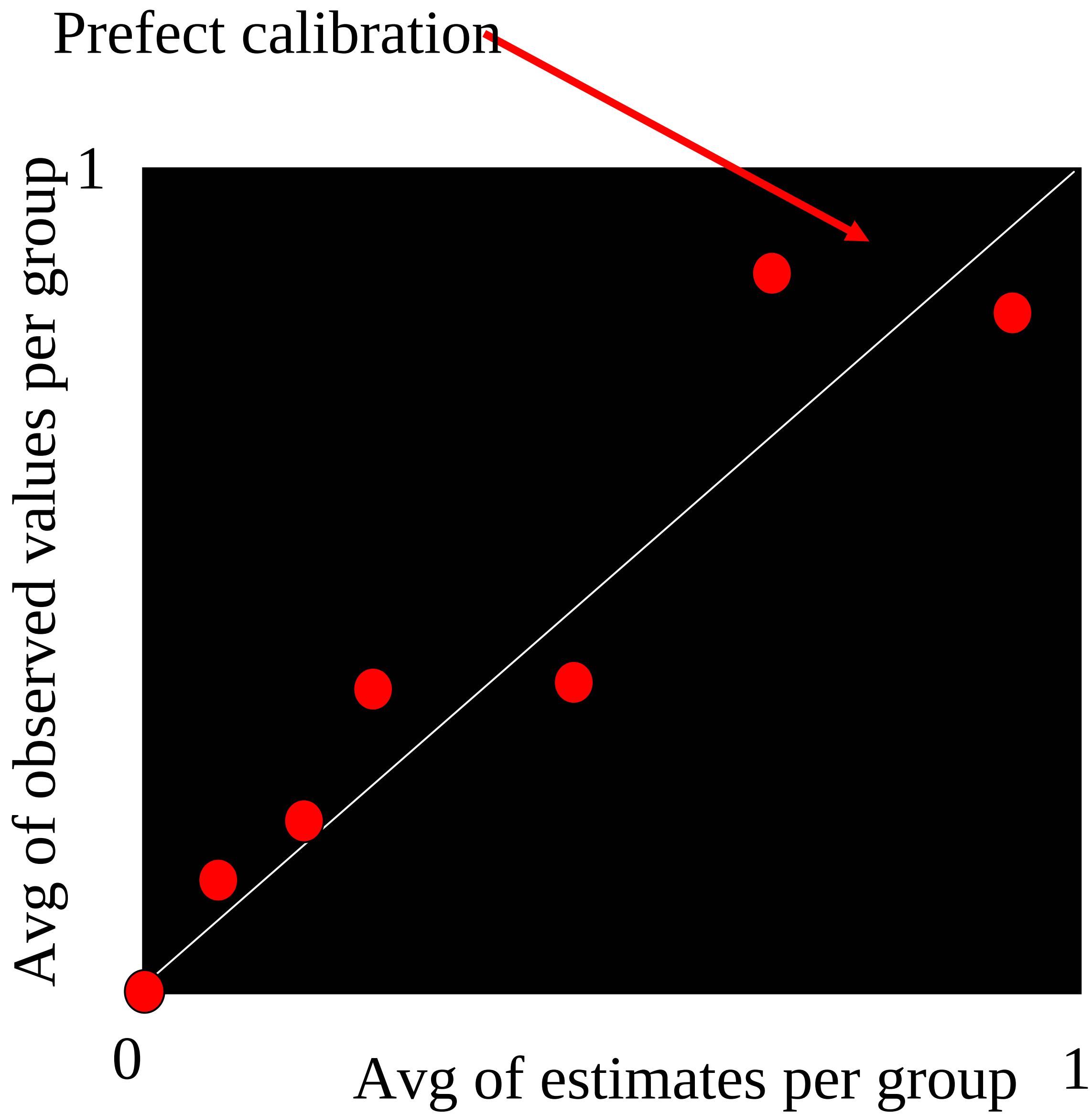
3

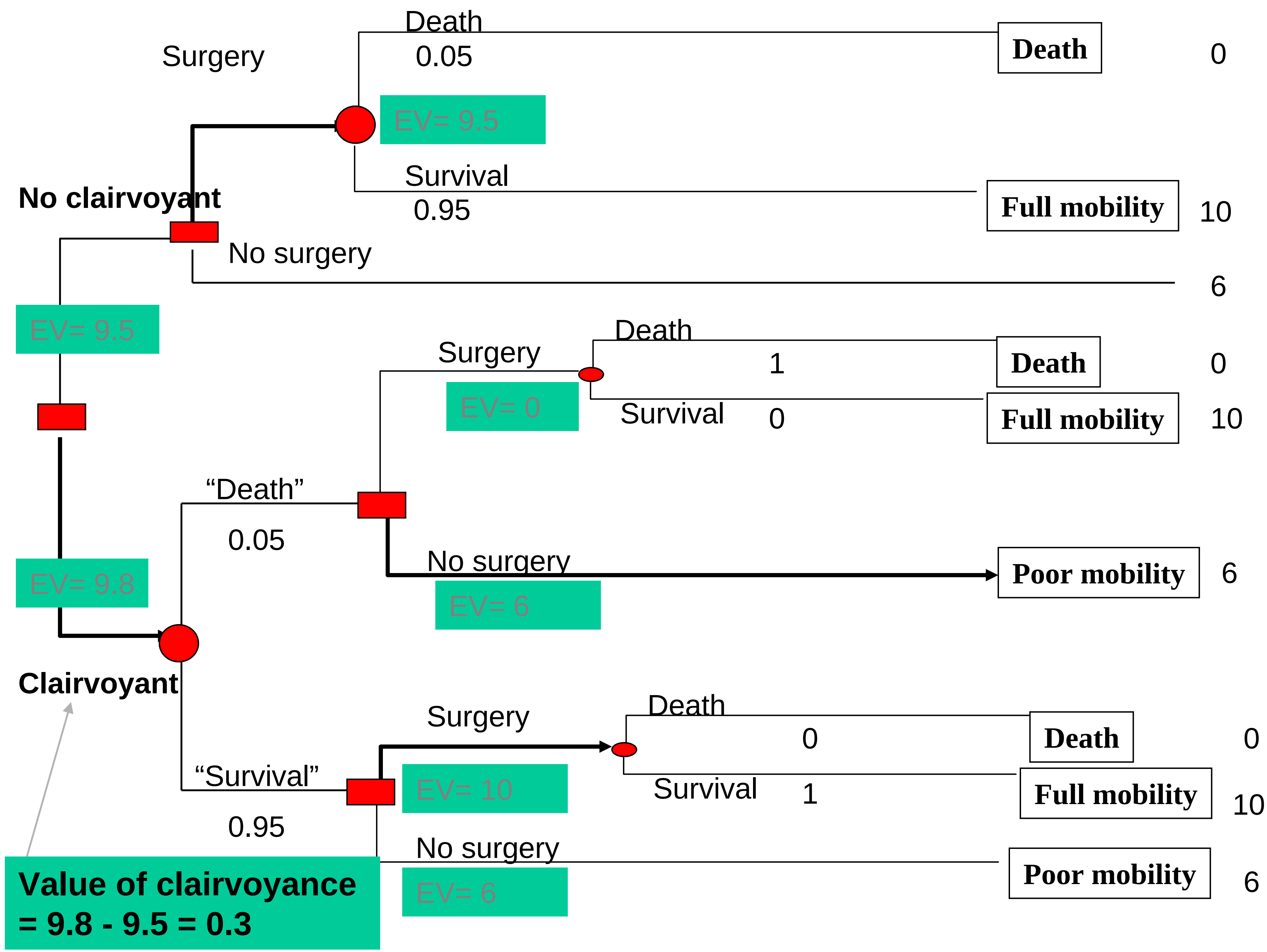
$$\text{C -index} = \frac{\text{Concordant} + 1/2\ \text{Ties}}{\text{All pairs}} = \frac{18 + 1.5}{25}$$

Calibration

Sorted pairs by systems' estimates			Real outcomes
0.1			**0**
0.2			**0**
0.2	**sum of group = 0.5**	**1**	**sum = 1**
0.3			**0**
0.5			**0**
0.5	**sum of group = 1.3**	**1**	**sum = 1**
0.7			**0**
0.7			**1**
0.8			**1**
0.9	**sum of group = 3.1**	**1**	**sum = 3**

