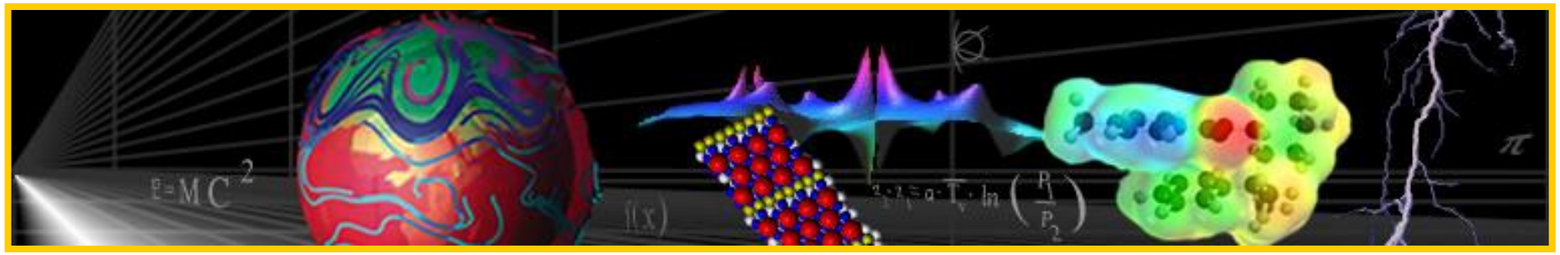
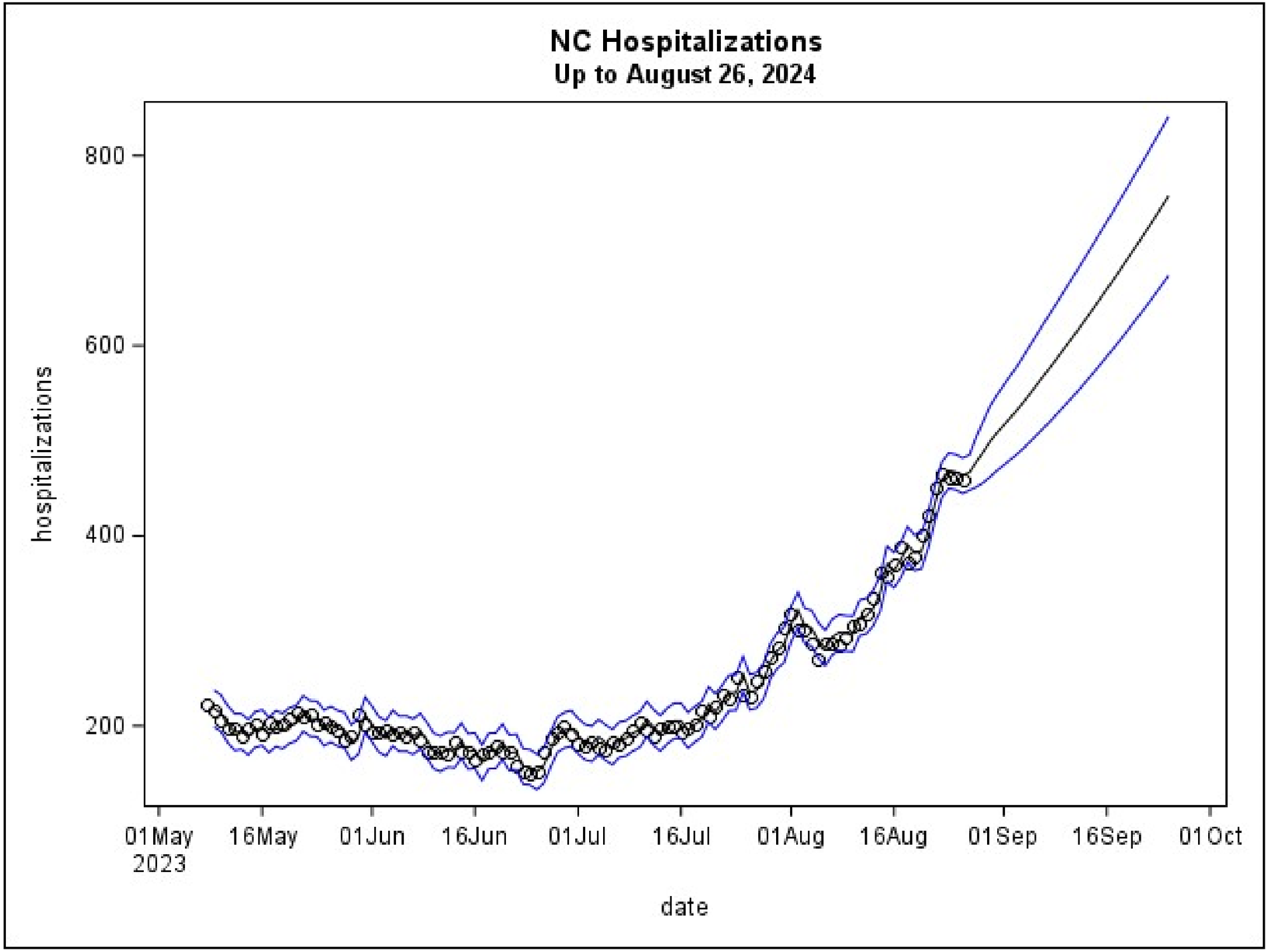
# An Overview of Data Mining

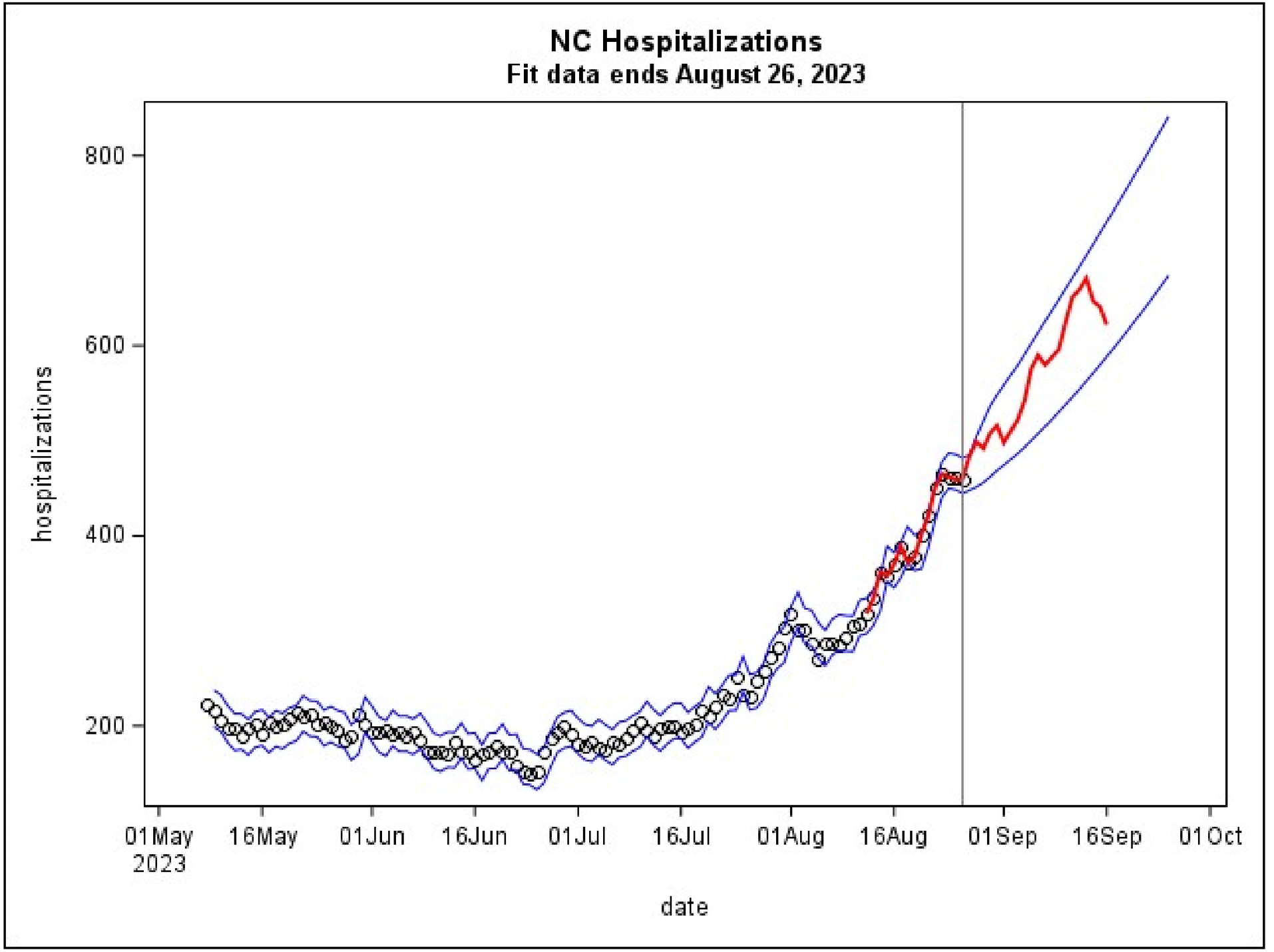



D. A. Dickey

IAA

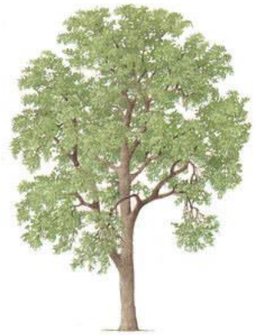
September 2023


NC Hospitalizations
Up to August 26, 2024
800
600
400
200
hospitalizations
01May
2023
16May
01Jun
16Jun
01Jul
16Jul
01Aug
16Aug
01Sep
16Sep
01Oct
date


NC Hospitalizations
Fit data ends August 26, 2023
hospitalizations
800
600
400
200
01May
2023
16May
01Jun
16Jun
01Jul
16Jul
01Aug
16Aug
01Sep
16Sep
01Oct
date

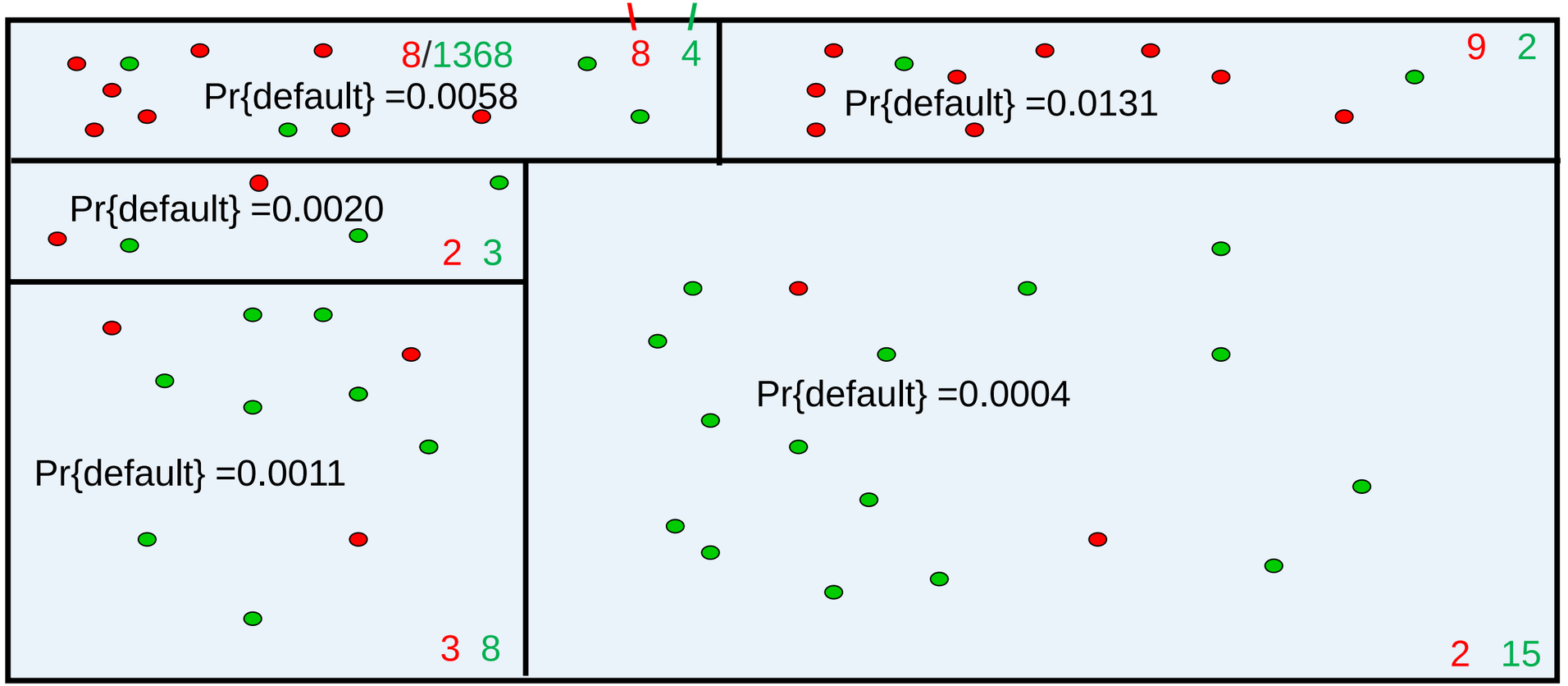
# Decision Trees

- A “divisive” method (splits)
- Start with “root node” – all in one group
- Get splitting rules
- Response often binary
- Result is a “tree”
- Example: Loan Defaults
- Example: Framingham Heart Study
- Example: Covid vaccine %
- Optional: Automobile Accidents



# Recursive Splitting

Representing 8 defaulters, 4*340=1360 nondefaulters




X1=Debt To Income Ratio

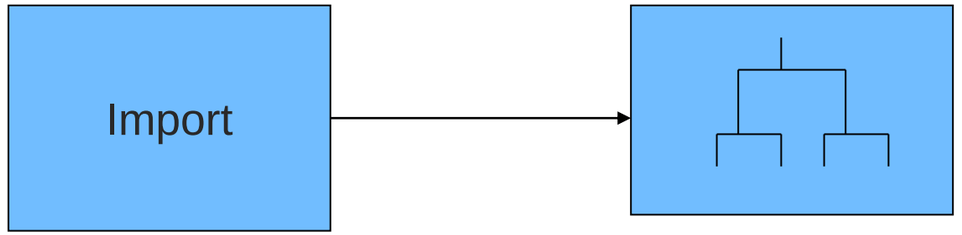
X2 = Age

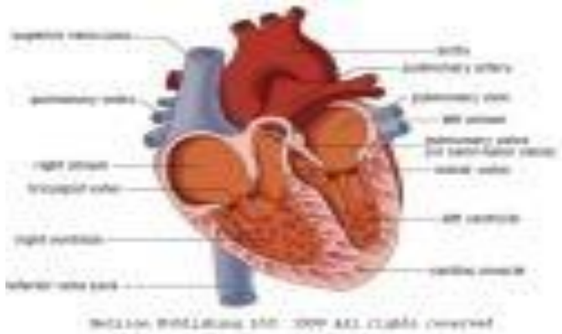
No default (1 dot per 340)

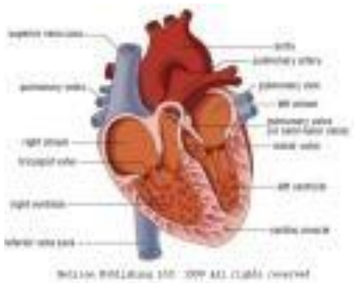
Default (1 dot per defaulter)

# Some Actual Data

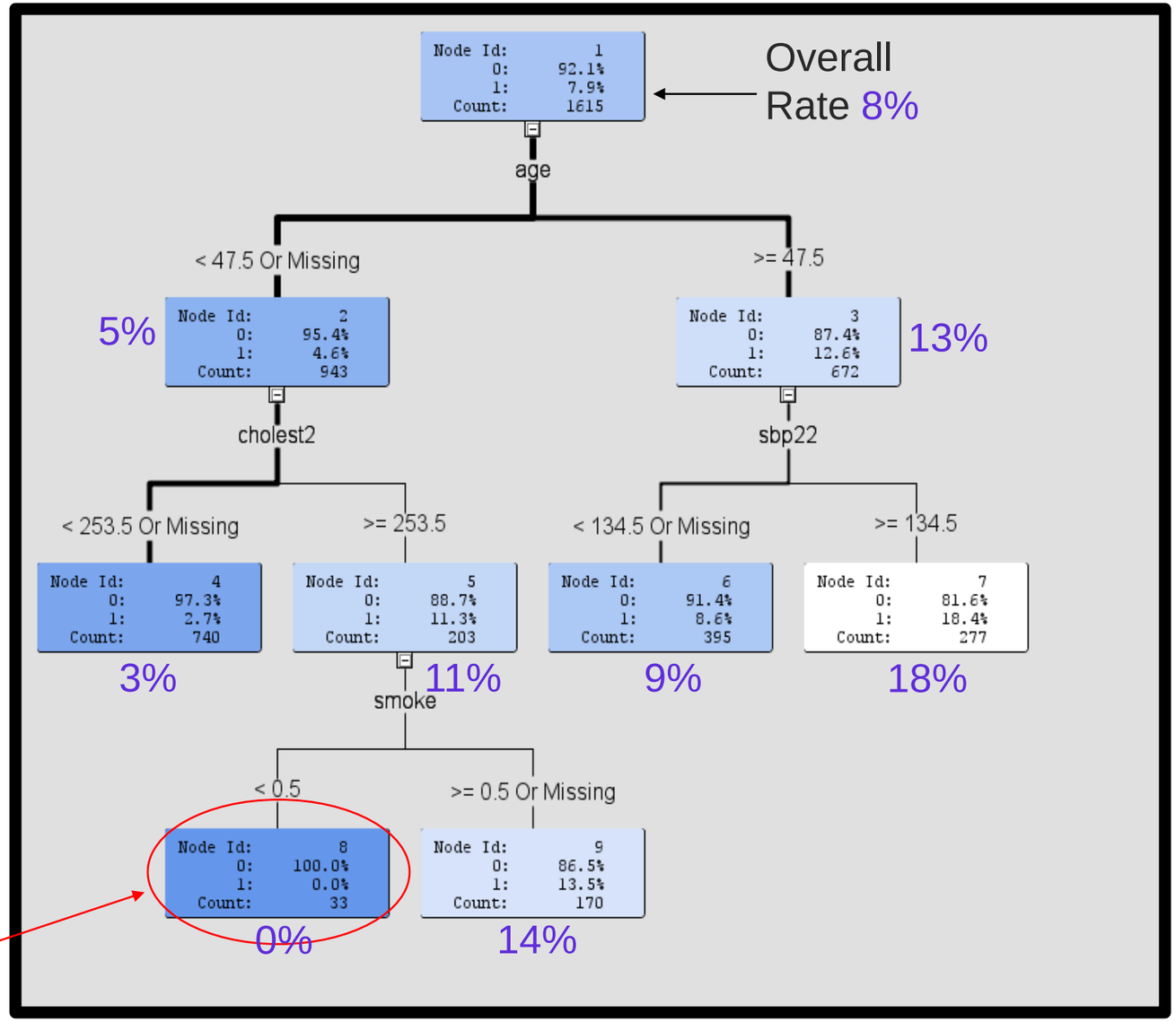
- Framingham Heart Study
- First Stage Coronary Heart Disease
  - P{CHD} = Function of:
    - » Age  - no drug yet! ☹
    - » Cholesterol
    - » Systolic BP
    - » Smoking