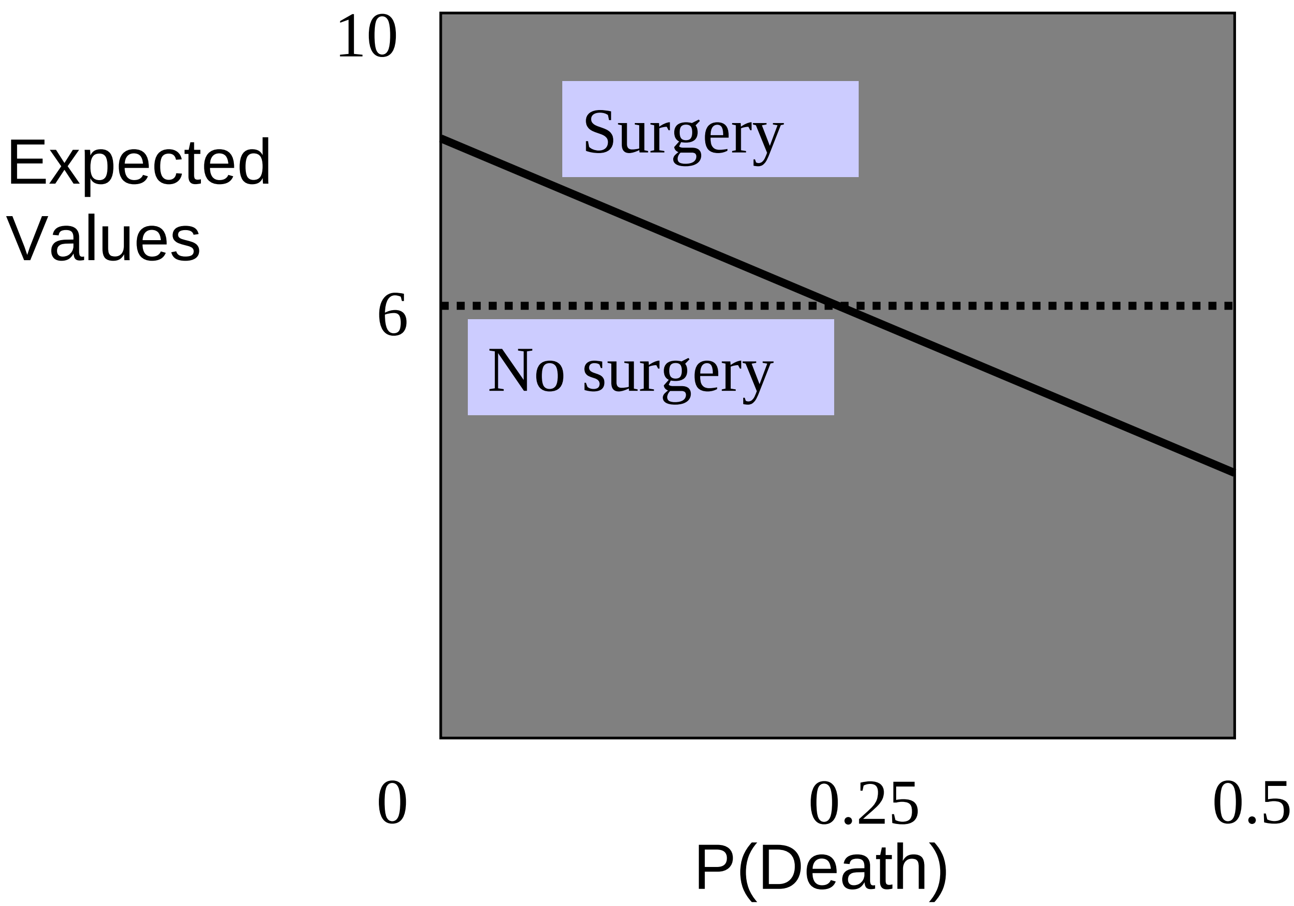
Calibration plot




Surgery
Death
0.05
Death
0
EV= 9.5
Survival
0.95
Full mobility
10
No clairvoyant
No surgery
6
EV= 9.5
Surgery
Death
1
Death
0
EV= 0
Survival
0
Full mobility
10
"Death"
0.05
No surgery
Poor mobility
6
EV= 9.8
EV= 6
Clairvoyant
Surgery
Death
0
Death
0
"Survival"
EV= 10
Survival
1
Full mobility
10
0.95
No surgery
Poor mobility
6
EV= 6
Value of clairvoyance
= 9.8 - 9.5 = 0.3

Sensitivity Analysis

- Effect of probabilities in the decision



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

$$correlation\_coefficient$$

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho$$

$$VARIANCE$$

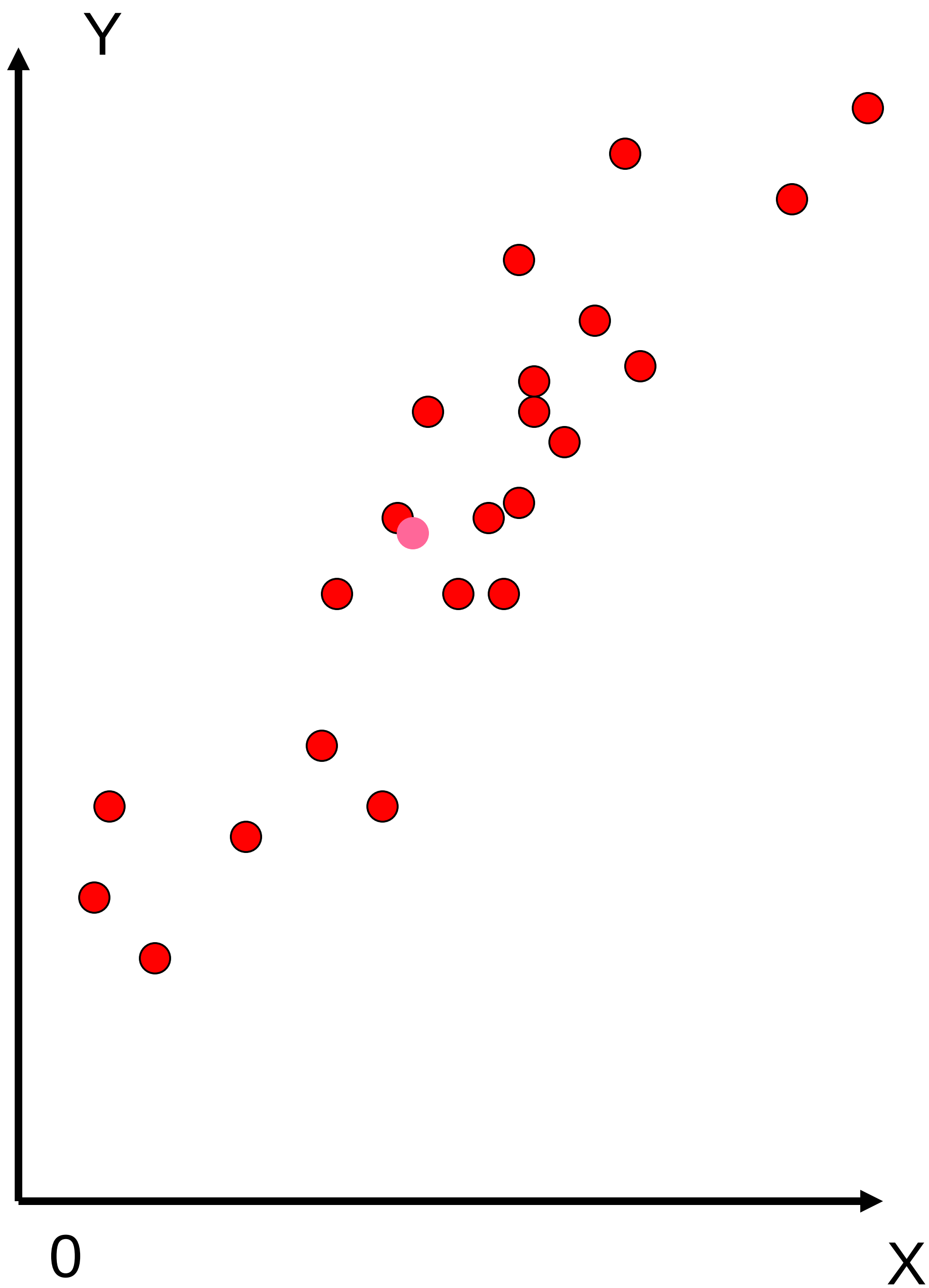
$$\sigma_{XX} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})}{n-1}$$

$$COVARIANCE$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

$$st\_deviation$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})}{n-1}}$$



Covariance and Correlation Matrices

$$cov = \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{bmatrix}$$

$$corr = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

$$\sigma_{XX} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})}{n-1}$$

Slope from linear regression is asymmetric, covariance and ρ are symmetric

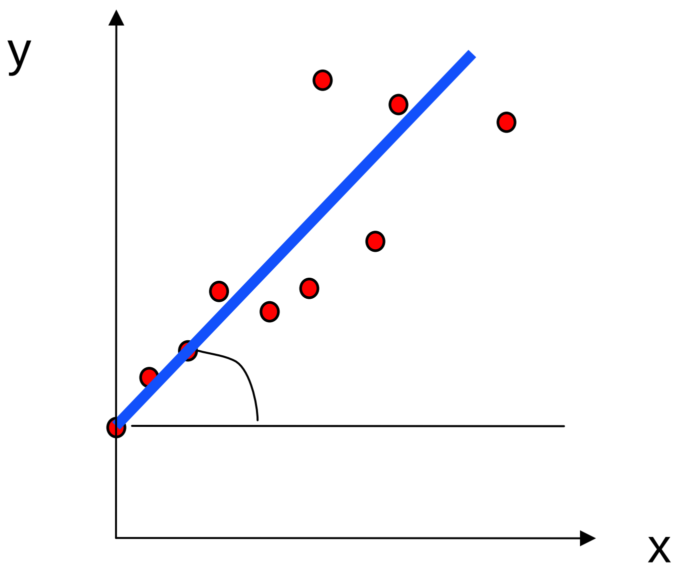
$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$y = \beta_0 + \beta_1 x$$

$$y = 2 + 4x$$

$$x = y/4 - 2$$



$$\mathrm{cov} = \begin{bmatrix} 0.86 & 0.35 \\ 0.35 & 15.69 \end{bmatrix} = \Sigma$$

$$corr = \begin{bmatrix} 1 & 0.96 \\ 0.96 & 1 \end{bmatrix}$$

Solve system of normal equations

$$\beta_0 n + \beta_1 \sum x = \sum y$$

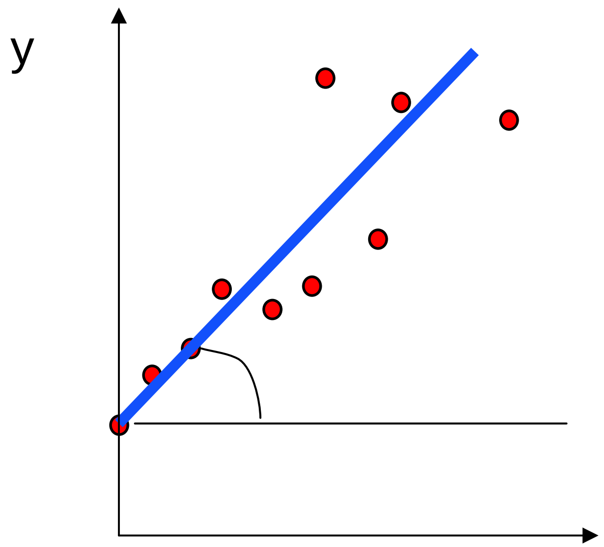
Normal equation 1

$$\beta_0 \sum x + \beta_1 \sum x^2 = \sum yx$$

Normal equation 2

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

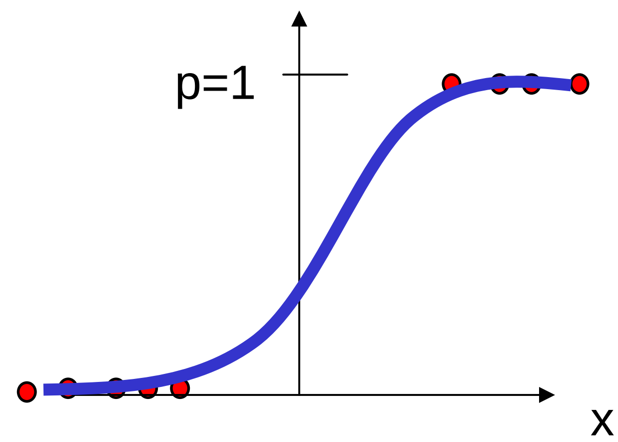
$$\beta_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$



Logit Model

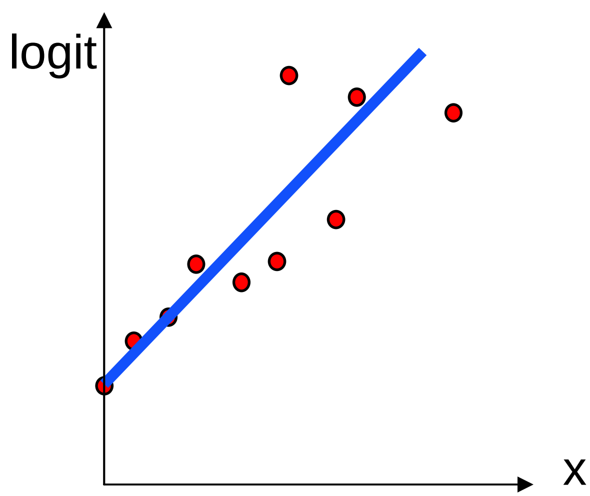
$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{e^{\beta_0 + \beta_1 x_i} + 1}$$



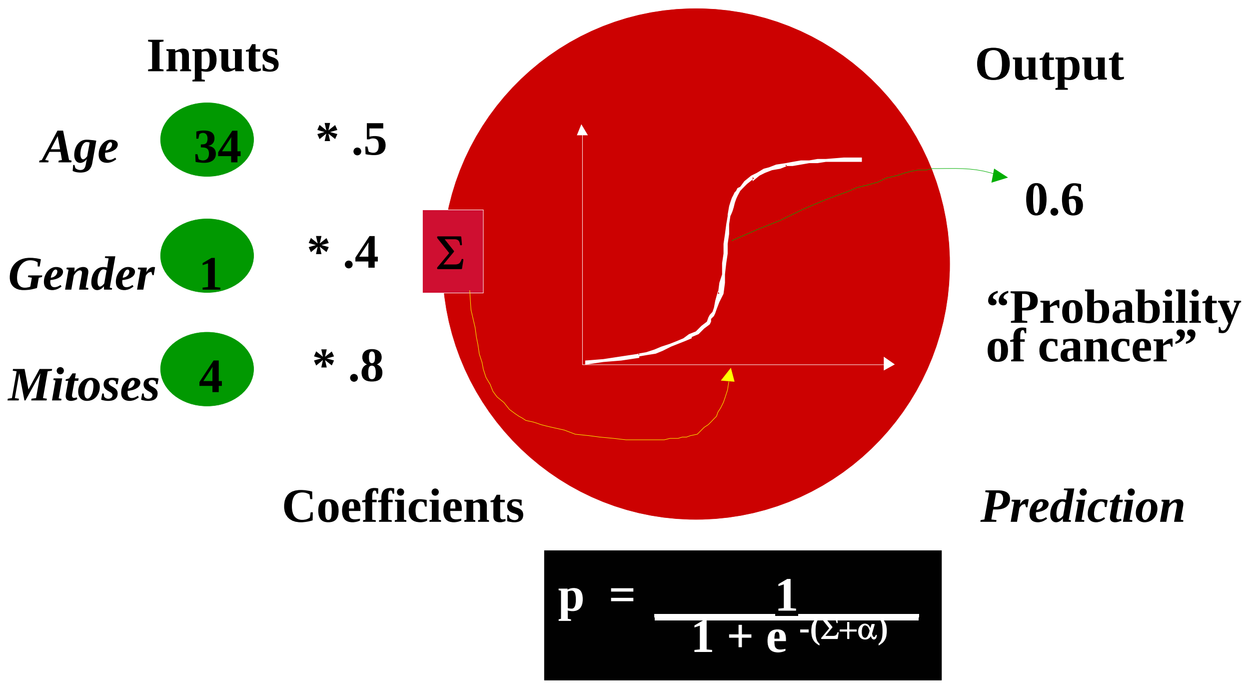
$$\log\left[\frac{p_i}{1 - p_i}\right] = \beta_0 + \beta_1 x_i$$

$$\log\left[\frac{p_i}{1 - p_i}\right] = \sum_i \beta x_i$$



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



What do coefficients mean?

$$e^{\beta_{age}} = OR_{age}$$

	Age49	Age50	
Death	28	22	50
Life	45	52	97
Total	73	74	147

$$OR = \frac{\dfrac{p_{death|age=50}}{1 - p_{death|age=50}}}{\dfrac{p_{death|age=49}}{1 - p_{death|age=49}}}$$

What do coefficients mean?

$$e^{\beta_{color}} = OR_{color}$$

	Blue	Green	
Death	28	22	50
Life	45	52	97
Total	73	74	147

$$OR = \frac{28/45}{22/52} = 1.47$$

$$e^{\beta_{color}} = 1.47$$

$$\beta_{color} = 0.385$$

$$p_{blue} = \frac{1}{1+e^{-(-0.8616+0.385)}} = 0.383$$

$$p_{green} = \frac{1}{1+e^{0.8616}} = 0.297$$

Maximum Likelihood Estimation

- Steps:
  - Define expression for the probability of data as a function of the parameters
  - Find the values of the parameters that maximize this expression