# Example of a “tree” → Using Default Settings




No first stage Coronary Heart Disease (n=33)

Narrowest possible 95% confidence interval is 0 < Pr{0 out of 33} < 0.0868 (for *random* sample of 33 with no events)

# How to make splits?

- Contingency tables

180 ?
240?

| | Heart Disease: No | Heart Disease: Yes | |
|---|---|---|---|
| Low BP | 95 | 5 | 100 |
| High BP | 55 | 45 | 100 |
| | 150 | 50 | |

DEPENDENT (effect)

| | Heart Disease: No | Heart Disease: Yes | |
|---|---|---|---|
| Low BP | 75 | 25 | 100 |
| High BP | 75 | 25 | 100 |
| | 150 | 50 | |

INDEPENDENT (no effect)

# How to make splits?

- Contingency tables

180 ? 240?

| | Heart Disease: No | Heart Disease: Yes | |
|---|---|---|---|
| Low BP | 95 / 75 | 5 / 25 | 100 |
| High BP | 55 / 75 | 45 / 25 | 100 |
| | 150 | 50 | |

DEPENDENT (effect)

$$\chi^2 = \sum_{allcells} \frac{(\text{Observed - Expected})^2}{\text{Expected}} =$$

2(400/75)+ 2(400/25) = 42.67

Compare to $\chi^2$ tables –> P=0.00000000064

$H_0$: No association

$H_1$: BP and heart disease *are* associated

# How to make splits?

- Which variable to use?
- Where to split?
  - Cholesterol > ____
  - Systolic BP > _____
- Idea – Pick BP cutoff <u>to minimize p-value</u> for $\chi^2$
- Split point data-derived!
- What does "<u>significance</u>" mean now?
- Fix (Bonferroni): (p-value)x(number of splits) < 0.05

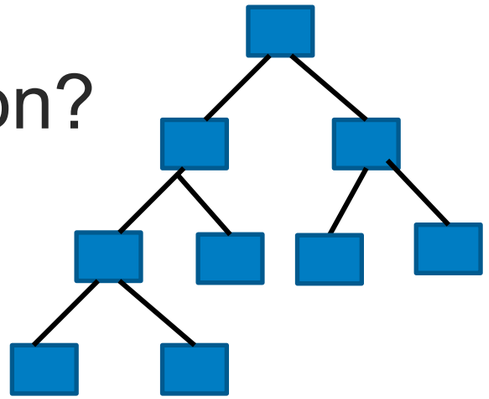
Carlo Emilio Bonferroni

# Validation


- Traditional stats – small dataset, need all observations to estimate parameters of interest.
- Data mining – loads of data, can afford “holdout sample”
- Variation: n-fold cross validation
  - Randomly divide data into n sets
  - Estimate on n-1, validate on 1
  - Repeat n times, using each set as holdout.
  - Framingham example did not use holdout.

# Pruning

- Grow bushy tree on the “fit data”
- Classify validation (holdout) data
- Likely farthest out branches do not improve, possibly hurt fit on validation data
- Prune non-helpful branches.
- What is “helpful”?  What is good discriminator criterion?

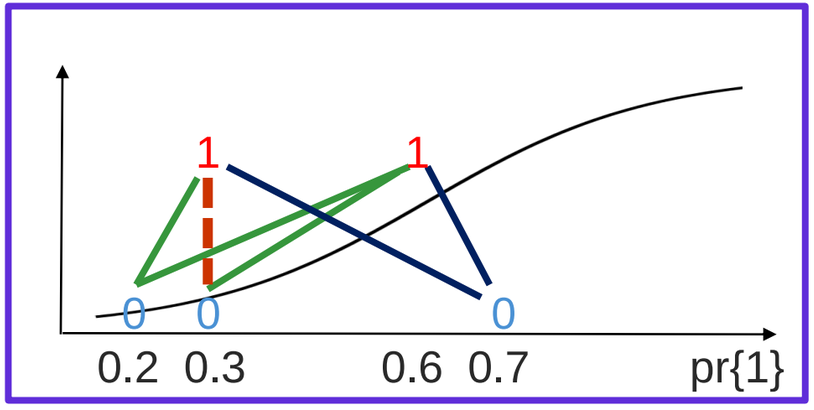


# Goals

- Split (or keep split) if diversity in parent “node” > summed diversities in child nodes
- Prune to optimize
  - Estimates
  - Decisions
  - Ranking
- in validation data

# Assessment for:

- Decisions  Minimize incorrect model decisions (versus realized)
- Estimates  Error Mean Square (average squared error)
- Ranking C (concordance) statistic = proportion concordant + ½ (proportion  tied)
  - Obs number  1  2  3  4  5
  - Actual  0  (0,1)  1  0  (1 means “event”)
  - Probability of 1  0.2  0.3  0.6  0.7  (← from model)
    - » Concordant Pairs: (1,3)  (1,4)  (2,4)
    - » Discordant  Pairs: (3,5) (4,5)
    - » Tied (2,3)
    - » 6 ways to get pair with 2 different responses
  - C = 3/6 + (1/2)(1/6) = 7/12=0.5833= ROC area



# Accounting for Costs

- Pardon me (sir, ma’am) can you spare some change?
- Say “sir” and prefers male  +$2.00
- Say “ma’am” and prefers female +$5.00
- Say “sir” but prefers female -$1.00 (balm for slapped face)
- Say “ma’am” but prefers male -$10.00 (nose splint)

# Including Probabilities

Leaf has Pr(**M**)=**.4**, Pr(**F**)=**.6**

You say:

| Actual Preference | Sir | Ma'am |
|---|---|---|
| **M 40%** | **0.4** (+2) | **0.4** (**-10**) |
| **F 60%** | **0.6** (**-1**) | **0.6** (+5) |
| | +$0.20 | –$1.00 |

Expected profit is
2(0.4)-1(0.6) = $0.20
if I say "sir"

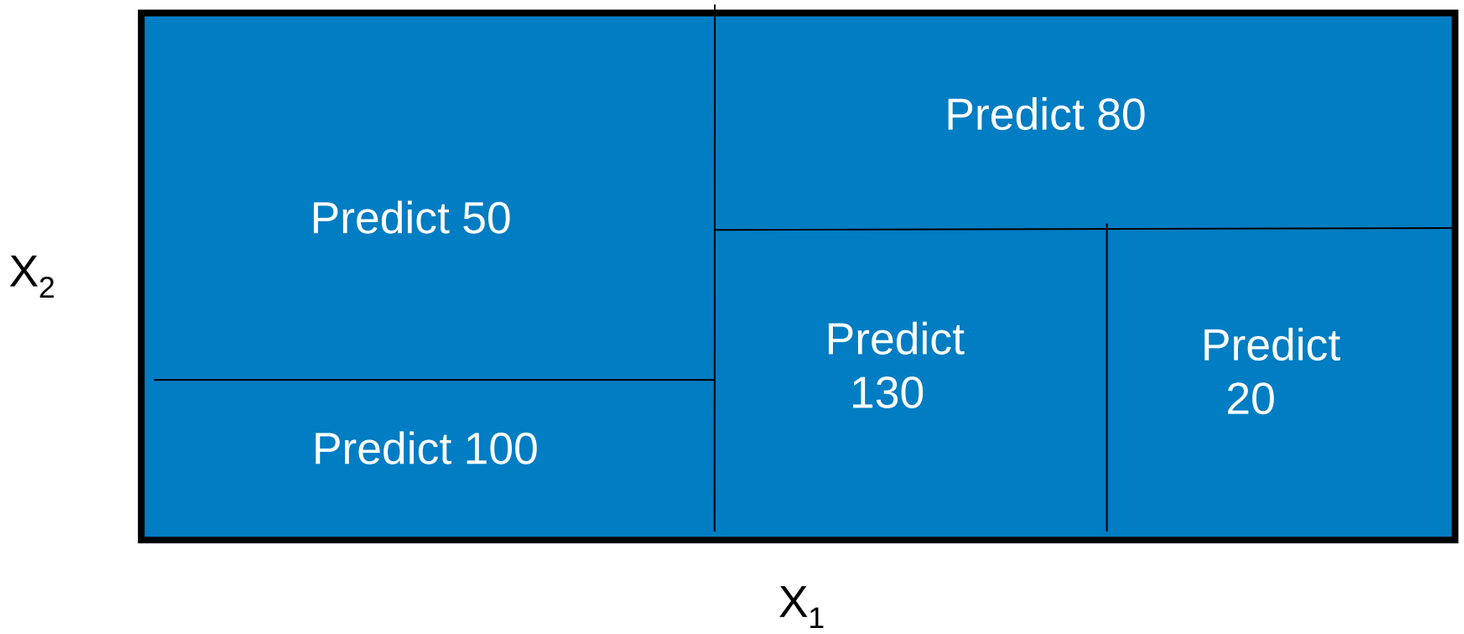
Expected profit is
-4+3 = -$1.00 (a loss)
if I say "Ma'am"

Weight leaf profits by leaf size (# obsns.) and sum.

Prune (and split) to maximize profits.

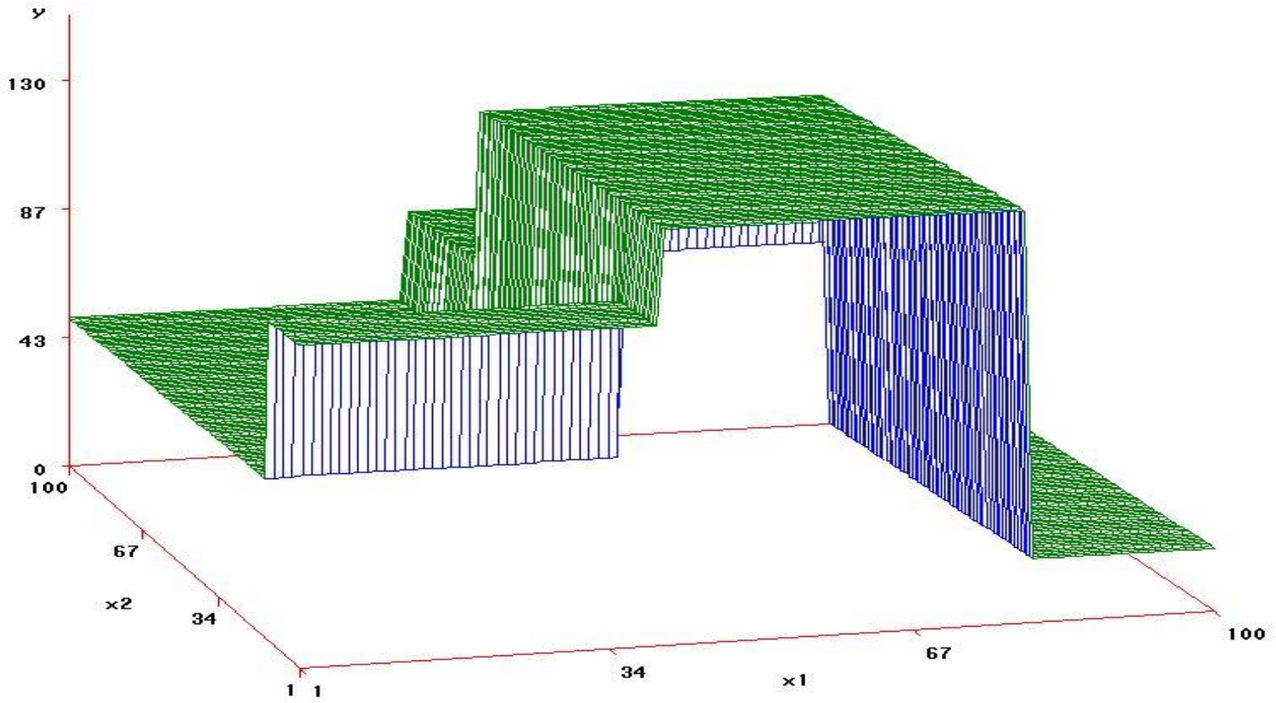
# Regression Trees

- *Continuous* response Y
- Predicted response $P_i$ constant in regions i=1, …, 5



# Regression Trees

- Predict $P_i$ in cell i.
- $Y_{ij}$ $j^{th}$ response in cell i.
- Split to minimize $\Sigma_i \Sigma_j (Y_{ij}-P_i)^2$



# % Vaccinated by State


Results - Node: Decision Tree Diag

File Edit View Window

Tree

Node Id: 1
Average: 53.3765
Count: 51

←SSE = 3295

rep2dem_ratio

< 0.9007

>= 0.9007 Or Missing

Node Id: 2
Average: 60.4618
Count: 22

Node Id: 3
Average: 48.0014
Count: 29

evangelical_pct

bachelor_deg

< 14.5

>= 14.5 Or Missing

< 29.205

>= 29.205 Or Missing

Node Id: 4
Average: 65.0311
Count: 9

Node Id: 5
Average: 57.2985
Count: 13

Node Id: 6
Average: 43.9636
Count: 11

Node Id: 7
Average: 50.4689
Count: 18

gov_age

republican_gov

< 62.5 Or Missing

>= 62.5

< 0.5

>= 0.5 Or Missing

Node Id: 8
Average: 45.9800
Count: 6

Node Id: 9
Average: 41.5440
Count: 5

Node Id: 10
Average: 54.6160
Count: 5

Node Id: 11
Average: 48.8738
Count: 13

$\sum_{\text{Leaf=1}}^{6}$ (SSEs) = 572

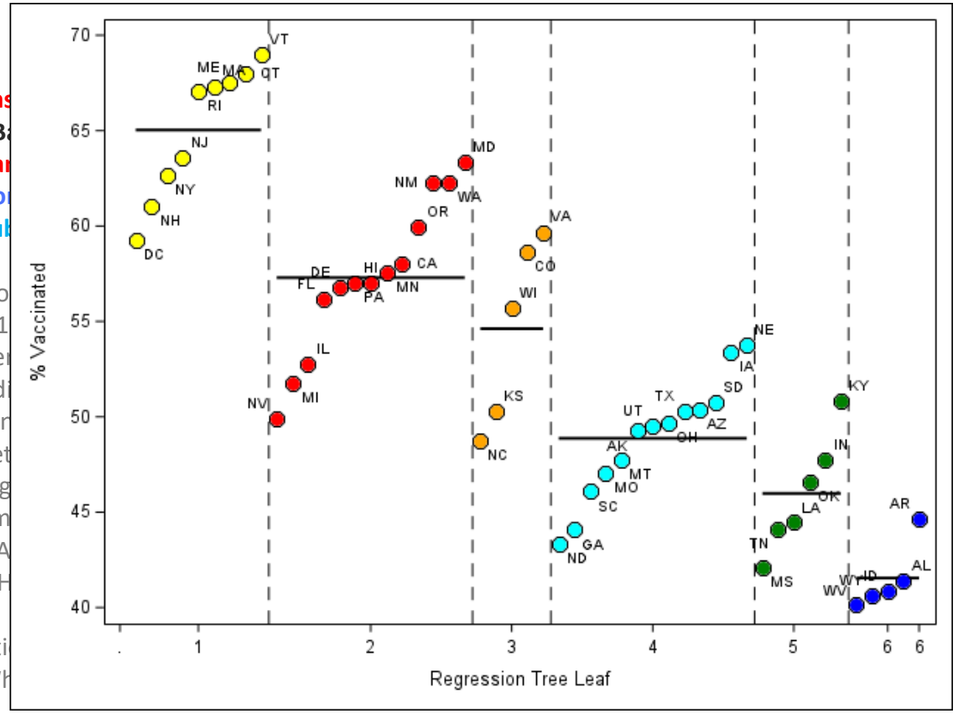
statecode

**vaxpct** **Target**

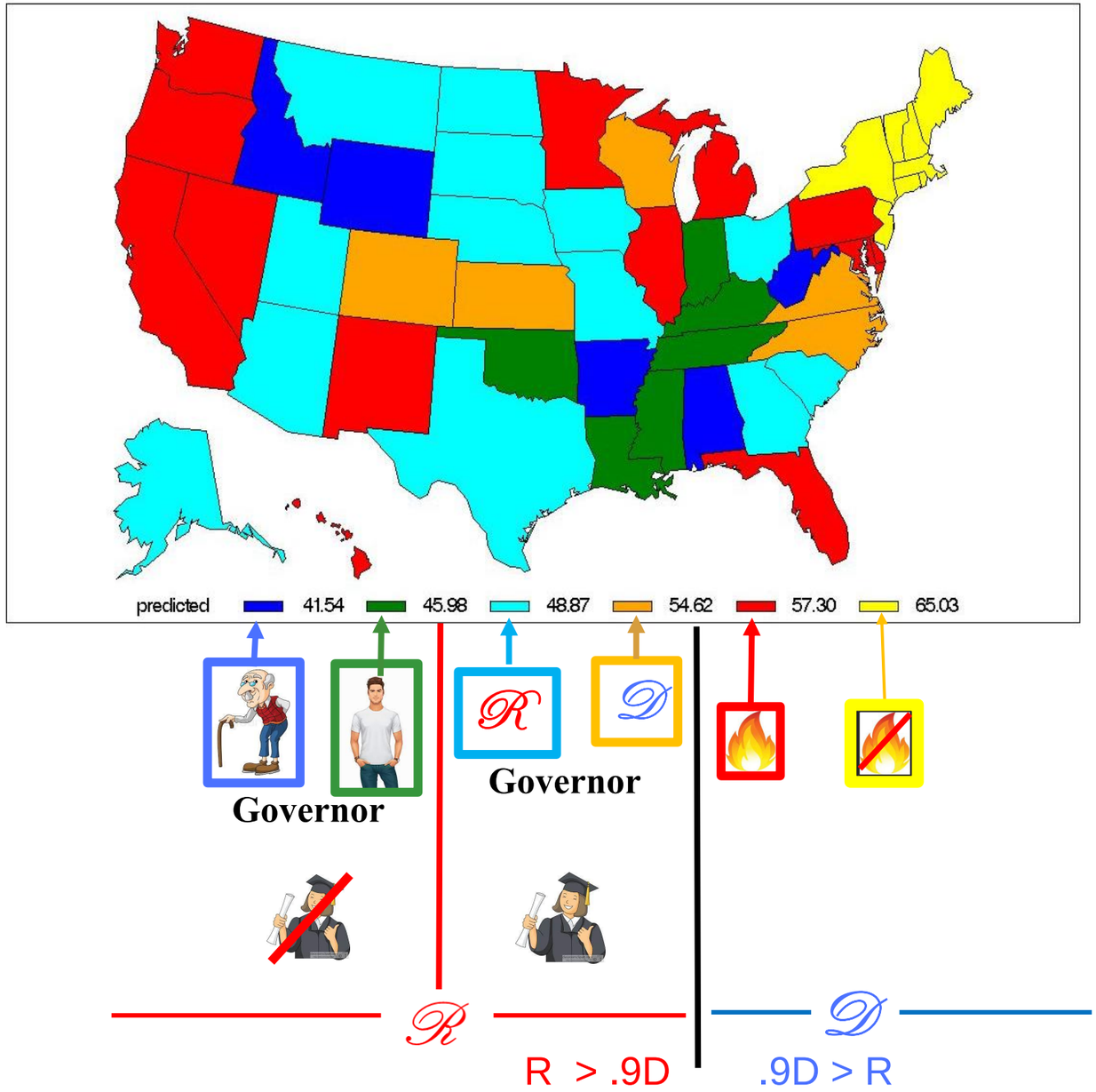
**rep2dem_ratio** **(leans**
**bachelor_deg** **% w Ba**
**evangelical_pct** **% Eva**
**gov_age** **Governo**
**republican_gov** **Repub**

catholic_pct % Catho
Birth Births per 1
Death Deaths per
Med_Hhold_Inc Medi
MedianAge Median
Methodist_pct % Met
US_IO Net immig
World_IO Net imm
advanced_deg % w A
highschool_deg % w H
male_pct % Male
population Populati
white_pct_2012 % Wh