Likelihood Function

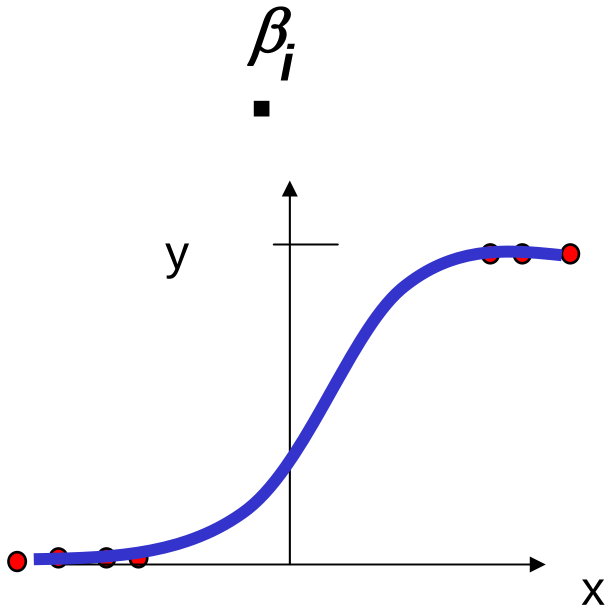
$$L = \Pr(Y)$$

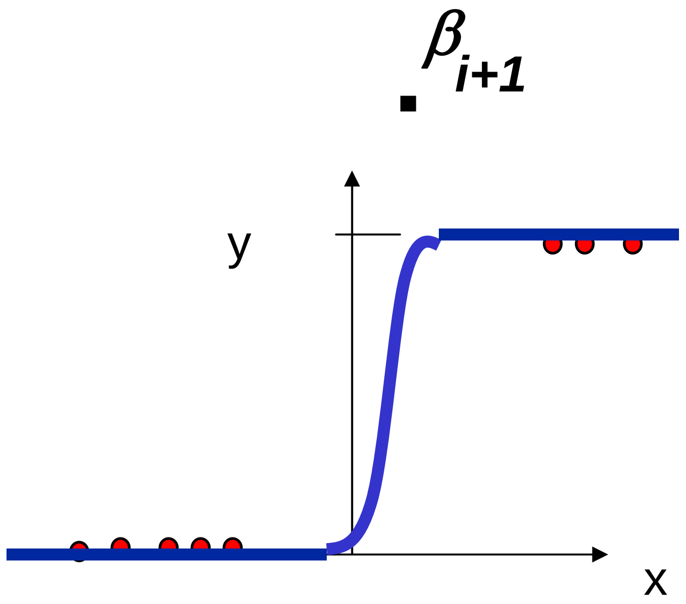
$$L = \Pr(y_1, y_2, \ldots, y_n)$$

$$L = \Pr(y_1)\Pr(y_2)\ldots\Pr(y_n) = \prod_{i=1}^{n}\Pr(y_i)$$

Complete separation

MLE does not exist (ie, it is infinite)





x

Logistic Regression and non-linearly-separable problems

- Simple form below cannot deal with it
- $Y = 1/(1+\exp-(ax_1+bx_2))$
- Adding interaction terms transforms the space such that problem may become linearly separable
- $Y = 1/(1+\exp-(ax_1 + bx_2 + cx_1x_2))$

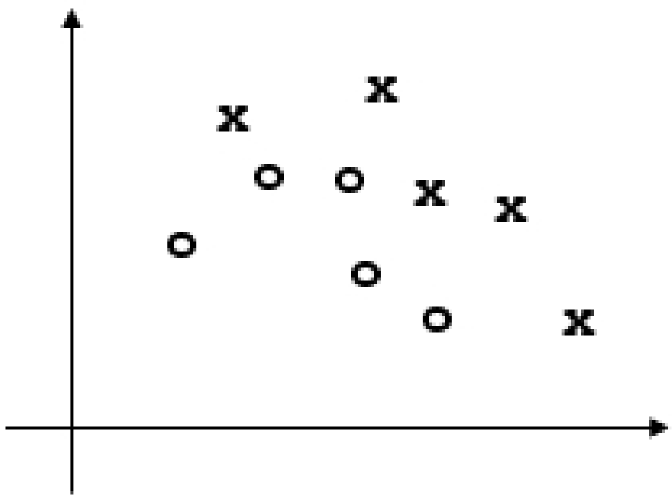
Figures removed due to copyright reasons.

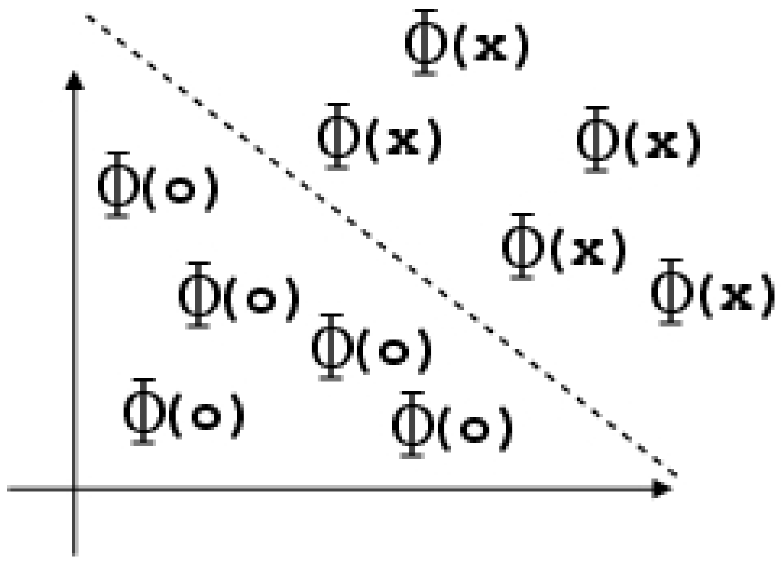
Please see:

Khan, J., et. al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nat Med* 7, no. 6 (June 2001): 673-9.