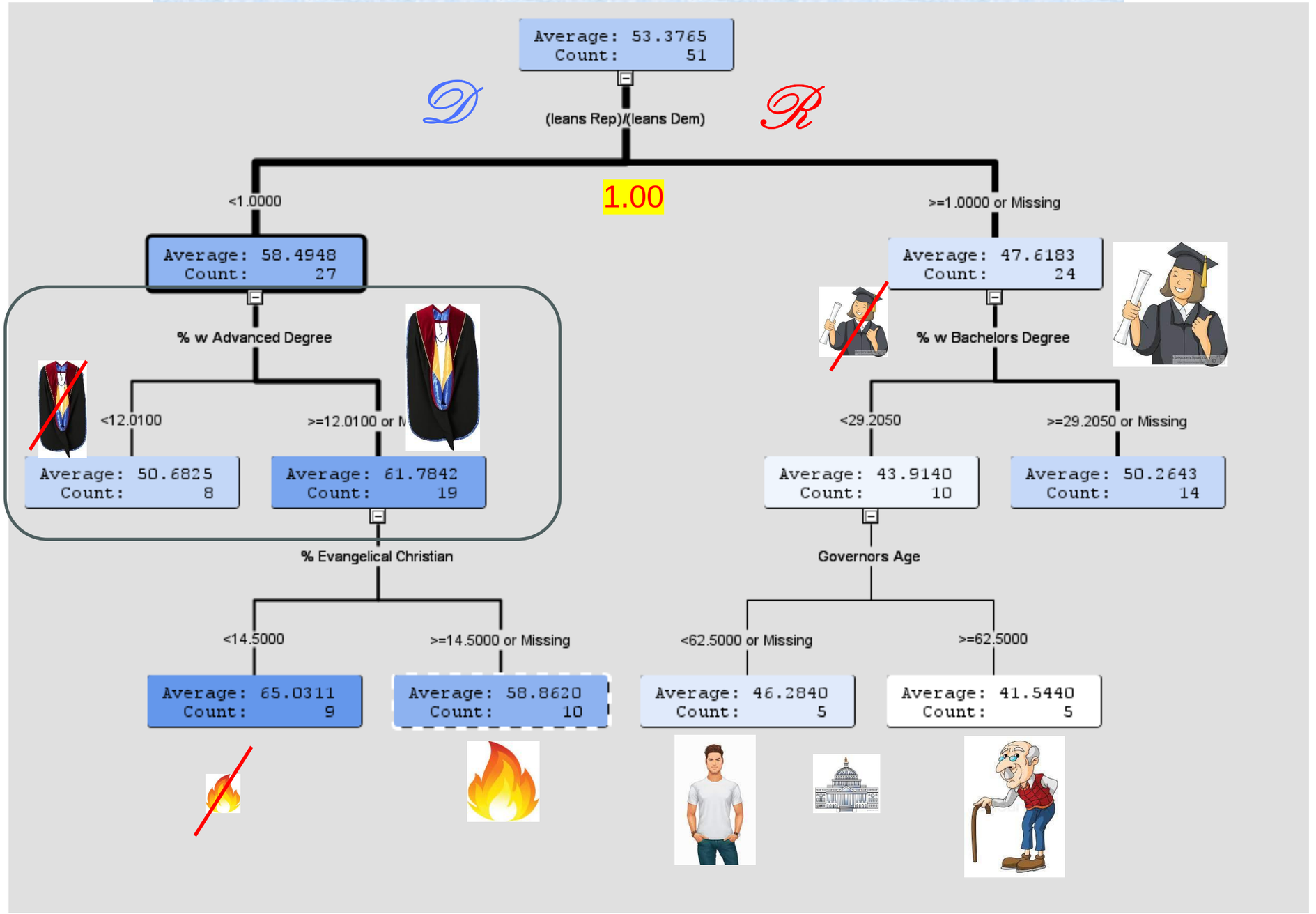
# Actual and Predicted vaccination %



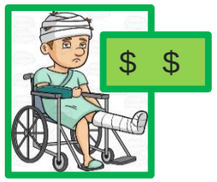

# Predicted vaccination %



# Interactive – force first split at R/D=1

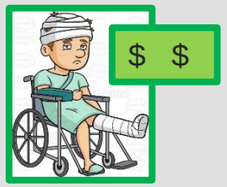



# Real data example: Traffic accidents in Portugal*




Y = injury induced "cost to society" (in multiples of cost of 1 death)

According to the NSC, the cost per death is
D = $1.41 million







* Tree developed by Guilhermina Torrao, (used with permission)
NCSU Institute for Transportation Research & Education

Help - I ran Into a "tree"

Help - I ran Into a "tree"

Extensions of tree based methods:

(1) Boosting - build a tree, take residuals. Build a second tree on the residuals from first to fix mistakes. Take residuals, build a new tree etc.

(2) Random Forests – sample from the observations and variables.
Build multiple trees, one for each sample
Combine information from all trees.
(i.e. use an ensemble model)

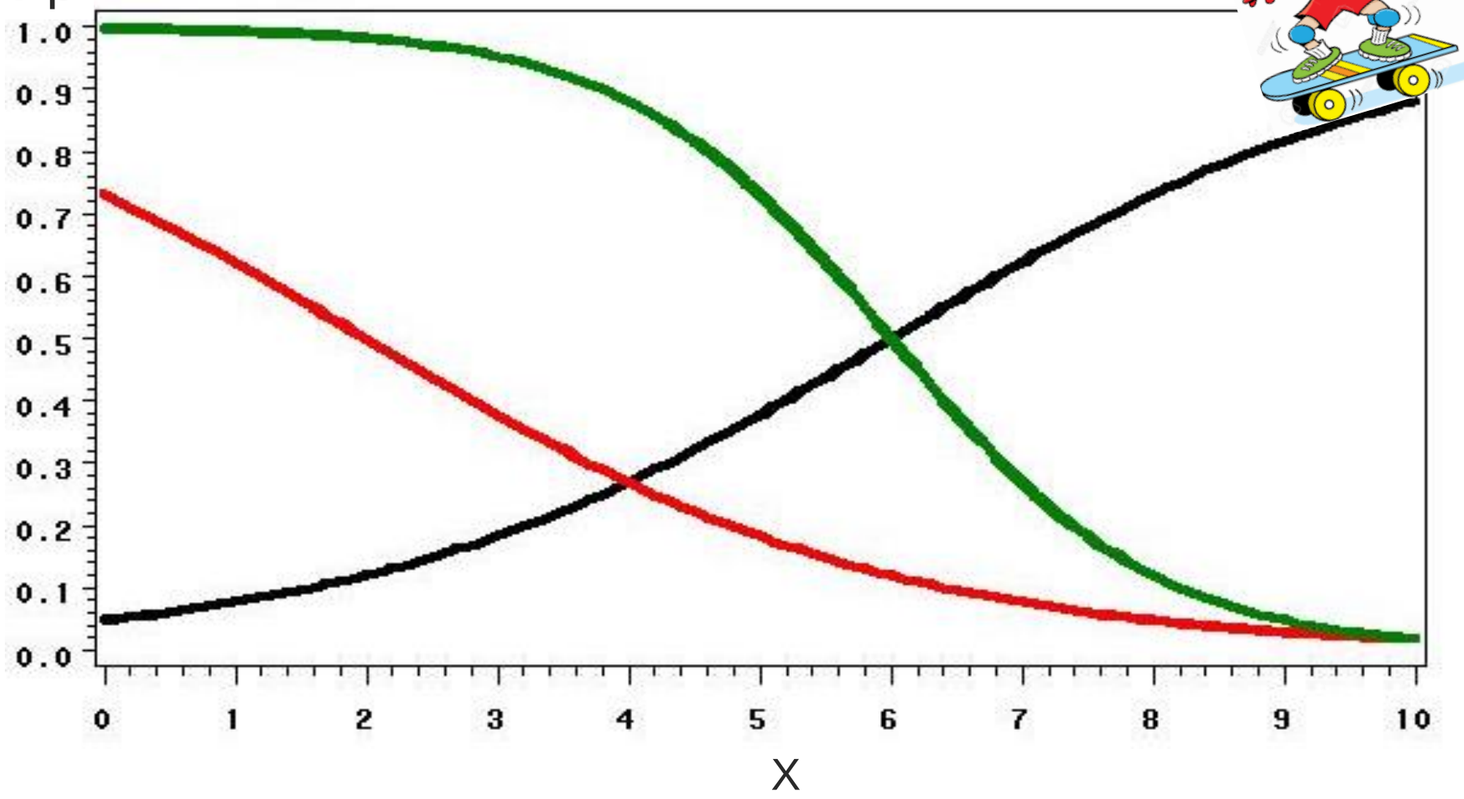
# Logistic Regression

- Logistic – another classifier
- Older – “tried & true” method
- Predict probability of response from input variables (“Features”)
- Linear regression gives infinite range of predictions
- 0 < probability < 1 so not linear regression.

# Logistic Regression

$$\frac{e^{a+bX}}{(1+e^{a+bX})} \rightarrow p$$

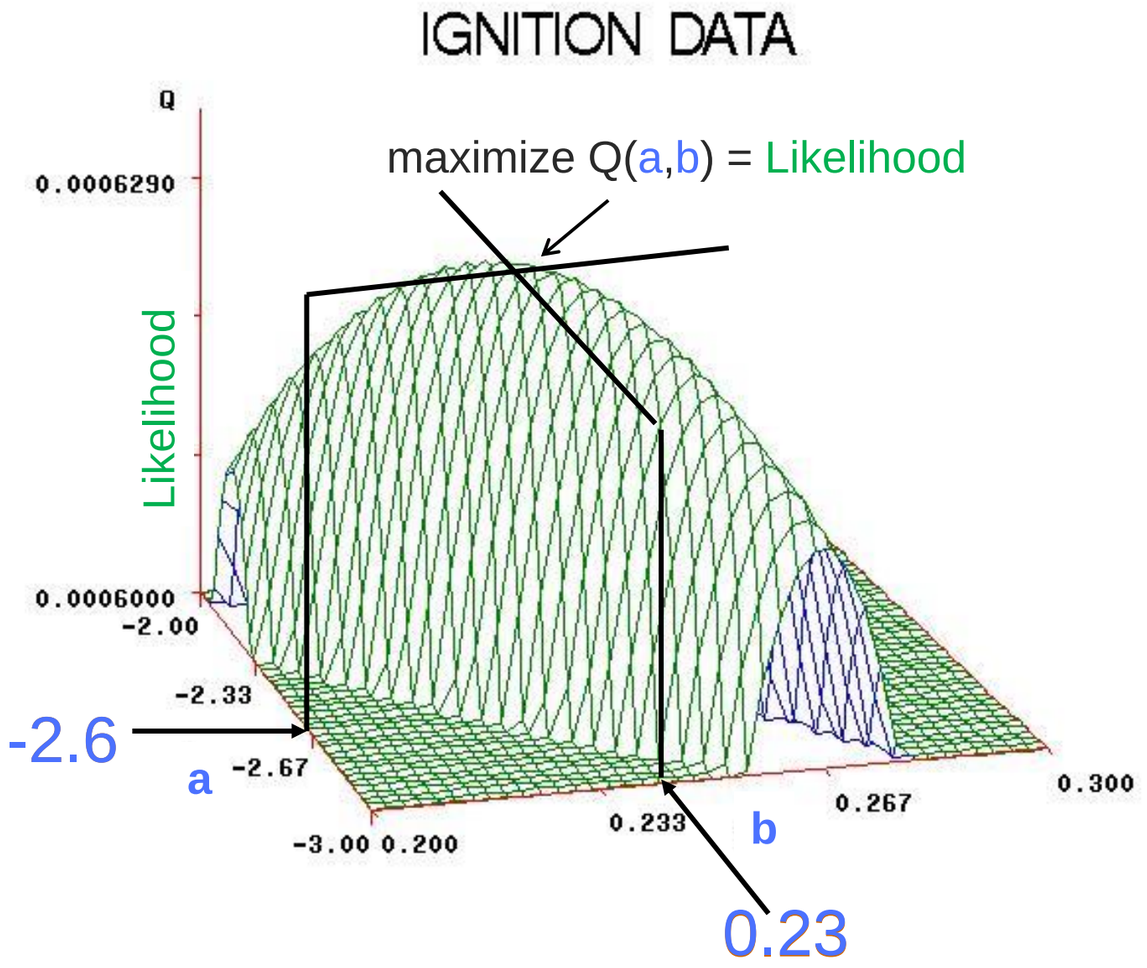
## Three Logistic Curves



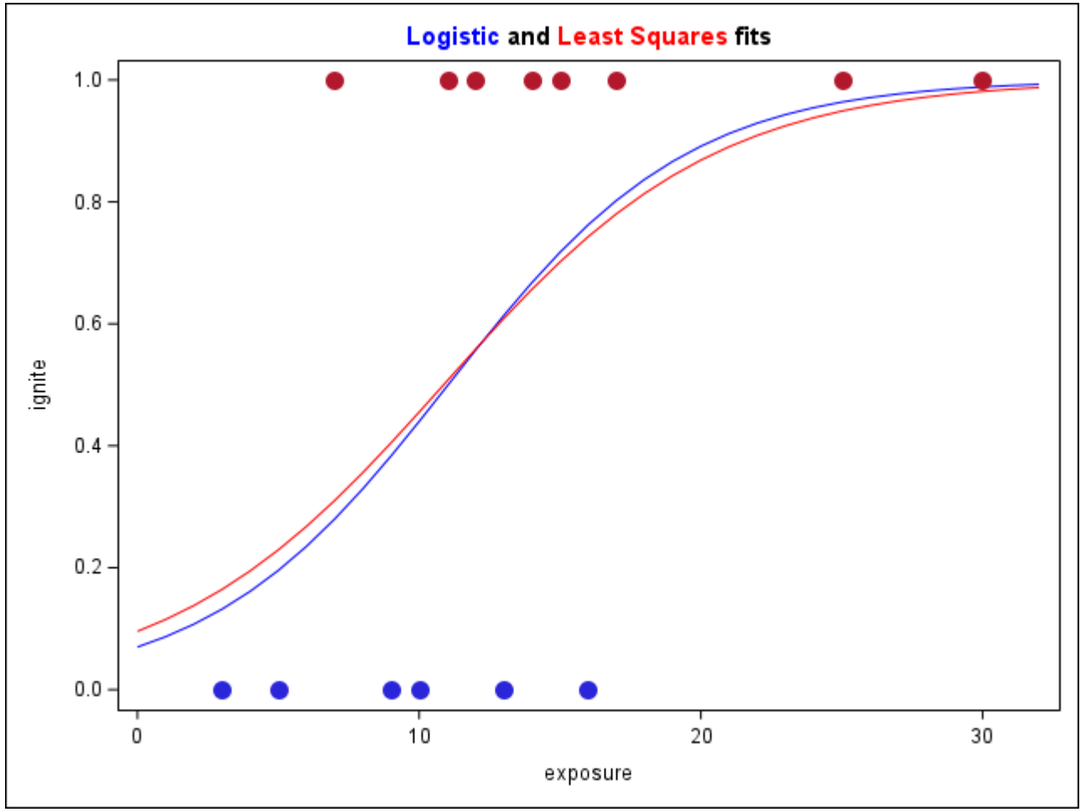

# Example: Seat Fabric Ignition

- Flame exposure time = X
- Y=1 → ignited,  Y=0→ did not ignite
  - Y=0, X= 3, 5, 9 10 , 13, 16
  - Y=1, X = 7 11, 12 14, 15, 17, 25, 30
- $Q=(1-p_1)(1-p_2)p_3(1-p_4)(1-p_5)p_6p_7(1-p_8)p_9p_{10}(1-p_{11})p_{12}p_{13}p_{14}$
- p's all different $p_i = f(a+bX_i) = e^{a+bXi}/(1+e^{a+bXi})$
- Find a,b to maximize Q(a,b)

- Logistic idea:
- Given temperature X, compute L(x)=a+bX then $p = e^{L}/(1+e^{L})$
- $p(i) = e^{a+bX_i}/(1+e^{a+bX_i})$
- Write p(i) if event, 1-p(i) if not
- Multiply all n of these together, find a,b to maximize this “likelihood”
- Logistic: L = -2.59 + 0.23X