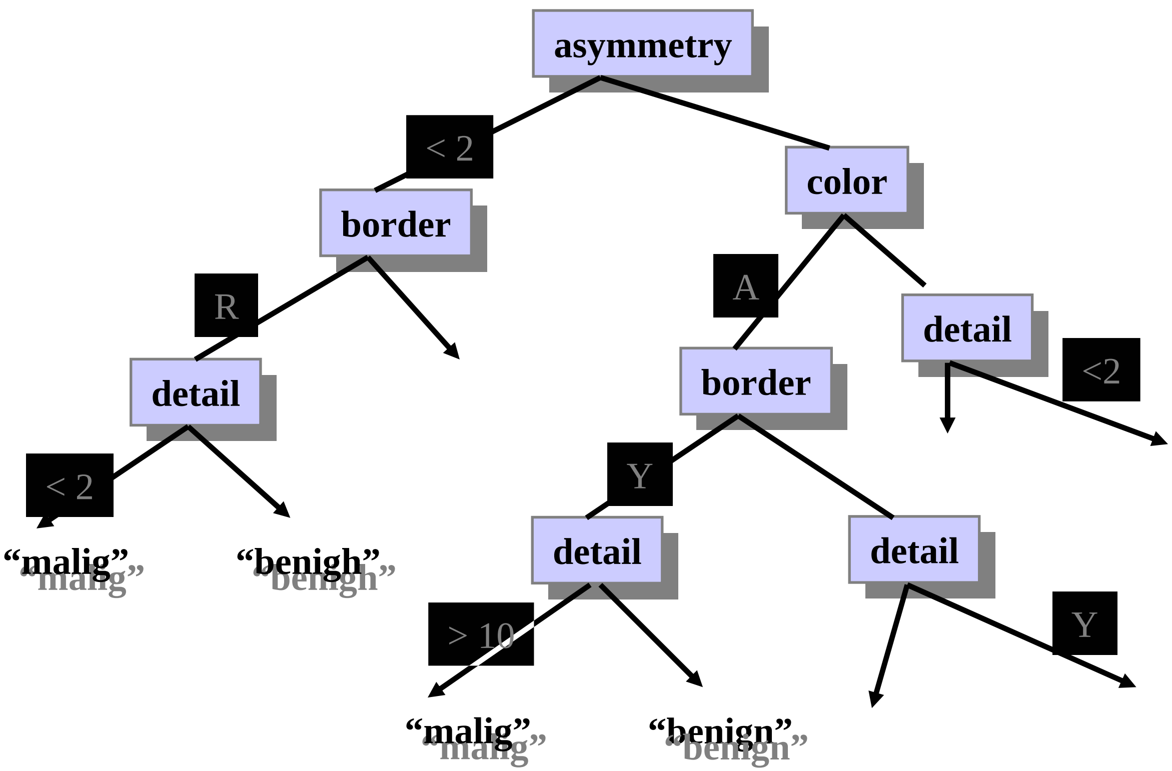
Kernel trick

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





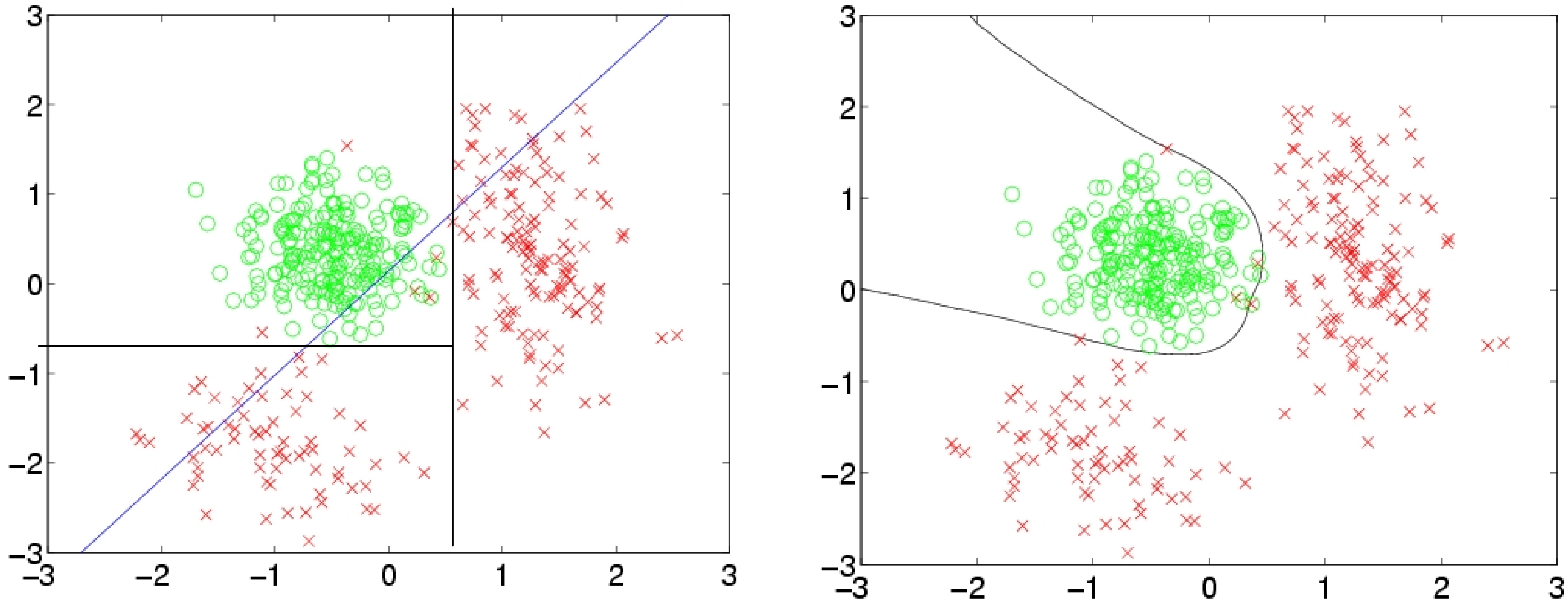
Classification Trees




```
0- TEST: null  VALUE: null Num Cases: 700.0 Num Dsrd: 241.0
  2- TEST: breath  VALUE: 1 Num Cases: 75.0 Num Dsrd: 1.0
    ********PRUNED!!!
    ********PRUNED!!!
  1- TEST: breath  VALUE: 0 Num Cases: 625.0 Num Dsrd: 240.0
    4- TEST: CWtender  VALUE: 1 Num Cases: 11.0 Num Dsrd: .0
    3- TEST: CWtender  VALUE: 0 Num Cases: 614.0 Num Dsrd: 240.0
      8- TEST: age  VALUE: >32 Num Cases: 611.0 Num Dsrd: 240.0
        10- TEST: Duration  VALUE: >72 Num Cases: 3.0 Num Dsrd: .0
        9- TEST: Duration  VALUE: <=72 Num Cases: 608.0 Num Dsrd: 240.0
          12- TEST: Duration  VALUE: >48 Num Cases: 2.0 Num Dsrd: 2.0
          11- TEST: Duration  VALUE: <=48 Num Cases: 606.0 Num Dsrd: 238.0
            14- TEST: prevang  VALUE: 1 Num Cases: 340.0 Num Dsrd: 92.0
              18- TEST: Epis  VALUE: 1 Num Cases: 8.0 Num Dsrd: .0
              17- TEST: Epis  VALUE: 0 Num Cases: 332.0 Num Dsrd: 92.0
                22- TEST: Worsening  VALUE: >72 Num Cases: 6.0 Num Dsrd: .0
                21- TEST: Worsening  VALUE: <=72 Num Cases: 326.0 Num Dsrd: 92.0
                  28- TEST: Duration  VALUE: >36 Num Cases: 3.0 Num Dsrd: .0
                  27- TEST: Duration  VALUE: <=36 Num Cases: 323.0 Num Dsrd: 92.0
                    36- TEST: Worsening  VALUE: >28 Num Cases: 3.0 Num Dsrd: 2.0
                    35- TEST: Worsening  VALUE: <=28 Num Cases: 320.0 Num Dsrd: 90.0
                      44- TEST: age  VALUE: >55 Num Cases: 240.0 Num Dsrd: 81.0
                        52- TEST: Worsening  VALUE: >0 Num Cases: 238.0 Num Dsrd: 81.0
                          64- TEST: OldMI  VALUE: 1 Num Cases: 49.0 Num Dsrd: 9.0
                            74- TEST: Smokes  VALUE: 0 Num Cases: 37.0 Num Dsrd: 9.0
                              86- TEST: age  VALUE: >65 Num Cases: 30.0 Num Dsrd: 5.0
                                ********PRUNED!!!
                                ********PRUNED!!!
                              85- TEST: age  VALUE: <=65 Num Cases: 7.0 Num Dsrd: 4.0
                                98- TEST: Worsening  VALUE: >2 Num Cases: 5.0 Num Dsrd: 2.0
                                97- TEST: Worsening  VALUE: <=2 Num Cases: 2.0 Num Dsrd: 2.0
                            73- TEST: Smokes  VALUE: 1 Num Cases: 12.0 Num Dsrd: .0
                             63- TEST: OldMI  VALUE: 0 Num Cases: 189.0 Num Dsrd: 72.0
                            72- TEST: Nausea  VALUE: 0 Num Cases: 165.0 Num Dsrd: 57.0
                              84- TEST: Duration  VALUE: >16 Num Cases: 3.0 Num Dsrd: 2.0
                              83- TEST: Duration  VALUE: <=16 Num Cases: 162.0 Num Dsrd: 55.0
                                ********PRUNED!!!
                                ********PRUNED!!!
                             71- TEST: Nausea  VALUE: 1 Num Cases: 24.0 Num Dsrd: 15.0
                              82- TEST: Back  VALUE: 0 Num Cases: 21.0 Num Dsrd: 15.0
                                94- TEST: post  VALUE: 1 Num Cases: 1.0 Num Dsrd: .0
                                93- TEST: post  VALUE: 0 Num Cases: 20.0 Num Dsrd: 15.0
                              81- TEST: Back  VALUE: 1 Num Cases: 3.0 Num Dsrd: .0
                        51- TEST: Worsening  VALUE: <=0 Num Cases: 2.0 Num Dsrd: .0
                      43- TEST: age  VALUE: <=55 Num Cases: 80.0 Num Dsrd: 9.0
                        50- TEST: Worsening  VALUE: >1 Num Cases: 68.0 Num Dsrd: 5.0
                          ********PRUNED!!!
                          ********PRUNED!!!
                            ********PRUNED!!!
                            ********PRUNED!!!
                              ********PRUNED!!!
                                ********PRUNED!!!
                                ********PRUNED!!!
                              ********PRUNED!!!
                        49- TEST: Worsening  VALUE: <=1 Num Cases: 12.0 Num Dsrd: 4.0
                          60- TEST: age  VALUE: >47 Num Cases: 10.0 Num Dsrd: 2.0
                            68- TEST: OldMI  VALUE: 1 Num Cases: 1.0 Num Dsrd: 1.0
                            67- TEST: OldMI  VALUE: 0 Num Cases: 9.0 Num Dsrd: 1.0
                              ********PRUNED!!!
                              ********PRUNED!!!
                          59- TEST: age  VALUE: <=47 Num Cases: 2.0 Num Dsrd: 2.0
            13- TEST: prevang  VALUE: 0 Num Cases: 266.0 Num Dsrd: 146.0
              16- TEST: Duration  VALUE: >0 Num Cases: 259.0 Num Dsrd: 146.0
                20- TEST: post  VALUE: 1 Num Cases: 13.0 Num Dsrd: 2.0
                  26- TEST: Diabetes  VALUE: 1 Num Cases: 1.0 Num Dsrd: 1.0
                  25- TEST: Diabetes  VALUE: 0 Num Cases: 12.0 Num Dsrd: 1.0
                    ********PRUNED!!!
                    ********PRUNED!!!
                19- TEST: post  VALUE: 0 Num Cases: 246.0 Num Dsrd: 144.0
                  24- TEST: Nausea  VALUE: 0 Num Cases: 202.0 Num Dsrd: 105.0
                    32- TEST: OldMI  VALUE: 1 Num Cases: 13.0 Num Dsrd: 1.0
                      42- TEST: BP  VALUE: 1 Num Cases: 1.0 Num Dsrd: 1.0
                      41- TEST: BP  VALUE: 0 Num Cases: 12.0 Num Dsrd: .0
                    31- TEST: OldMI  VALUE: 0 Num Cases: 189.0 Num Dsrd: 104.0
                      40- TEST: age  VALUE: >37 Num Cases: 184.0 Num Dsrd: 103.0
                        48- TEST: Epis  VALUE: 1 Num Cases: 8.0 Num Dsrd: 2.0
                          58- TEST: Duration  VALUE: >8 Num Cases: 2.0 Num Dsrd: 2.0
                          57- TEST: Duration  VALUE: <=8 Num Cases: 6.0 Num Dsrd: .0
                        47- TEST: Epis  VALUE: 0 Num Cases: 176.0 Num Dsrd: 101.0
                          56- TEST: Duration  VALUE: >15 Num Cases: 2.0 Num Dsrd: .0
                          55- TEST: Duration  VALUE: <=15 Num Cases: 174.0 Num Dsrd: 101.0
                            66- TEST: Lipids  VALUE: 1 Num Cases: 1.0 Num Dsrd: 1.0
                            65- TEST: Lipids  VALUE: 0 Num Cases: 173.0 Num Dsrd: 100.0
                              76- TEST: Sweating  VALUE: 0 Num Cases: 73.0 Num Dsrd: 32.0
                                ********PRUNED!!!
                                ********PRUNED!!!
                              75- TEST: Sweating  VALUE: 1 Num Cases: 100.0 Num Dsrd: 68.0
                                88- TEST: Duration  VALUE: >8 Num Cases: 7.0 Num Dsrd: 2.0
                                  104- TEST: Rarm  VALUE: 0 Num Cases: 5.0 Num Dsrd: .0
                                  103- TEST: Rarm  VALUE: 1 Num Cases: 2.0 Num Dsrd: 2.0
                                87- TEST: Duration  VALUE: <=8 Num Cases: 93.0 Num Dsrd: 66.0
                                  ********PRUNED!!!
                                  ********PRUNED!!!
                      39- TEST: age  VALUE: <=37 Num Cases: 5.0 Num Dsrd: 1.0
                  23- TEST: Nausea  VALUE: 1 Num Cases: 44.0 Num Dsrd: 39.0
                    30- TEST: age  VALUE: >47 Num Cases: 41.0 Num Dsrd: 39.0
                      38- TEST: Duration  VALUE: >7 Num Cases: 7.0 Num Dsrd: 5.0
                        46- TEST: Larm  VALUE: 0 Num Cases: 1.0 Num Dsrd: .0
                        45- TEST: Larm  VALUE: 1 Num Cases: 6.0 Num Dsrd: 5.0
                          54- TEST: Rarm  VALUE: 0 Num Cases: 5.0 Num Dsrd: 5.0
                          53- TEST: Rarm  VALUE: 1 Num Cases: 1.0 Num Dsrd: .0
                      37- TEST: Duration  VALUE: <=7 Num Cases: 34.0 Num Dsrd: 34.0
                    29- TEST: age  VALUE: <=47 Num Cases: 3.0 Num Dsrd: .0
              15- TEST: Duration  VALUE: <=0 Num Cases: 7.0 Num Dsrd: .0
      7- TEST: age  VALUE: <=32 Num Cases: 3.0 Num Dsrd: .0
```