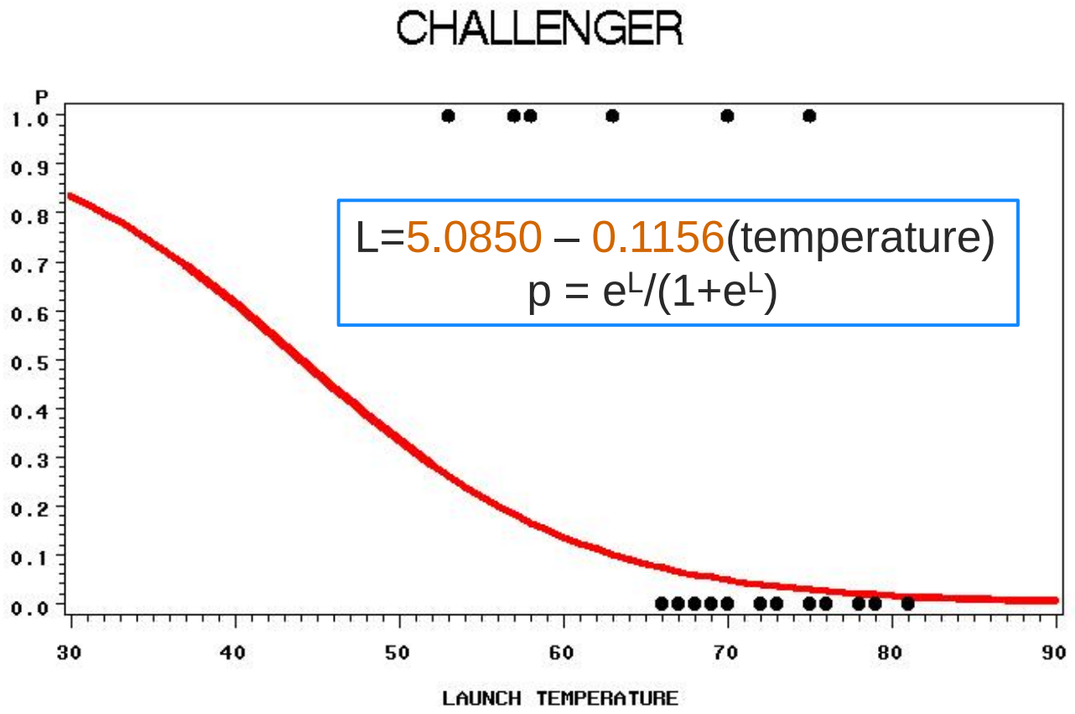


- Least squares idea:
- Given temperature X, compute L(x)=a+bX then p = $e^L/(1+e^L)$+error
- p(i) = $e^{a+bXi}/(1+e^{a+bXi})$
- Error = 1-p(i) if event, 0-p(i) if not
- Square errors and sum, find a,b to minimize this "SS(error)"
- Estimated L = -2.25+ 0.21X
- Versus L = -2.59 + 0.23X

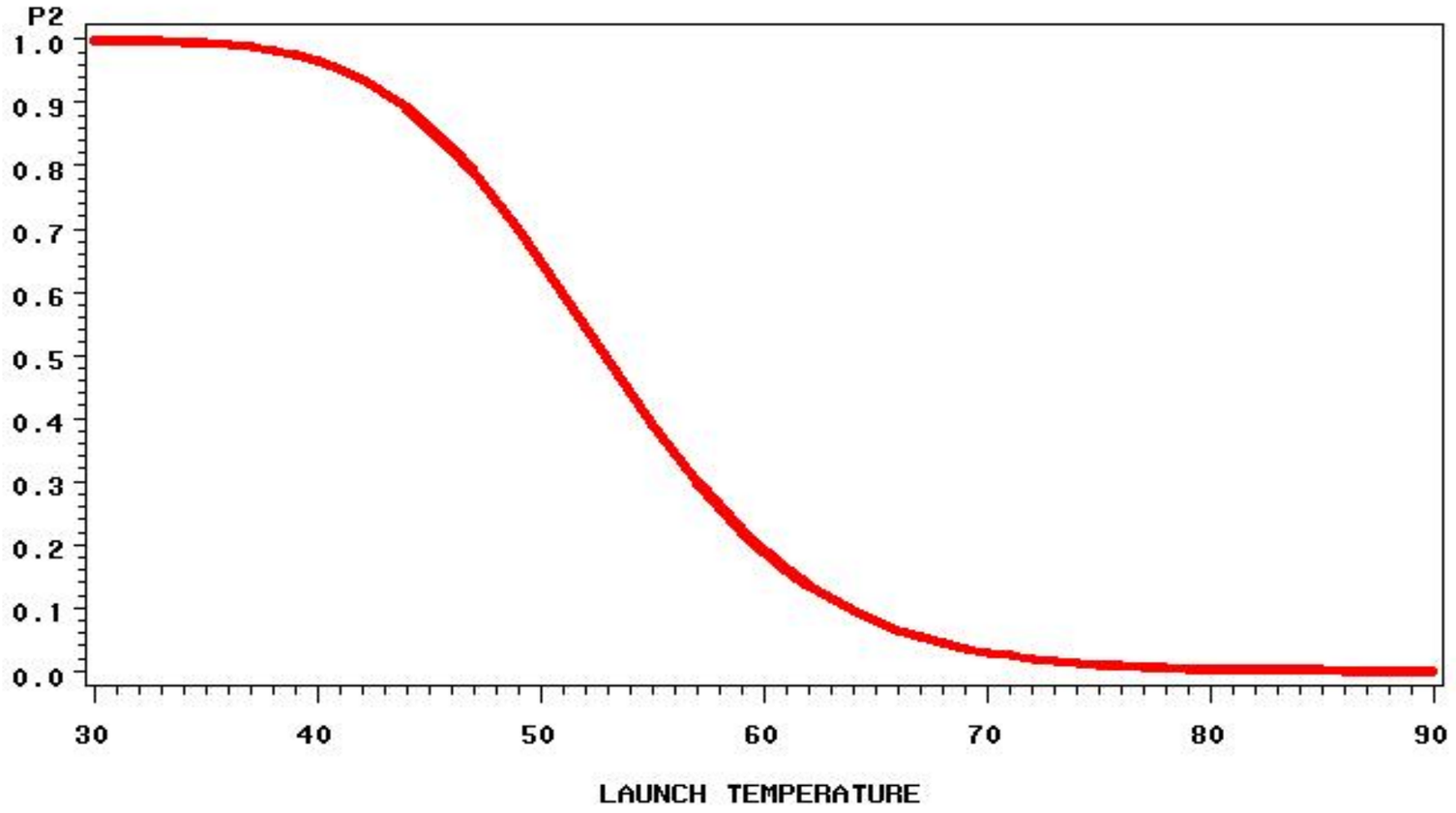


# Example: Shuttle Missions

- O-rings failed in Challenger disaster
- Prior flights “erosion” and “blowby” in O-rings (6 per mission)
- Feature: Temperature at liftoff
- Target: (1) - erosion or blowby vs. no problem (0)



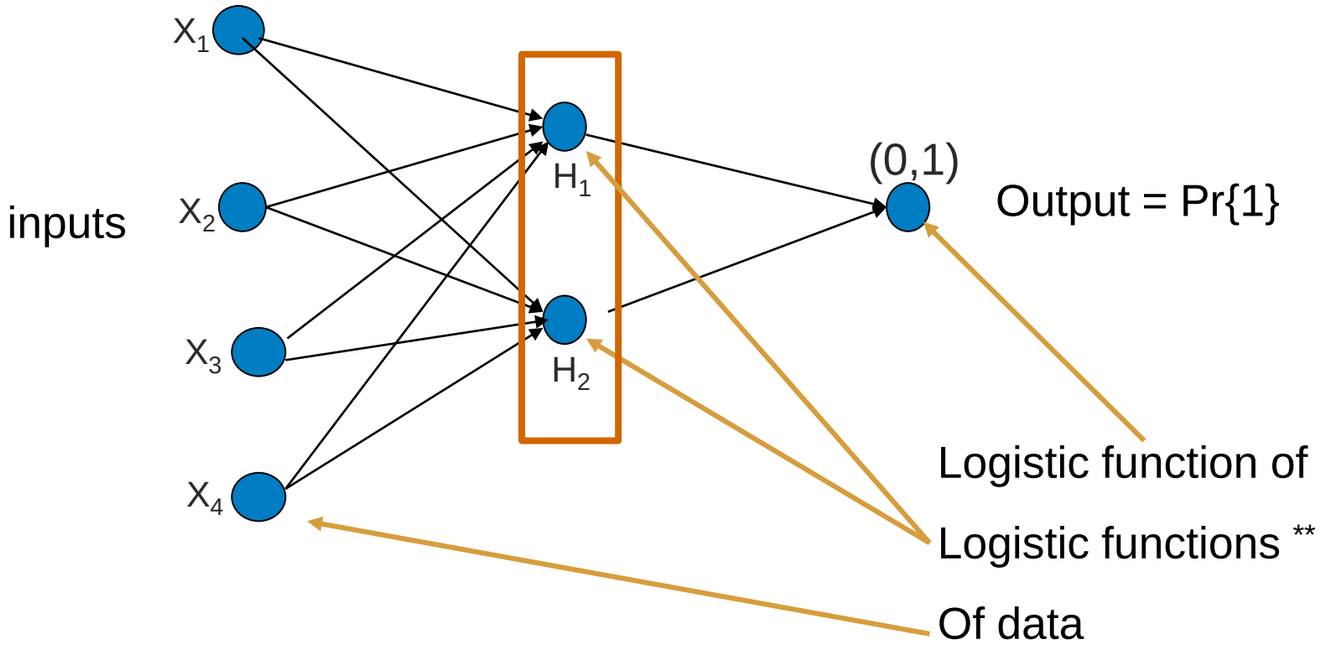
# Pr{2 OR MORE FAILURES}



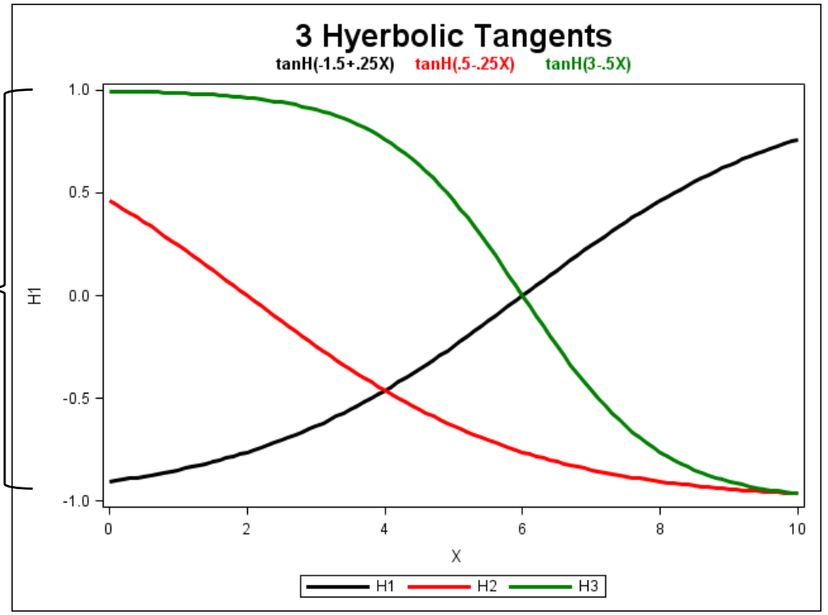

$$\Pr\{2 \ \ or \ \ more\} = 1 - p_X{}^0 (1 - p_X)^6 - 6 p_X (1 - p_X)^5$$

# Neural Networks

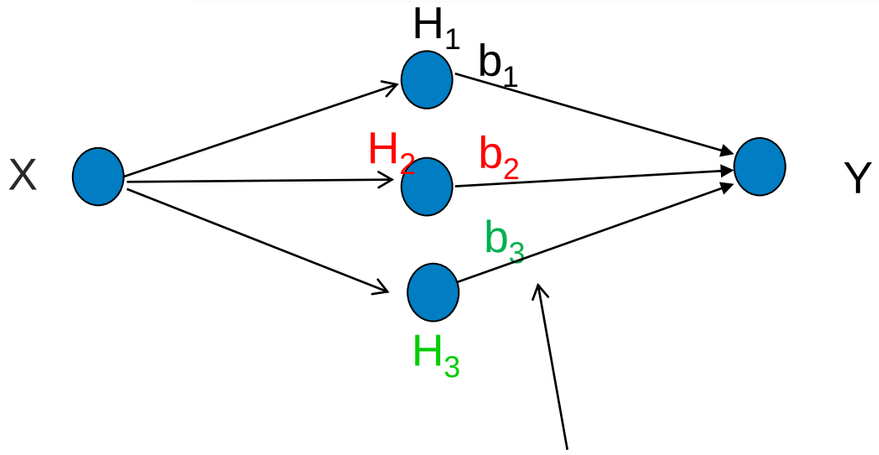
- Very flexible functions
- “Hidden Layers”
- “Multilayer Perceptron”




** (note: Hyperbolic tangent functions are just reparameterized logistic functions)




-1 to 1




Arrows on right represent linear combinations of “basis functions,” e.g. hyperbolic tangents (reparameterized logistic curves)

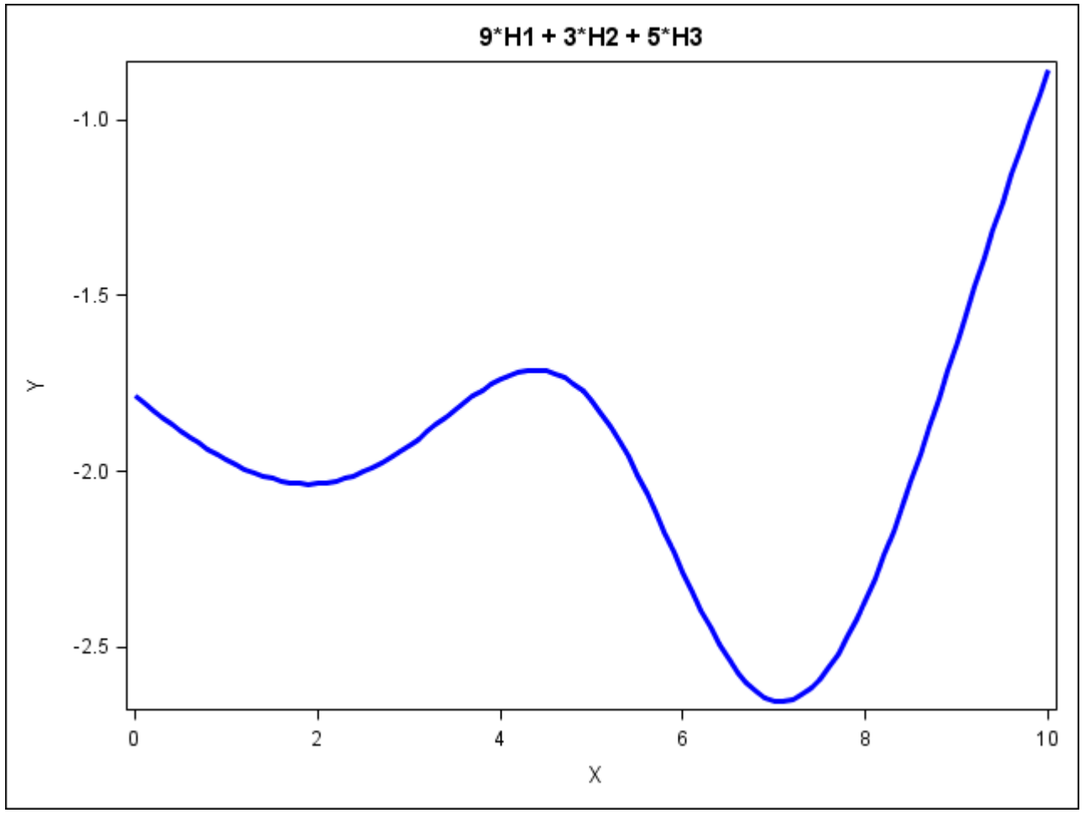
Example:

$Y = a + b_1\ H1 + b_2\ H2 + b_3\ H3$

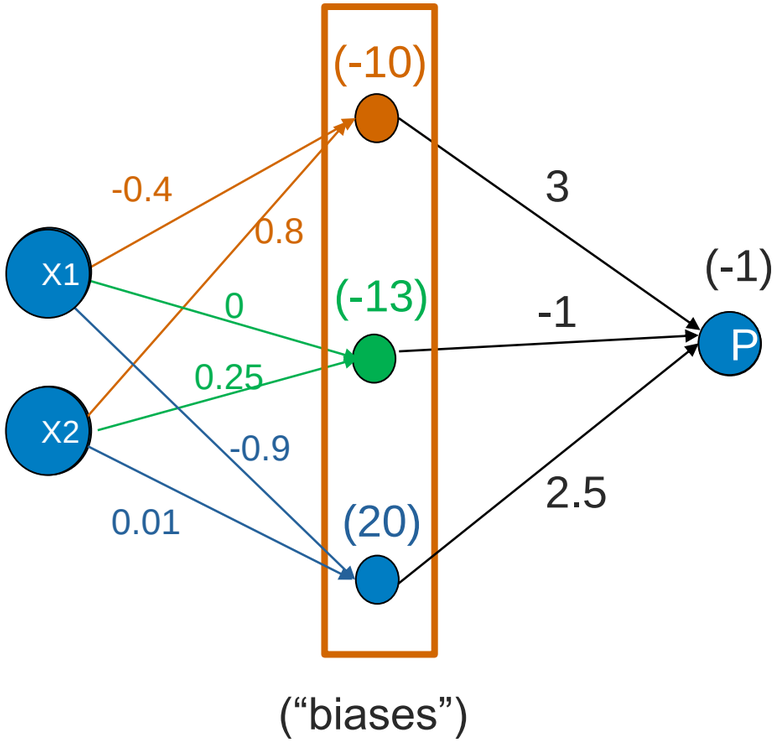
$Y = 0 + 9\ H1 + 3\ H2 + 5\ H3$

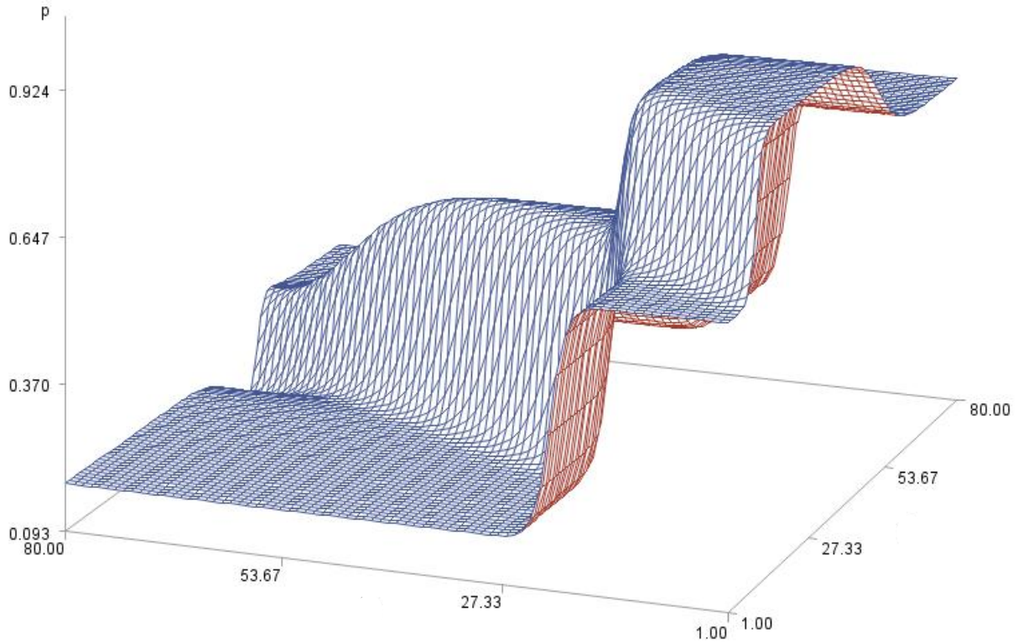
“bias”

“weights”



# A Complex Neural Network Surface



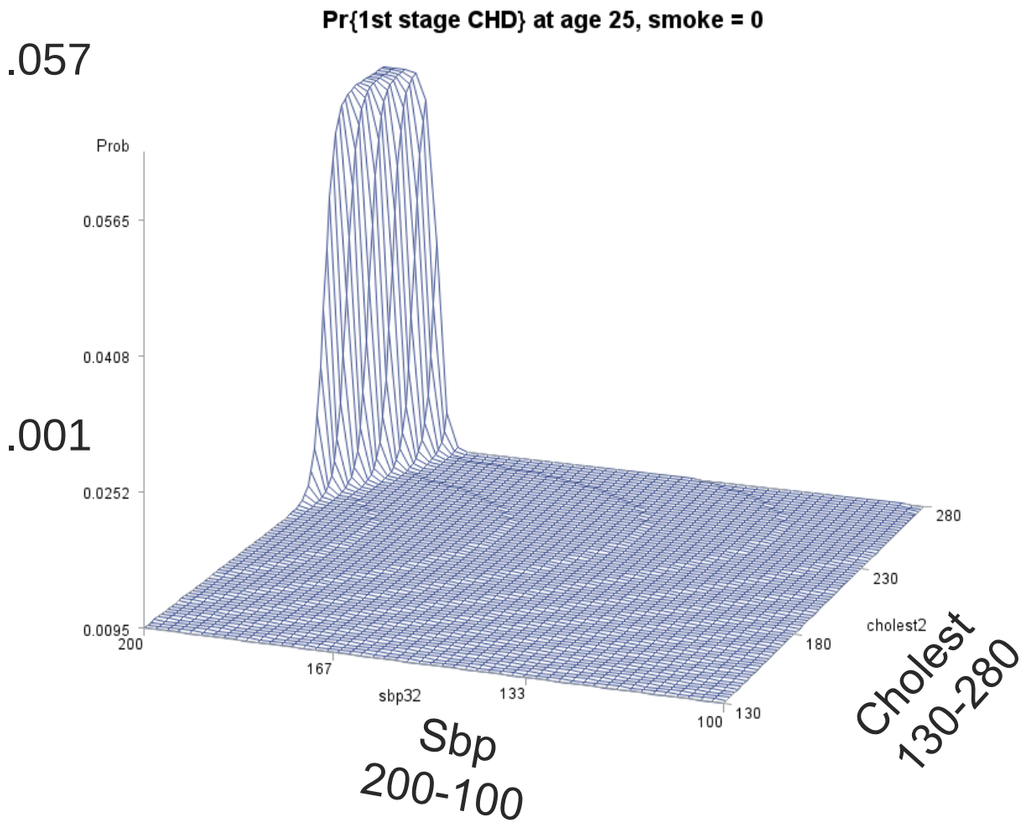
# Framingham Neural Network

Note: ⚠ No validation data used – surfaces are unnecessarily complex

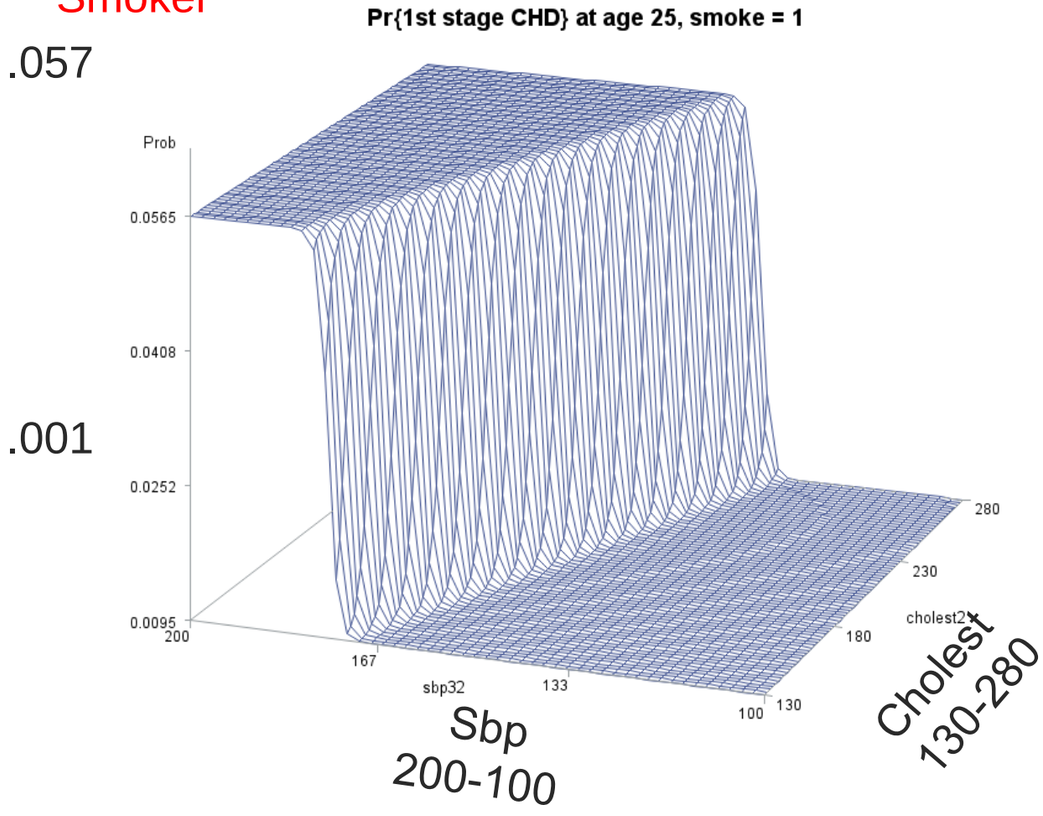
Inputs: Age, Systolic BP, Cholesterol, Smoker

"Slice" of 5-D surface at age=25

Age 25
Nonsmoker



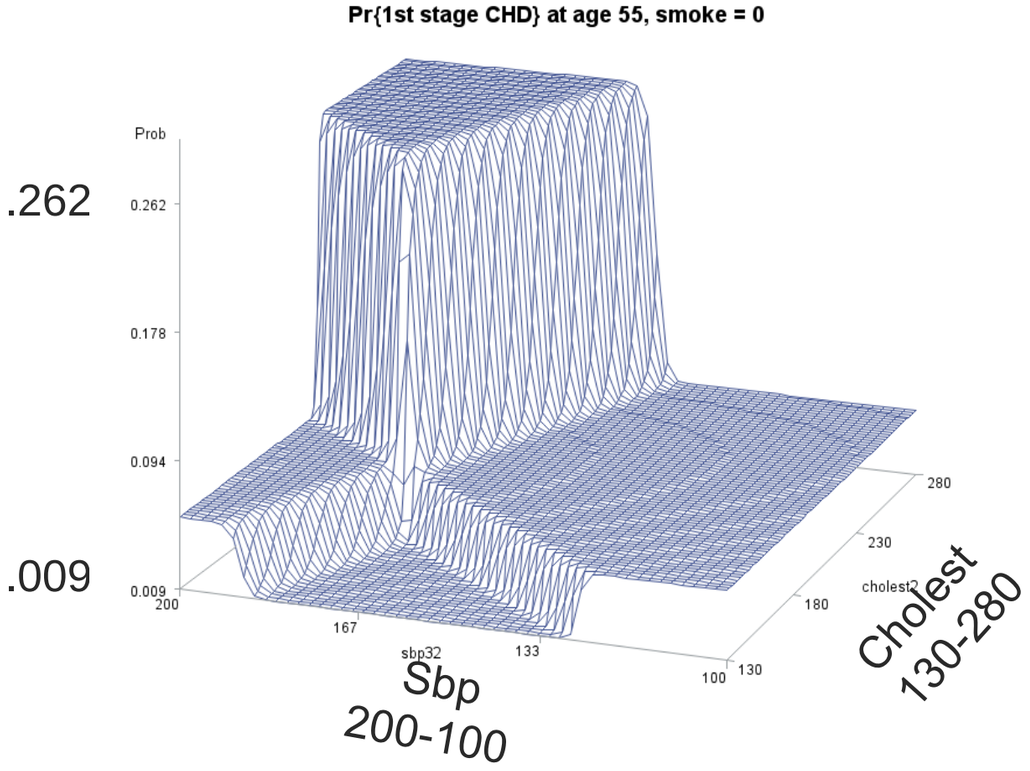

Age 25
Smoker



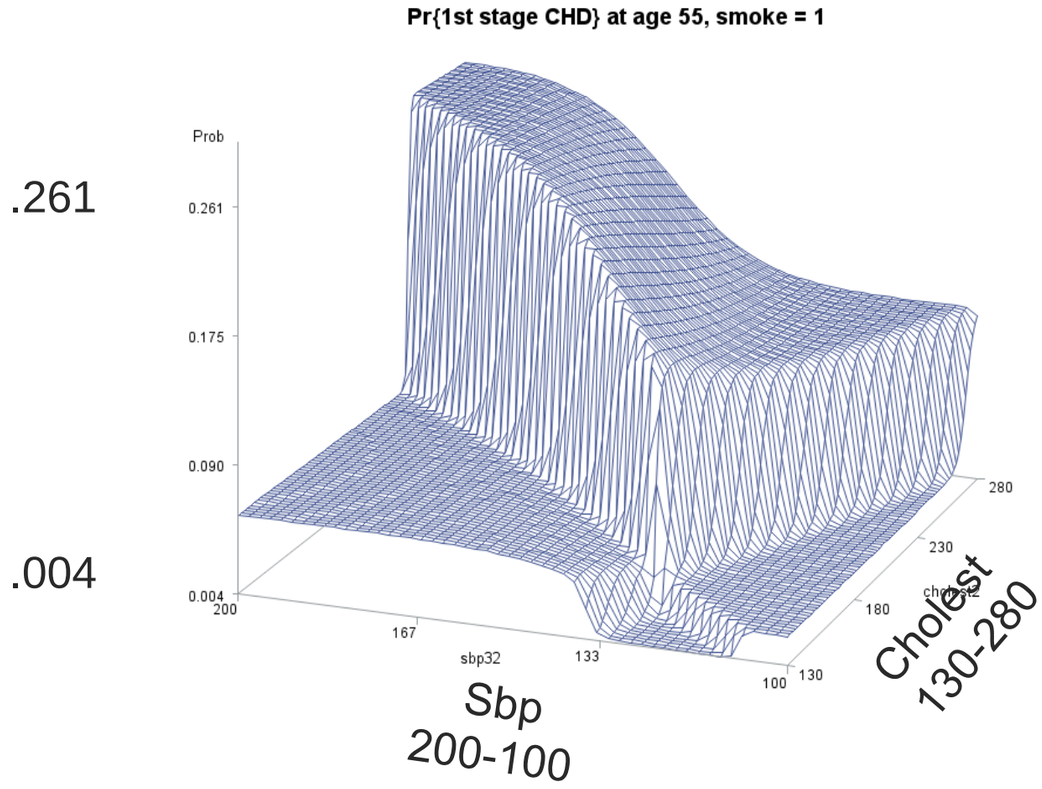
# Framingham Neural Network Surfaces

"Slice" at age 55 – are these patterns real?

Age 55
Nonsmoker




Age 55
Smoker



# Handling Neural Net Complexity

(1) Use validation data, stop iterating when fit gets worse on validation data.

(2) Use regression to omit predictor variables (“features”) and their parameters. This gives previous graphs.

(3) Do both (1) and (2).

# Extension of Neural Networks

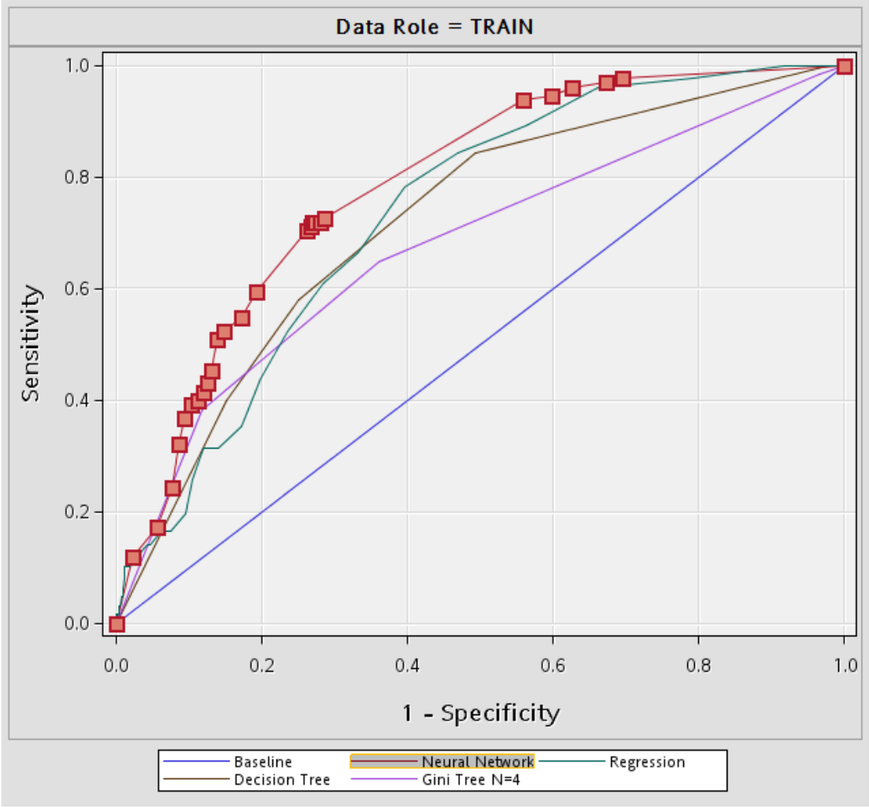
## “Deep learning”

A multilayer neural network

First layers get overall big picture

Latter layers fill in detail

There is feedback between layers




← Framingham ROC for 4 models and baseline 45 degree line
Neural net highlighted, area 0.780

Sensitivity: Pr{ calling a 1 a 1 }=Y coordinate
Specificity; Pr{ calling a 0 a 0 } = 1-X.

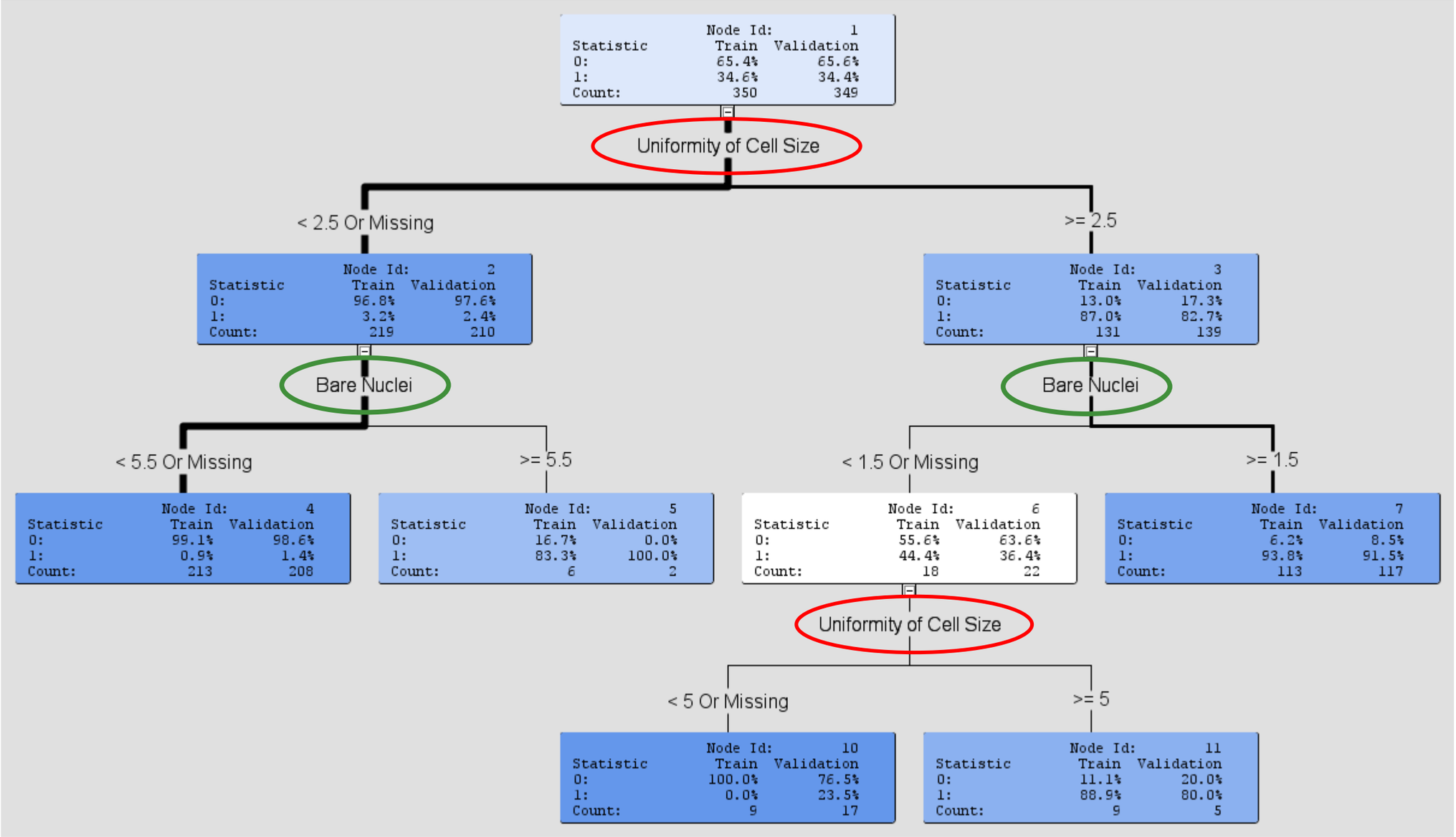
| Selected Model | Model Node | Model Description | Misclassification Rate | Avg. Squared Error | Area Under ROC |
|---|---|---|---|---|---|
| Y | Neural | Neural Network | 0.079257 | 0.066890 | 0.780 |
| | Tree | Decision Tree | 0.079257 | 0.069369 | 0.720 |
| | Tree2 | Gini Tree N=4 | 0.079257 | 0.069604 | 0.675 |
| | Reg | Regression | 0.080495 | 0.069779 | 0.734 |