From perceptrons to CART, to multilayer perceptrons

Why?



“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

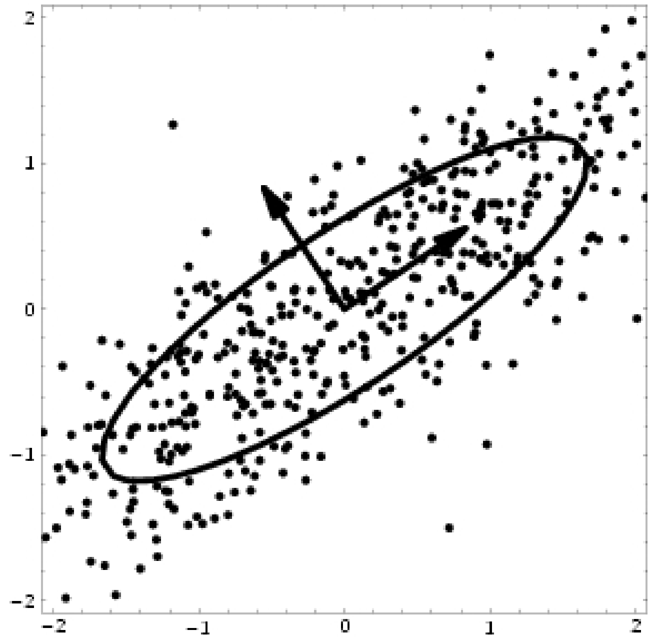
(a component is a linear combination of variables with specific coefficients)

•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

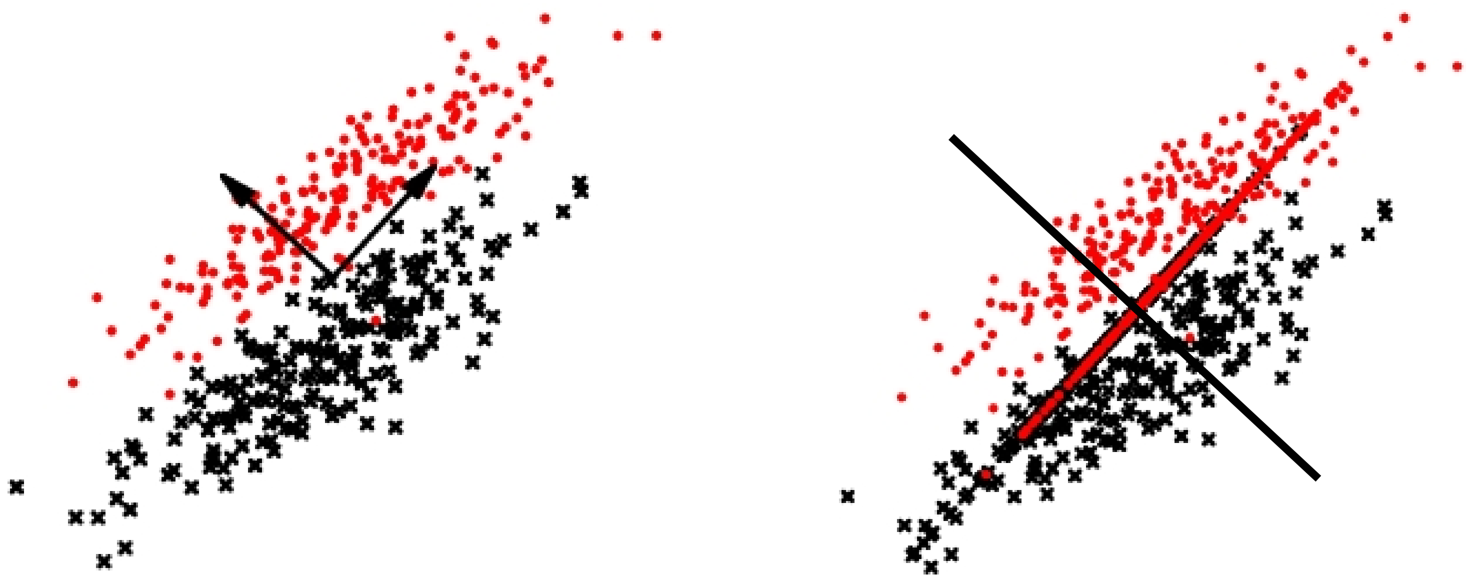
Principal Component Analysis

- 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



Related technique

- Partial Least Squares
  - PCA uses X to calculate directions of greater variation
  - PLS uses X and Y to calculate these directions
    - It is a variation of multiple linear regression