# Three Breast Cancer Models

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

1. Sample code number id number
2. Clump Thickness 1 - 10
3. Uniformity of Cell Size 1 - 10
4. Uniformity of Cell Shape 1 - 10
5. Marginal Adhesion 1 - 10
6. Single Epithelial Cell Size 1 - 10
7. Bare Nuclei 1 - 10
8. Bland Chromatin 1 - 10
9. Normal Nucleoli 1 - 10
10. Mitoses 1 - 10
11. Class: (2 for benign, 4 for malignant)

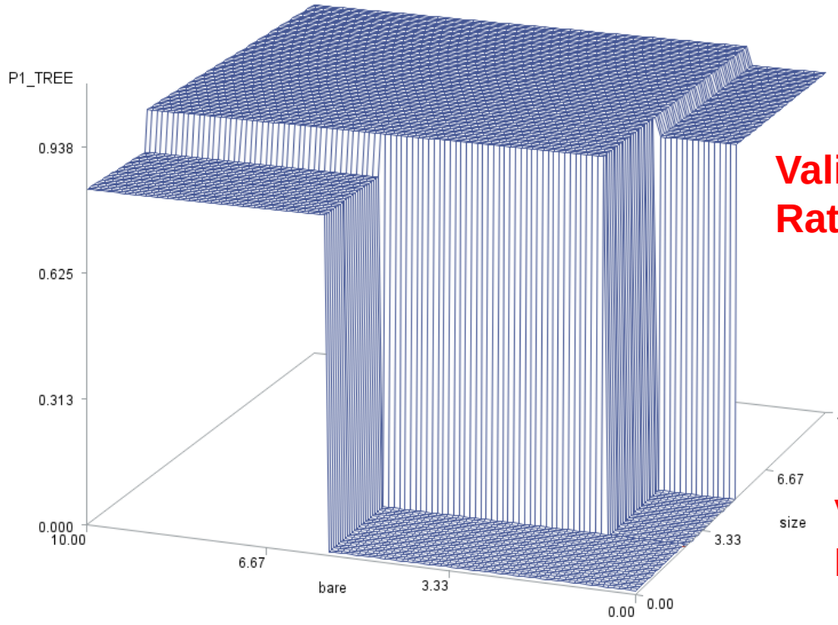
# Decision Tree Needs Only BARE & SIZE



# Cancer Screening Results: Model Comparison Node

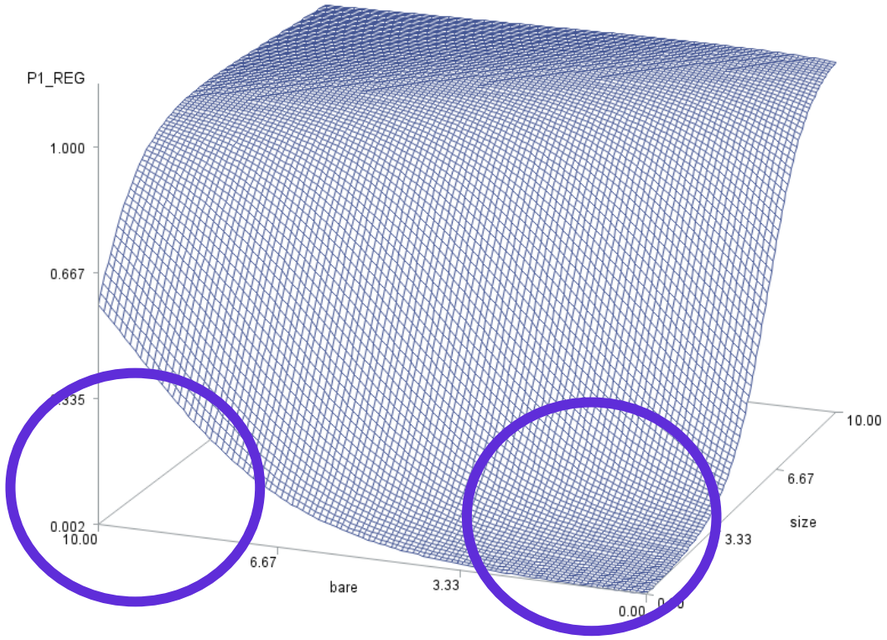
| Selected Model | Model Node | Model Description | Train Misclassification Rate | Train Avg. Squared Error | Train Area Under ROC | **Validation Misclassification Rate** | Validation Avg. Squared Error | Validation Area Under ROC |
|---|---|---|---|---|---|---|---|---|
| **Y** | Tree | **Decision Tree** | 0.031429 | 0.029342 | 0.975 | **0.051576** | 0.048903 | 0.933 |
|  | Neural | Neural Network | 0.031429 | 0.025755 | 0.994 | 0.054441 | 0.037838 | 0.985 |
|  | Reg | Regression | 0.031429 | 0.029322 | 0.992 | 0.054441 | 0.040768 | 0.987 |

**Decision Tree**

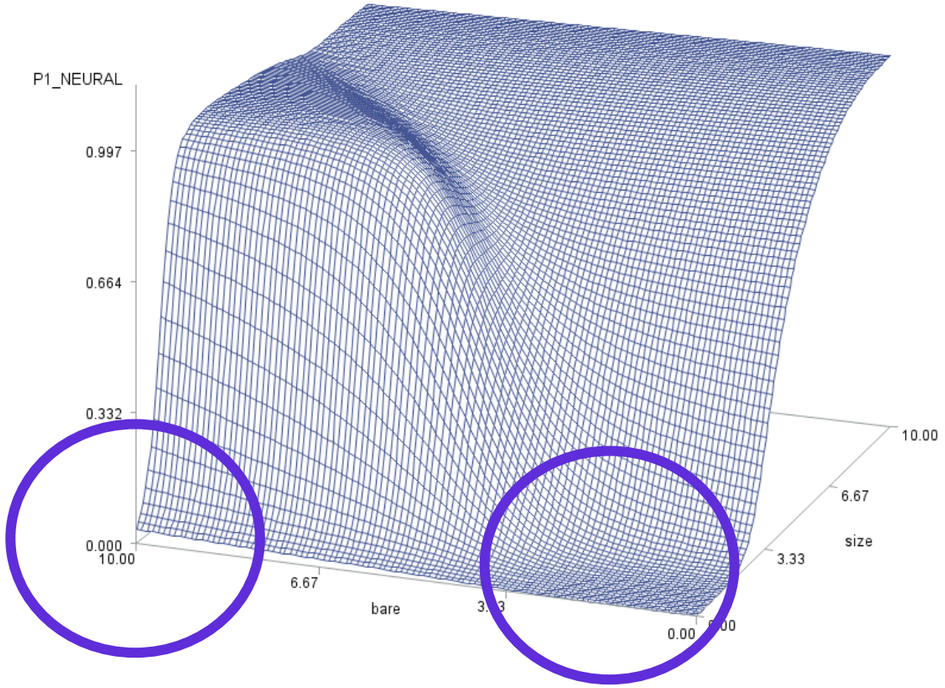



**Validation Misclassification Rate 0.0516**

**Validation Misclassification Rate 0.0544** **Logistic Regression**




**Validation Misclassification Rate 0.0544** **Neural Network**



# Unsupervised Learning

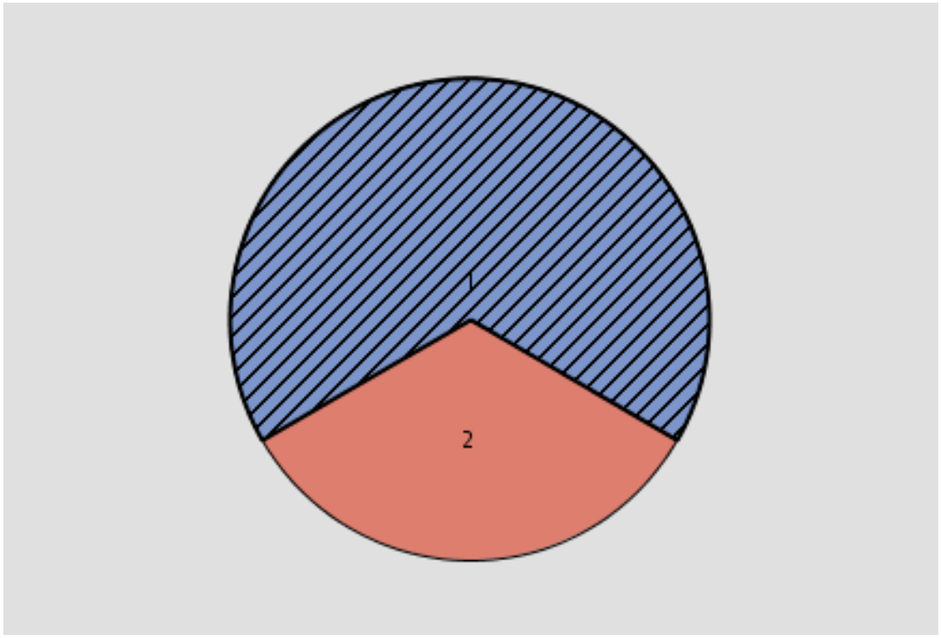
- We have the “features” (predictors)
- We do NOT have the response even on a training data set (*UN*supervised)
- Another name for clustering
- SAS Enterprise Miner
  - Large number of clusters with k-means (k clusters)
  - Ward’s method to combine (less clusters , say r<k)
  - One more k means  for final r-cluster solution.

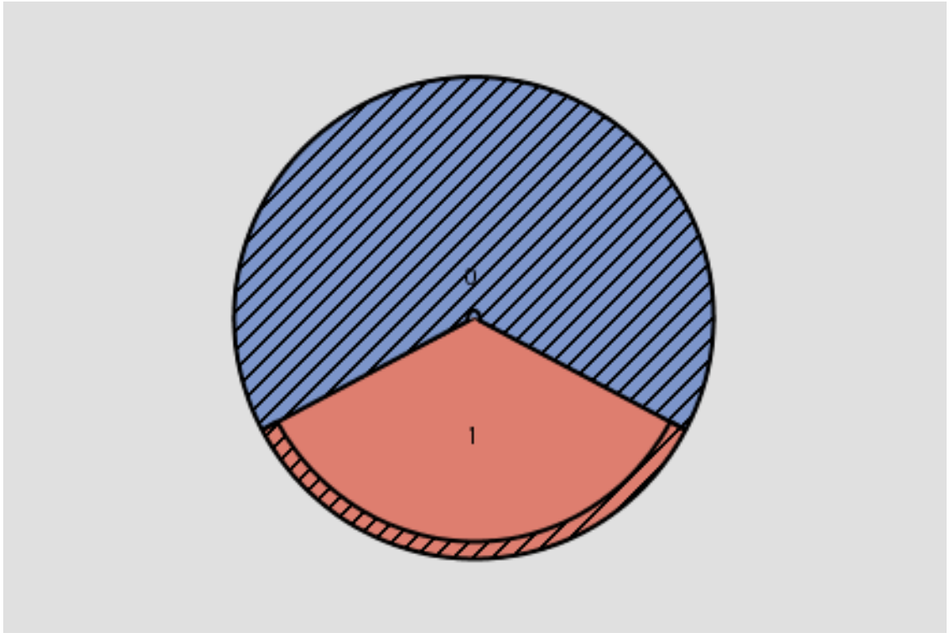
# Unsupervised Learning

- K means – place spread out seeds
  - Cluster to closest seed (centroid)
  - Compute new centroid
  - Repeat
- Ward – $\Sigma$ (sum of squared distances from centroids).
  - Combine 2 that minimally increase $\Sigma$ (sum of squares).
  - Repeat

Example:  Cluster the breast cancer data (disregard target)


Plot clusters and actual target values:

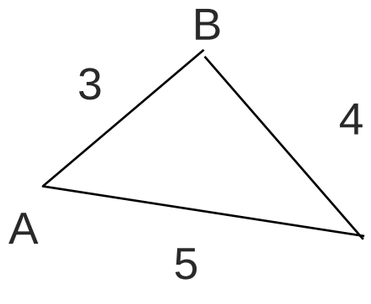


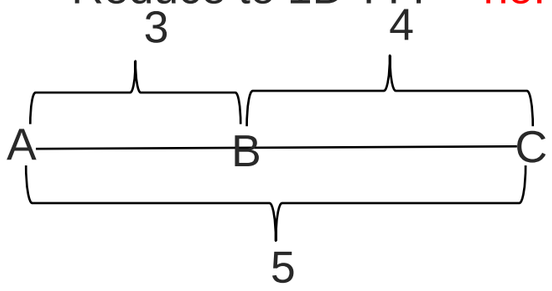
# Multidimensional Scaling

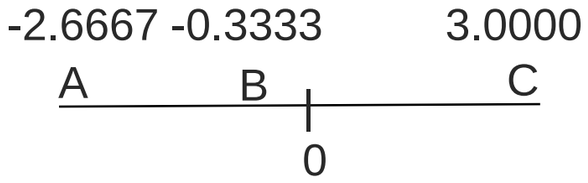
MORPH !

Figure in 2D




Reduce to 1D ??? - no!




Can we get <u>close</u> in some sense? (e.g. least squares)

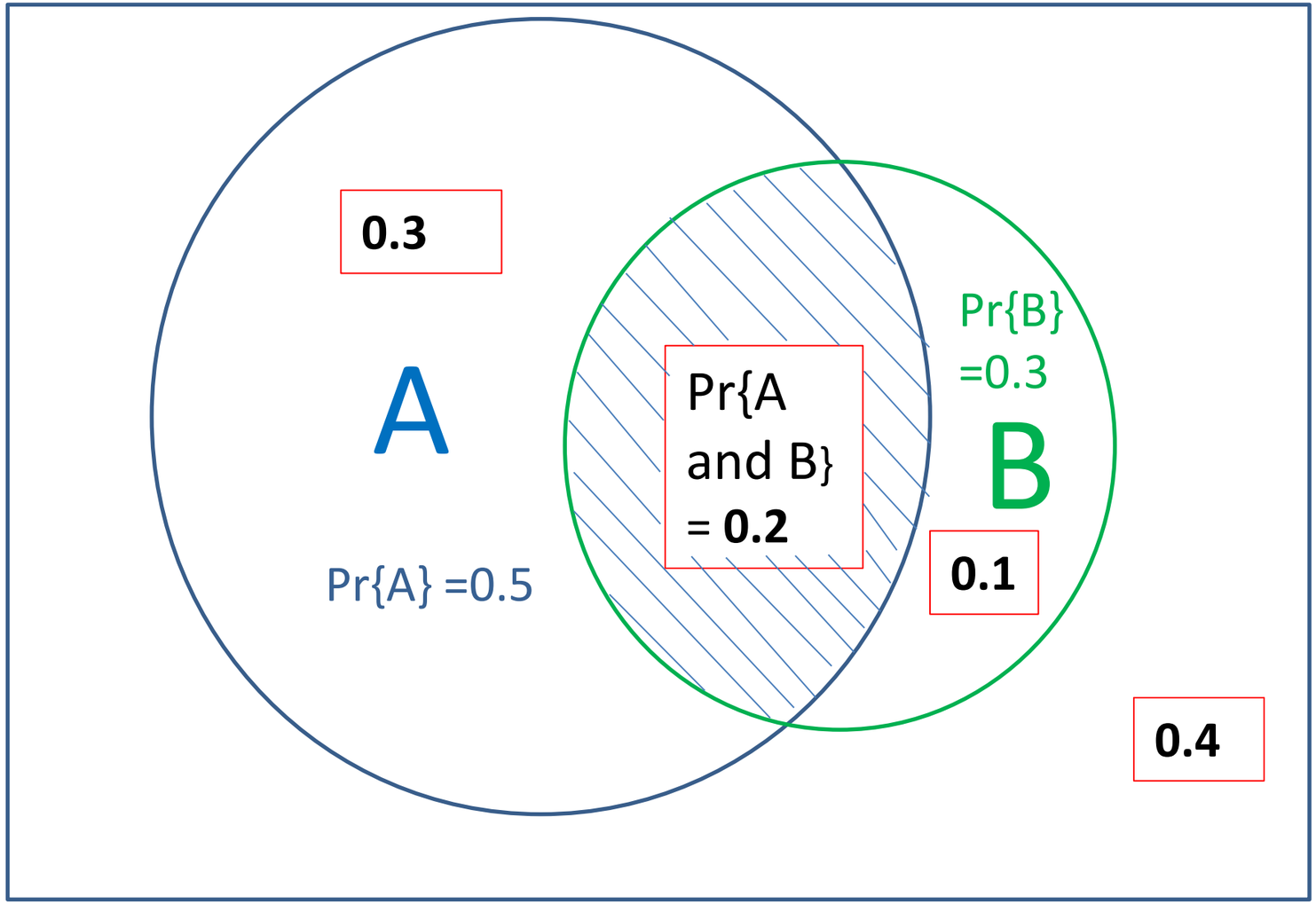
X matrix

| A | B | C | Y |
|---|---|---|---|
| -1 | 0 | 1 | 5 |
| -1 | 1 | 0 | 3 |
| 0 | -1 | 1 | 4 |
| 1 | 1 | 1 | 0 |

$$b = (X'X)^{-1}X'Y = \begin{pmatrix} -2.6667 \\ -0.3333 \\ 3.0000 \end{pmatrix}$$

| Obs | Dependent Variable | Predicted Value | Residual |
|---|---|---|---|
| 1 | 5 | 5.6667 | -0.6667 |
| 2 | 3 | 2.3333 | 0.6667 |
| 3 | 4 | 3.3333 | 0.6667 |
| 4 | 0 | 0 | 0 |

# Association Analysis is just elementary probability with new names



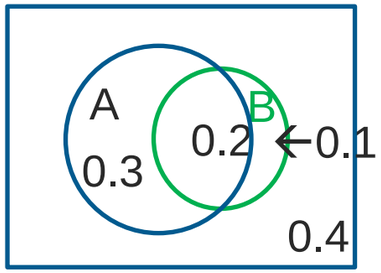

A: Purchase Milk

B: Purchase Cereal

0.3+0.2+0.1+0.4 = 1.0

Cereal=> Milk

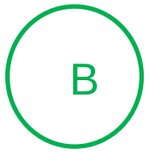
Rule B=> A “people who buy B will buy A”




Support:

Support= Pr{A and B} = **0.2**

Independence means that Pr{A|B} = Pr{A} = **0.5**
Pr{A} = **0.5** = Expected confidence if there is no relation to B.




Confidence:

Confidence = Pr{A|B}=Pr{A and B}/Pr{B}=**2/3**

??- Is the confidence in B=>A the same as the confidence in A=>B?? (yes, no)

Marketing A to the 30% of people who buy **B** will result in 33% better sales than marketing to a random 30% of the people.

Lift:

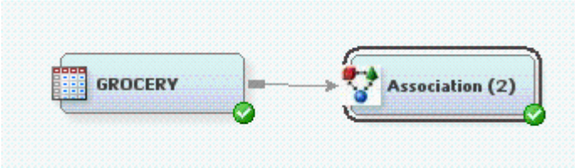
Lift = confidence / E{confidence} = (2/3) / (1/2) = **1.33**

Gain = 33%

# Example: Grocery cart items (hypothetical)

item cart

bread 1
milk 1
soap 2
meat 3
bread 3
bread 4
cereal 4
milk 4
soup 5
bread 5
cereal 5
milk 5



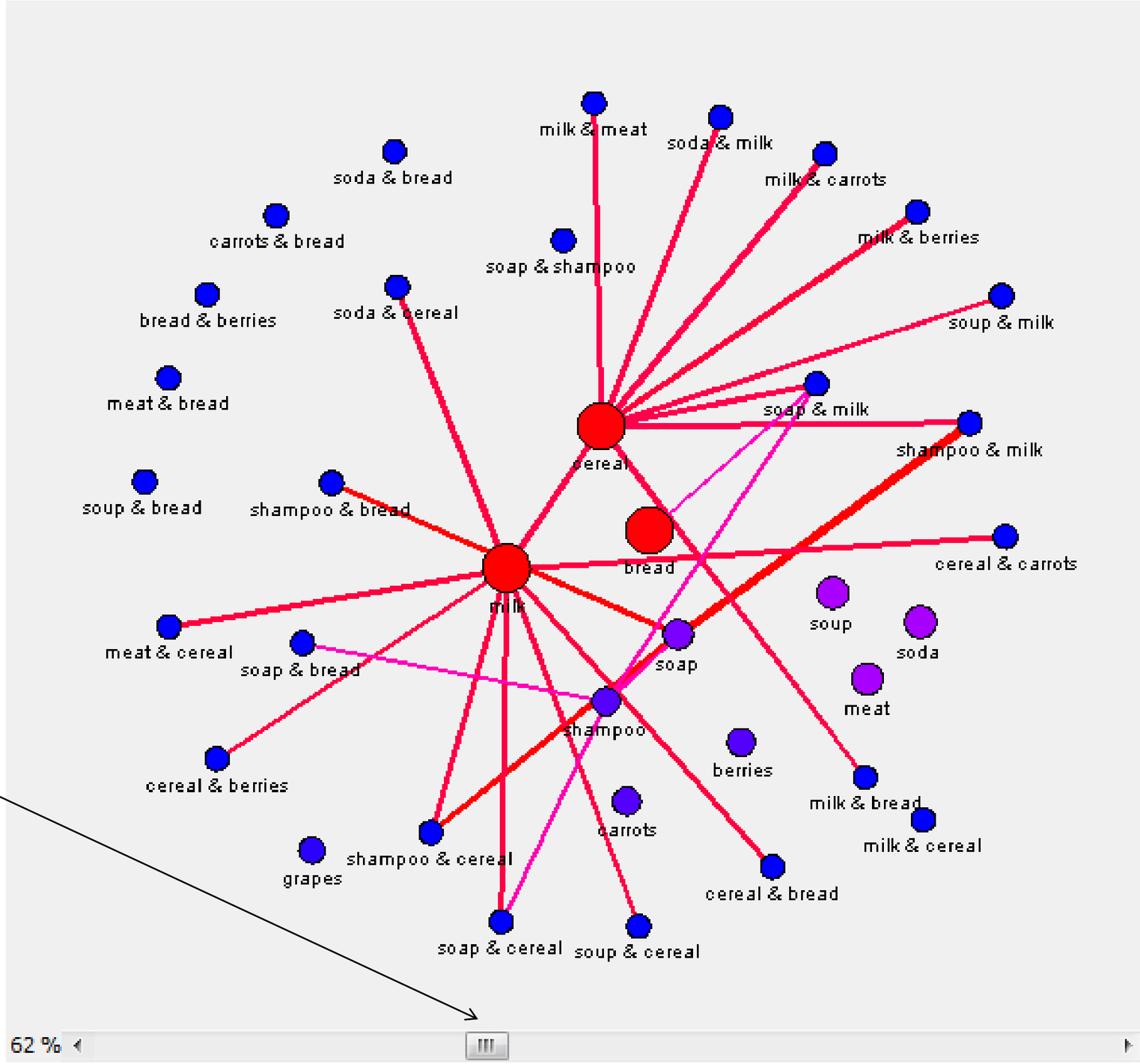

Sort Criterion: Confidence
Maximum Items: 2
Minimum (single item) Support: 20%

Association Report

| Relations | Expected Confidence (%) | Confidence (%) | Support (%) | Lift | Transaction Count | Rule |
|---|---|---|---|---|---|---|
| 2 | 63.43 | 84.21 | 52.68 | 1.33 | 8605 | cereal ==> milk |
| 2 | 62.56 | 83.06 | 52.68 | 1.33 | 8605 | milk ==> cereal |
| 2 | 63.43 | 61.73 | 12.86 | 0.97 | 2100 | meat ==> milk |
| 2 | 63.43 | 61.28 | 38.32 | 0.97 | 6260 | bread ==> milk |
| 2 | 63.43 | 60.77 | 12.81 | 0.96 | 2093 | soup ==> milk |
| 2 | 62.54 | 60.42 | 38.32 | 0.97 | 6260 | milk ==> bread |
| 2 | 62.56 | 60.28 | 37.70 | 0.96 | 6158 | bread ==> cereal |
| 2 | 62.54 | 60.26 | 37.70 | 0.96 | 6158 | cereal ==> bread |
| 2 | 62.56 | 60.16 | 12.69 | 0.96 | 2072 | soup ==> cereal |
| 2 | 21.08 | 20.28 | 12.69 | 0.96 | 2072 | cereal ==> soup |
| 2 | 20.83 | 20.27 | 12.86 | 0.97 | 2100 | milk ==> meat |
| 2 | 21.08 | 20.20 | 12.81 | 0.96 | 2093 | milk ==> soup |

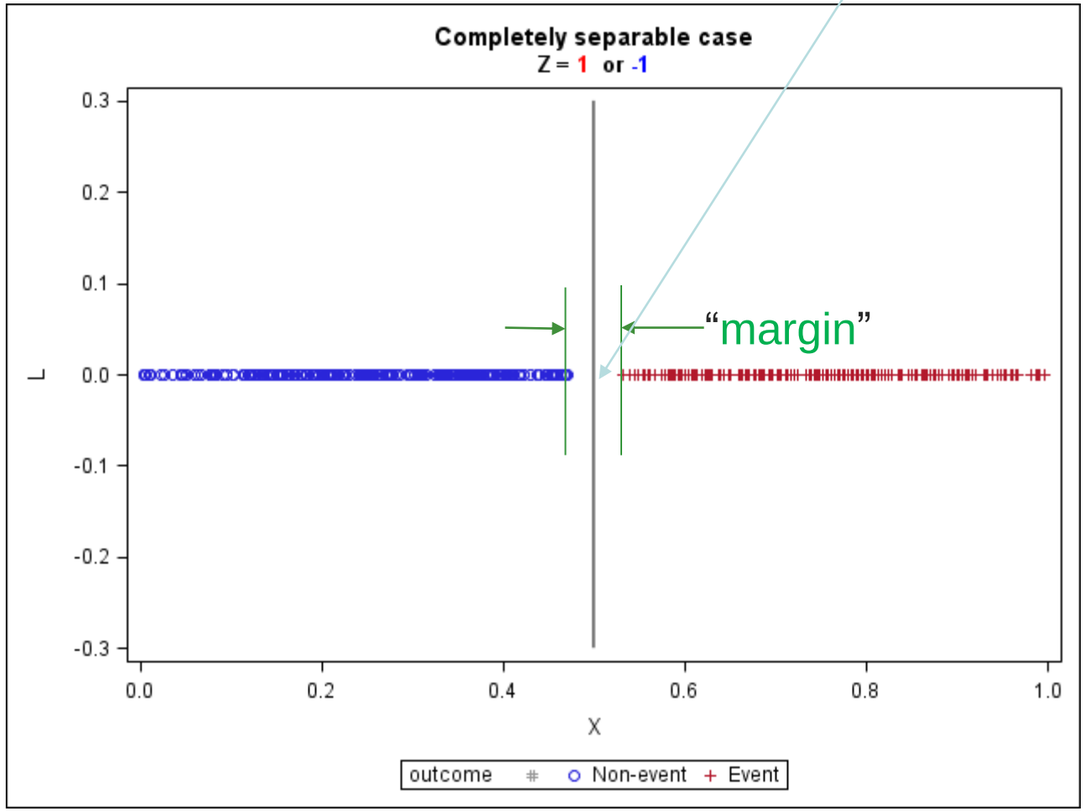
Link Graph →
Sort Criterion: Confidence
Maximum Items: 3
e.g. milk & cereal➔bread
Minimum Support: 5%

Slider bar at 62%



# Support Vector Machines

Find a point $X_0$ that "optimally" separates red from blue.
Optimally separate events from non-events.
Maximize the "margin" & take midpoint

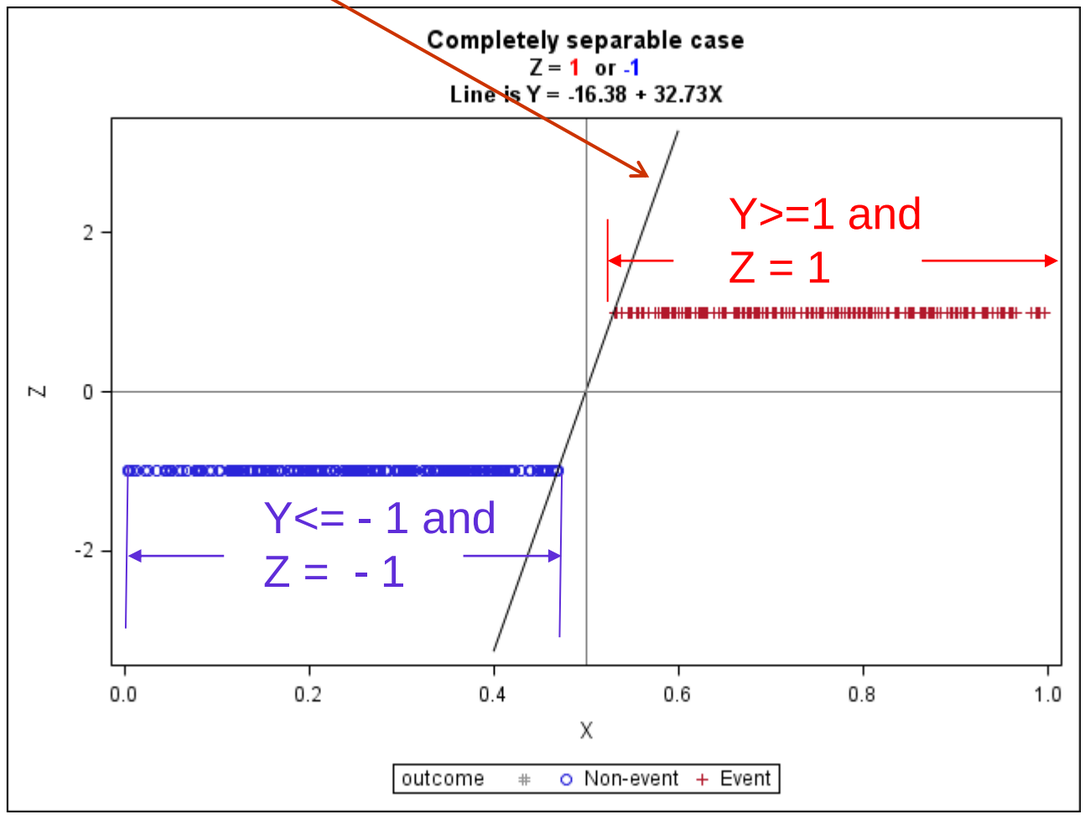


# Support Vector Machines

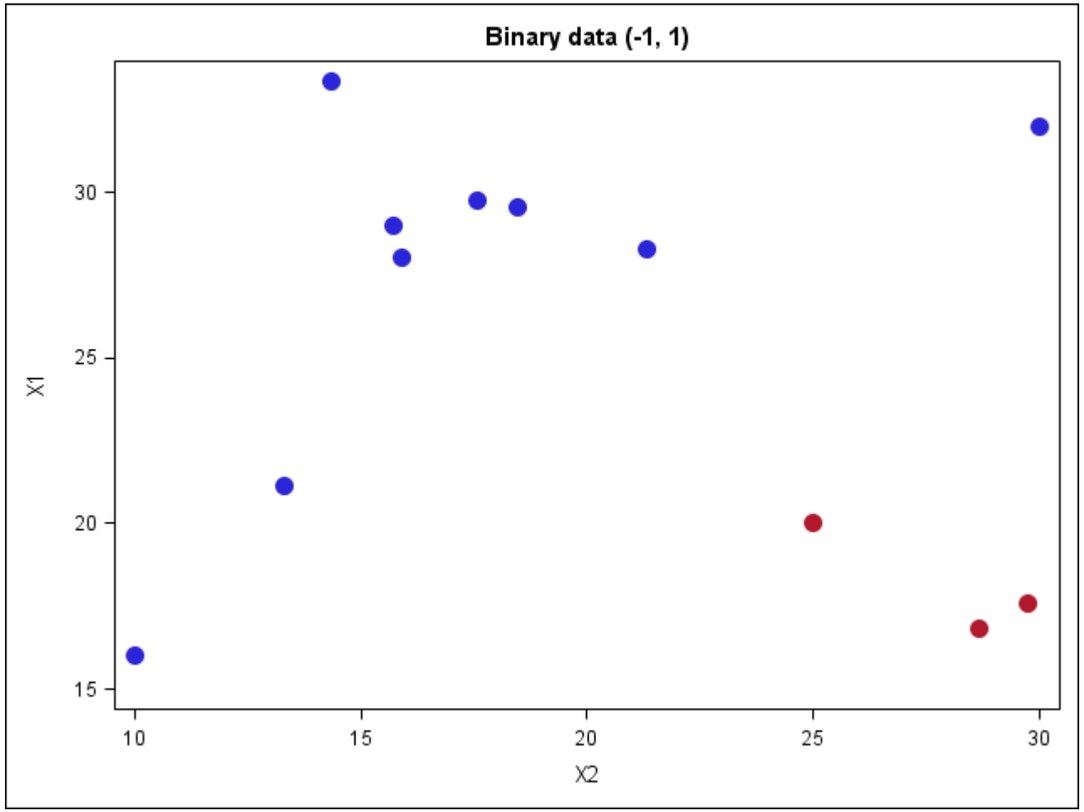
Let Z=1 for events, Z = -1 for non-events

Minimize slope of line subject to YZ>=1 everywhere

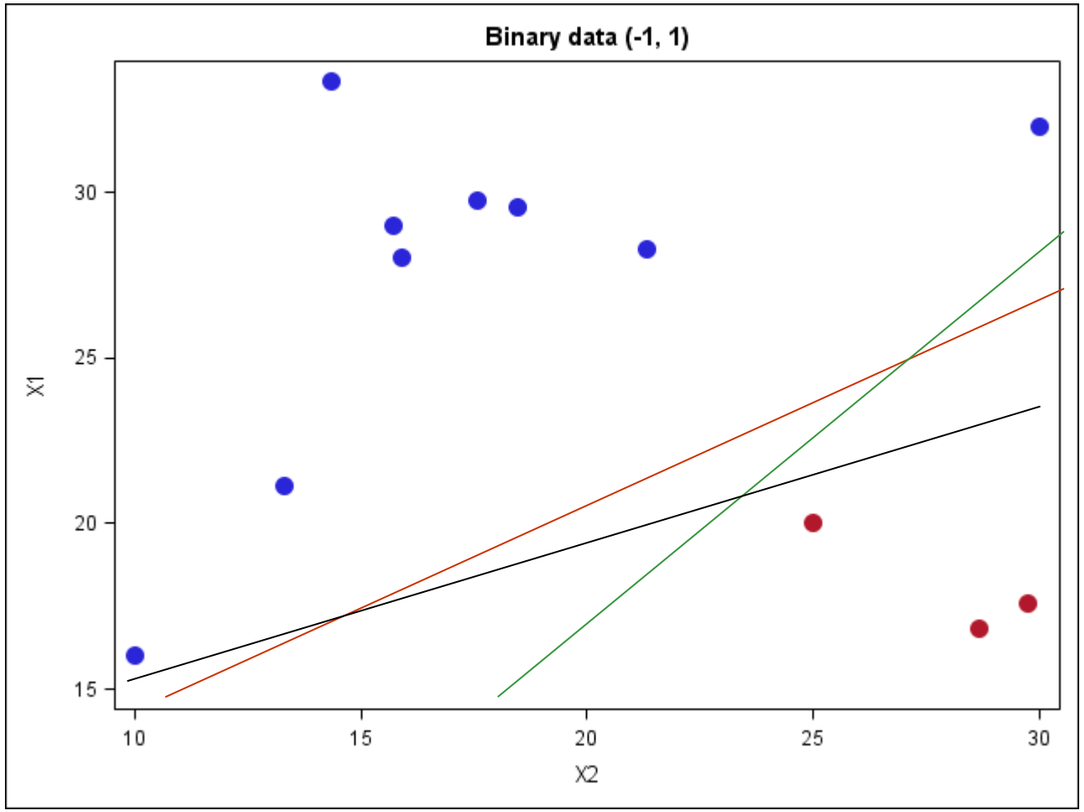
Y= -16.38 + 32.73 X so $X_0$=16.38/32.73=0.50