PCA maximizes $\mathrm{Var}(X\alpha)$,

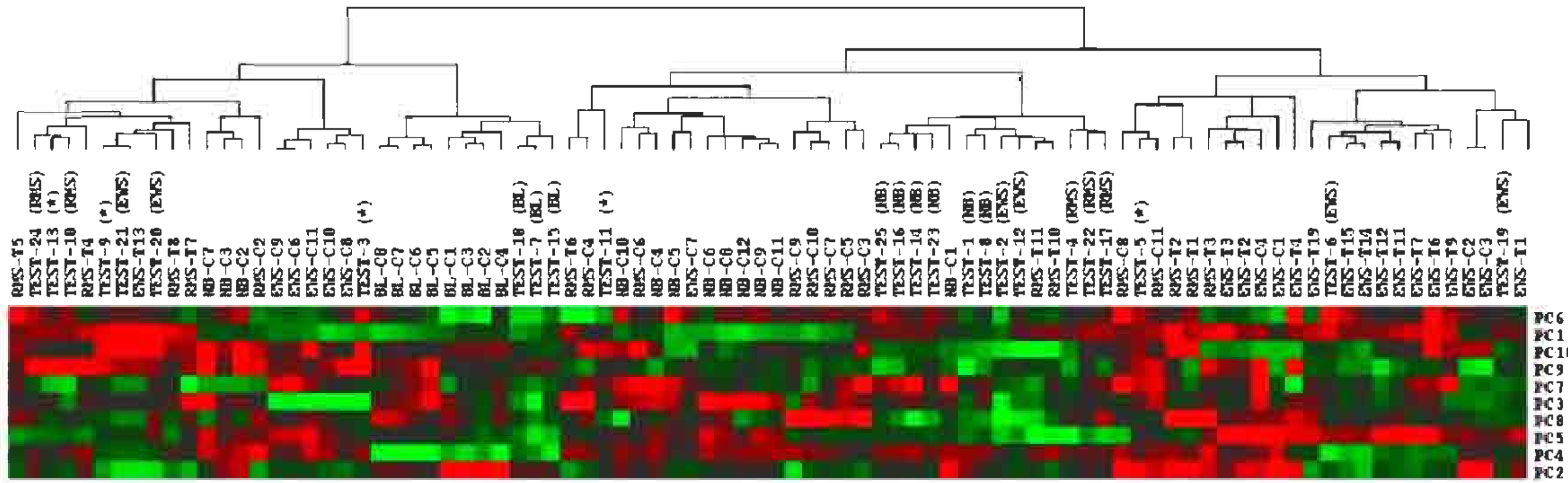
PLS maximizes $\mathrm{Corr}^2(y, X\alpha)\mathrm{Var}(X\alpha)$

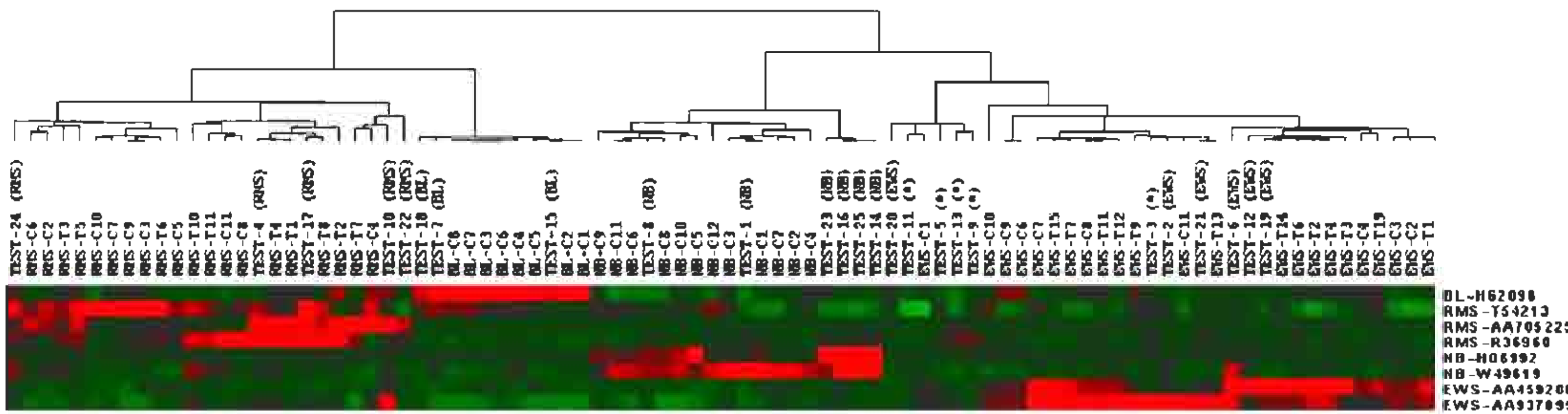
Variable Selection

- 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


PC6
PC1
PC10
PC9
PC7
PC3
PC8
PC5
PC4
PC2


BL-H62098
RMS-T54213
RMS-AA705225
RMS-R36960
NB-H06992
NB-W49619
EWS-AA459208
EWS-AA937895

Cross-validation

- Several training and test set pairs are created
- Results are pooled from all test sets

- “Leave-*n*-out”
- Jackknife (“Leave-1-out”)

Leave-N/3-out

1	23	54	0	1	1
2	43	23	1	0	1
3	34	35	0	0	0
4	20	21	1	1	1
5	19	03	1	1	0
6	78	04	0	1	0
7	98	03	0	1	1
8	35	05	1	1	1
9	99	23	0	0	1
10	23	34	0	0	0

Training Set

Model Building

Test Set

Evaluation

Bootstrap

- Efron (Stanford biostats) late 80's
  - "Pulling oneself up by one's bootstraps"
- Nonparametric approach to statistical inference
- Uses *computation* instead of traditional distributional assumptions and asymptotic results
- Can be used for non-linear statistics without known standard error formulas

Sample with Replacement

Sample	Y_1*	Y_2*	Y_3*	Y_4*	$\bar{Y}$*
1	6	6	6	6	6.00
2	6	6	6	-3	3.75
3	6	6	6	5	5.75
..					
100	-3	5	6	3	2.75
101	-3	5	-3	6	1.25
…					
255	-3	3	3	5	3.5
256	3	3	3	3	3.00

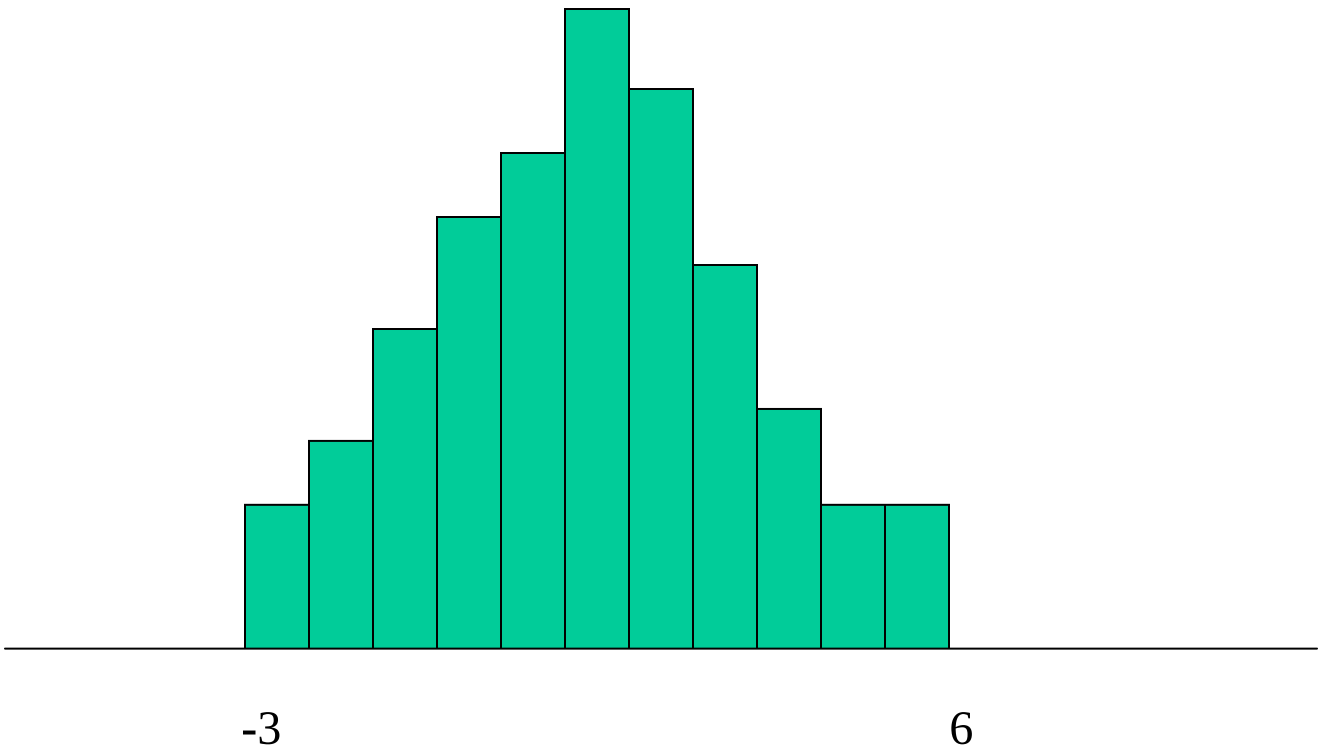
The population is to the sample

as

the sample is to the bootstrap samples

In practice (as opposed to previous example),
not all bootstrap samples are selected

Empirical distribution of Y



Bootstrap Confidence Intervals

- Percentile Intervals

  Example

  – 95% CI is calculated by taking

  – Lower = 0.025 x bootstrap replicates

  – Upper = 0.975 x bootstrap replicates

Bagging

- Breiman, 1996
- Derived from bootstrap (Efron, 1993)
- Create classifiers using training sets that are bootstrapped (drawn with replacement)
- Average results for each case

Boosting

- A family of methods
- Sequential production of classifiers
- Each classifier is dependent on the previous one, and focuses on the previous one’s errors
- Examples that are incorrectly predicted in previous classifiers are chosen more often or weighted more heavily

Visualization

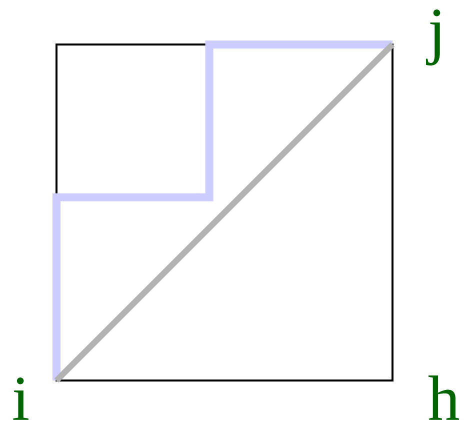
- Capabilities of predictive models in this area are limited
- Clustering is often good for visualization, but it is generally not very useful to separate data into pre-defined categories
  – Hierarchical trees
  – 2-D or 3-D multidimensional scaling plots
  – Self-organizing maps

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

Metric spaces

- Positivity Reflexivity $$d_{ij} > d_{ii} = 0$$
- Symmetry $$d_{ij} = d_{ji}$$
- Triangle inequality $$d_{ij} \le d_{ih} + d_{hj}$$



Figures removed due to copyright reasons.

Please see:

Khan, J., et. al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." ***Nat Med*** 7, no. 6 (June 2001): 673-9.

k-means clustering (Lloyd's algorithm)

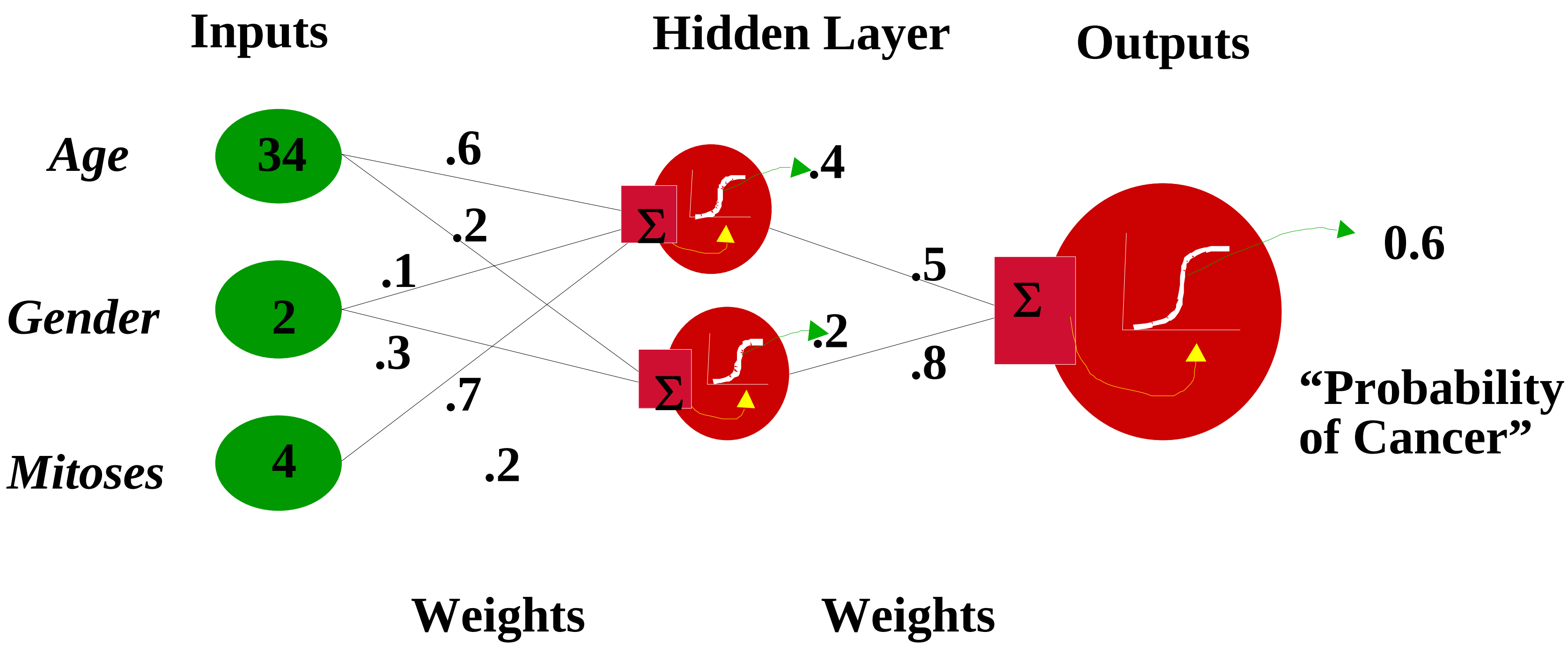
1. Select k (number of clusters)
2. Select k initial cluster centers $c_1,\ldots,c_k$
3. Iterate until convergence: For each i,
    1. Determine data vectors $v_{i1},\ldots,v_{in}$ closest to c_i (i.e., partition space)
    2. Update c_i as $c_i = 1/n\ (v_{i1}+\ldots+v_{in})$

Figures removed due to copyright reasons.

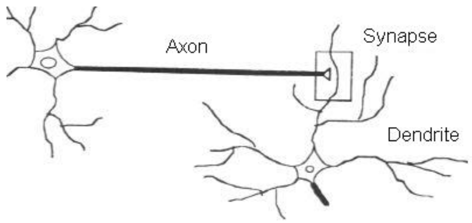
Please see:

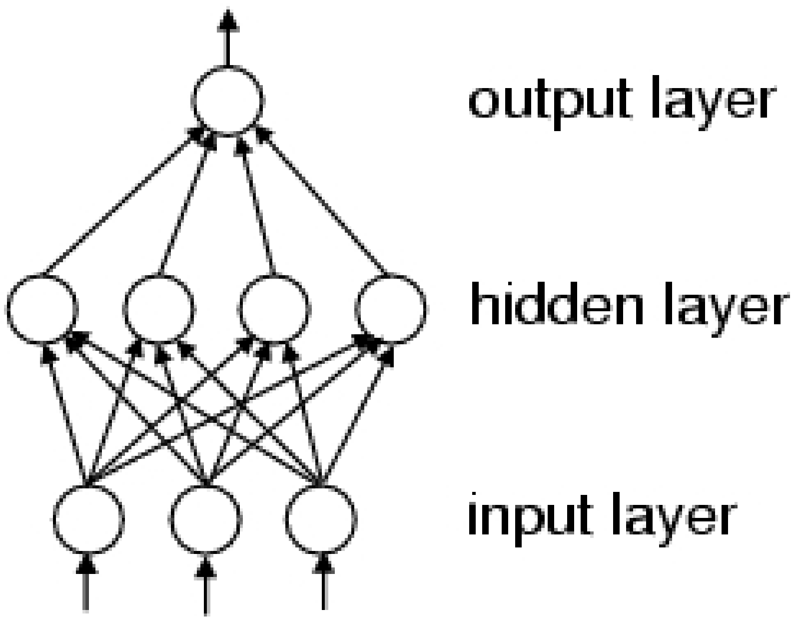
Khan, J., et. al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." ***Nat Med*** 7, no. 6 (June 2001): 673-9.

Neural Networks



Neural Networks





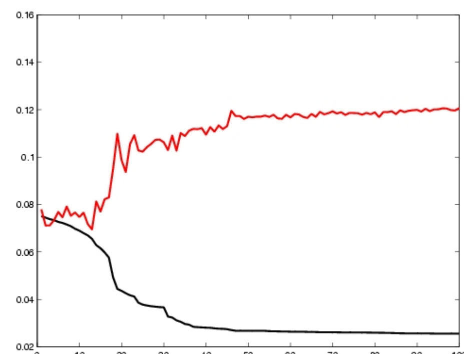
Work well even with non-linearly separable data

Overfitting control:

•Few weights

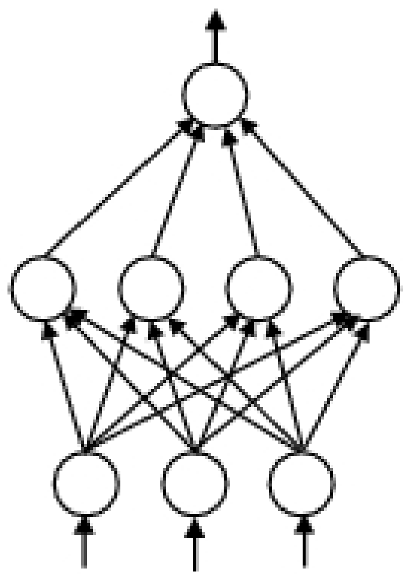
•Little training

•Penalty for large weights



Backpropagation algorithm

Classification	Regression
cross-entropy	sum-of-squares
sigmoidal neuron	linear neuron
sigmoidal neurons	sigmoidal neurons
linear neurons	linear neurons



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