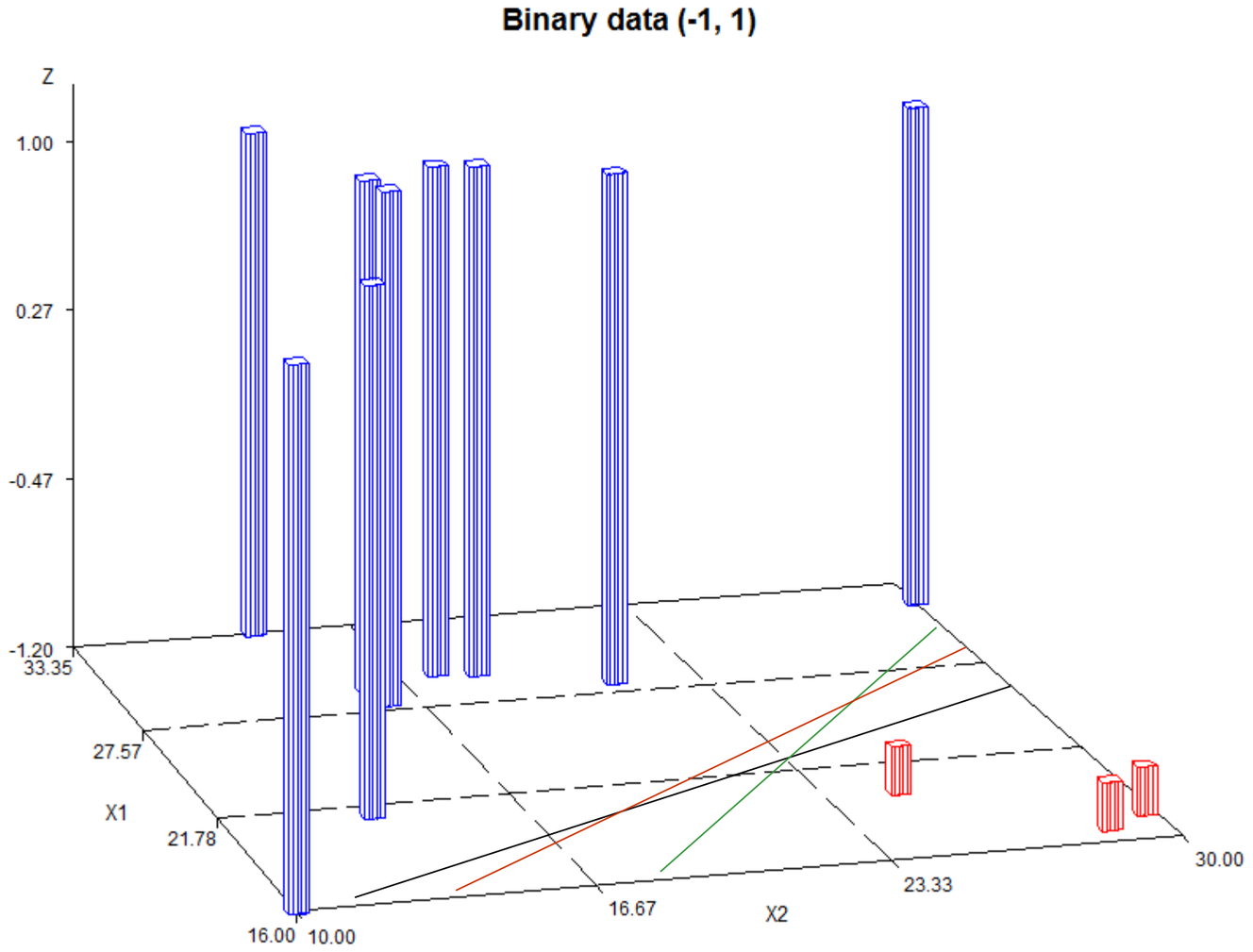
What about higher dimensions?  Separator is a line (not point).
Data completely separated by many lines
Which line maximizes margin?



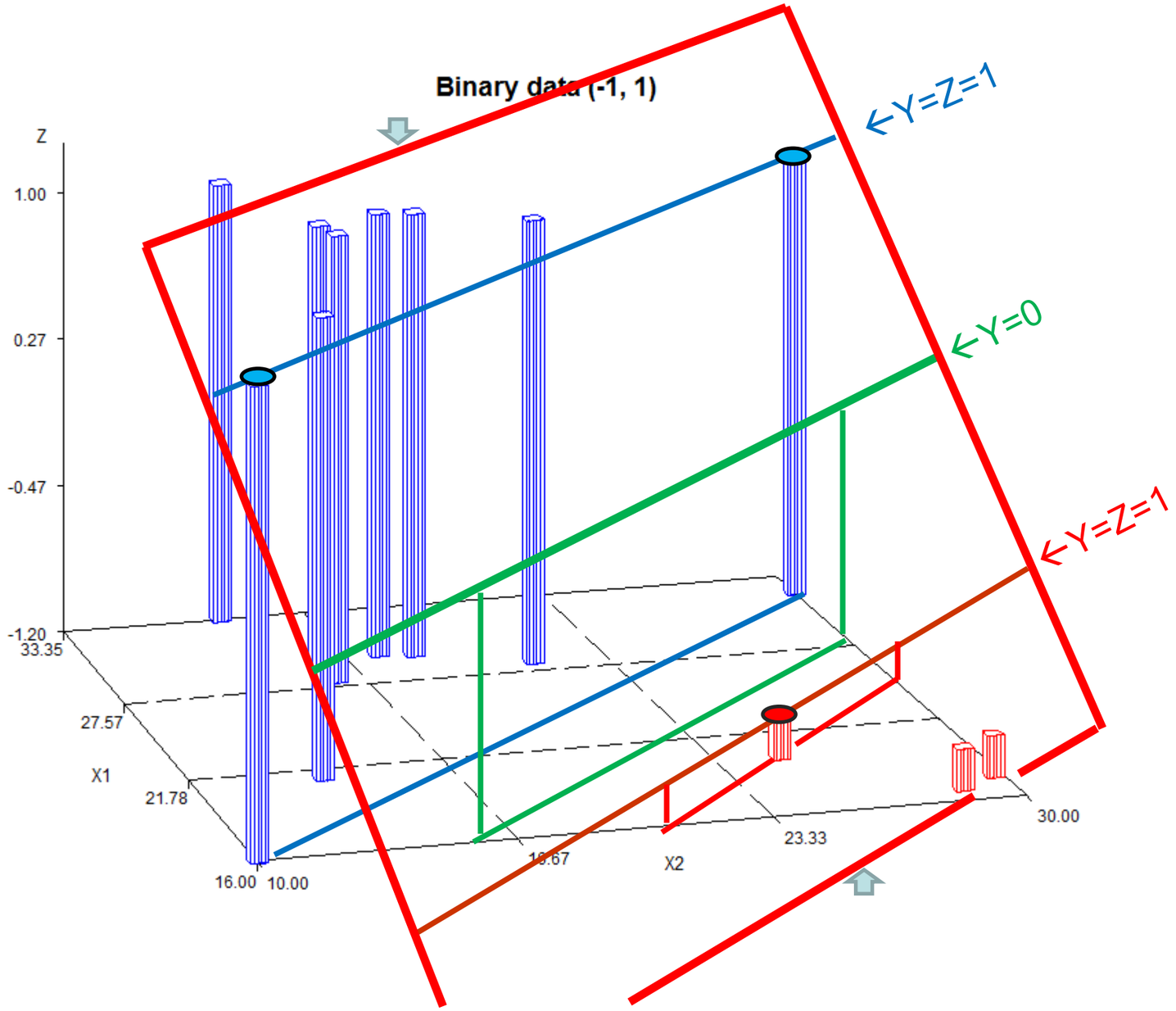
What about higher dimensions?  Separator is a line (not point).
Data completely separated by many lines
Which line maximizes margin?



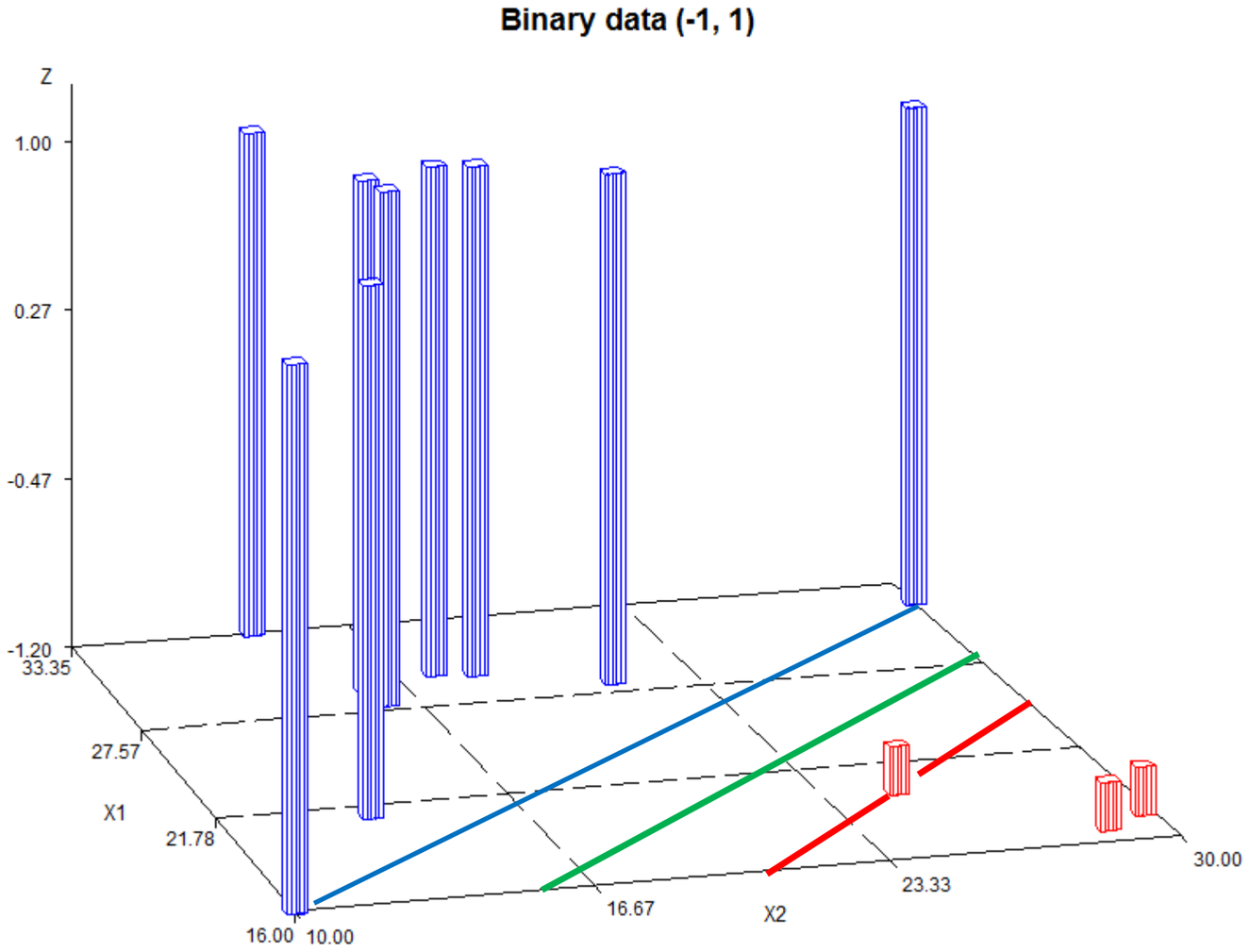
Add a Z dimension (Z=-1 or 1).
Which line maximizes margin?

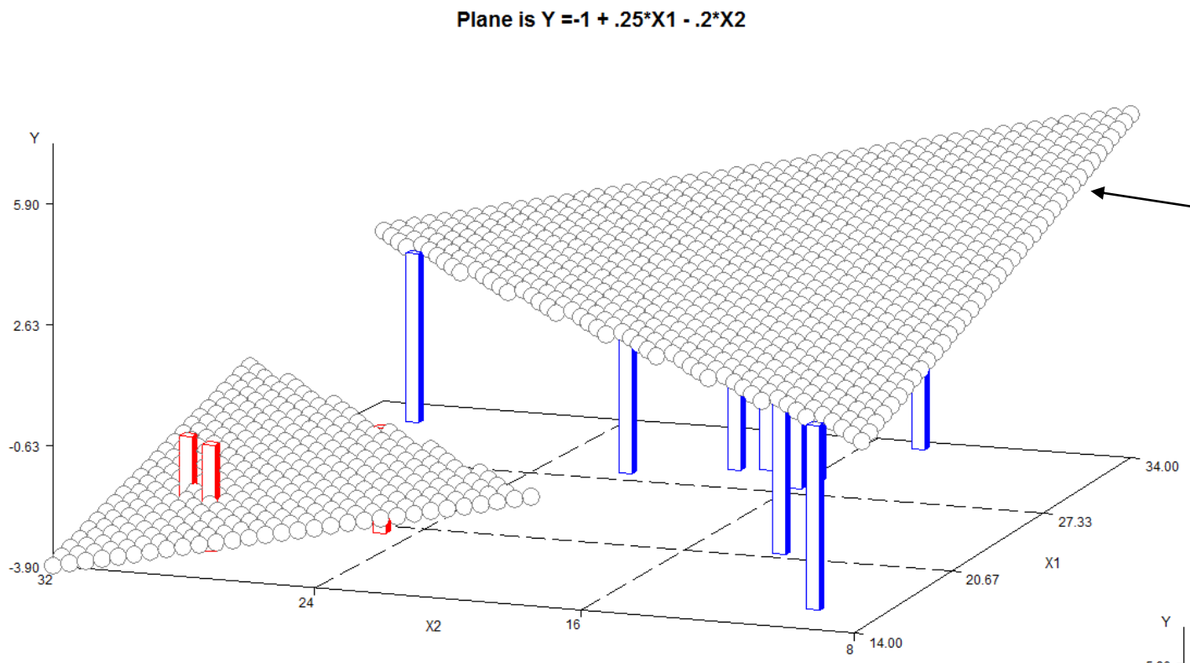


Add a Z dimension (Z=-1 or 1).
Plane: Y = -1 + 0.25*X1-0.20*X2



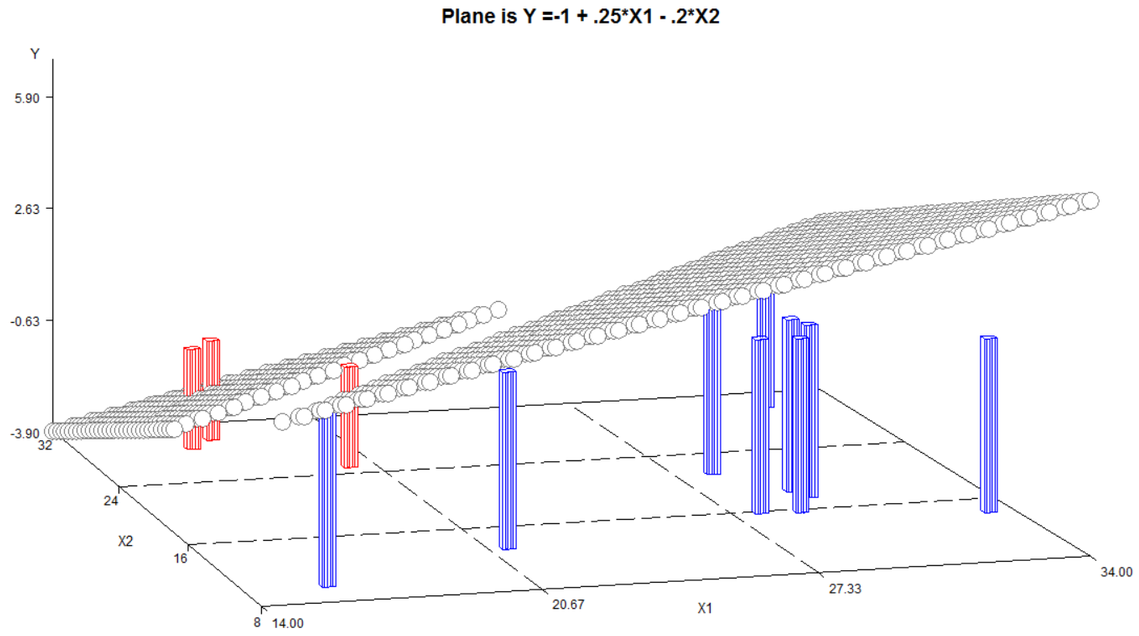
| = Maximum Margin Classifier (Z=-1 or 1).

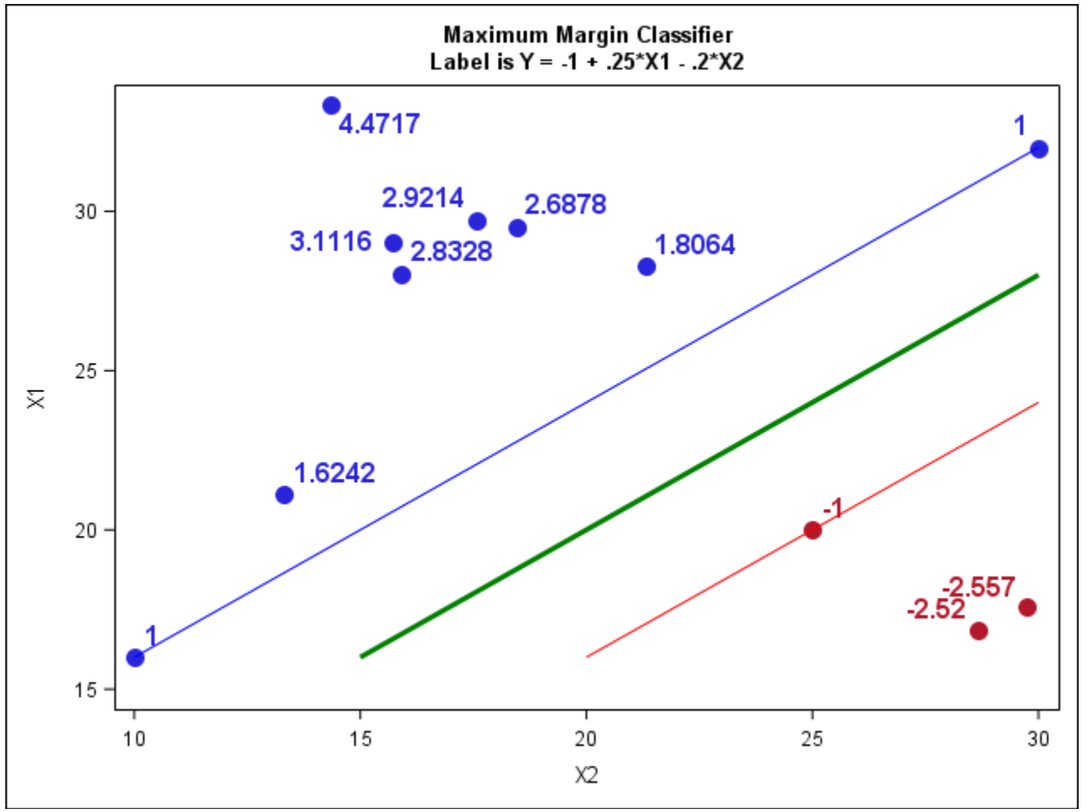




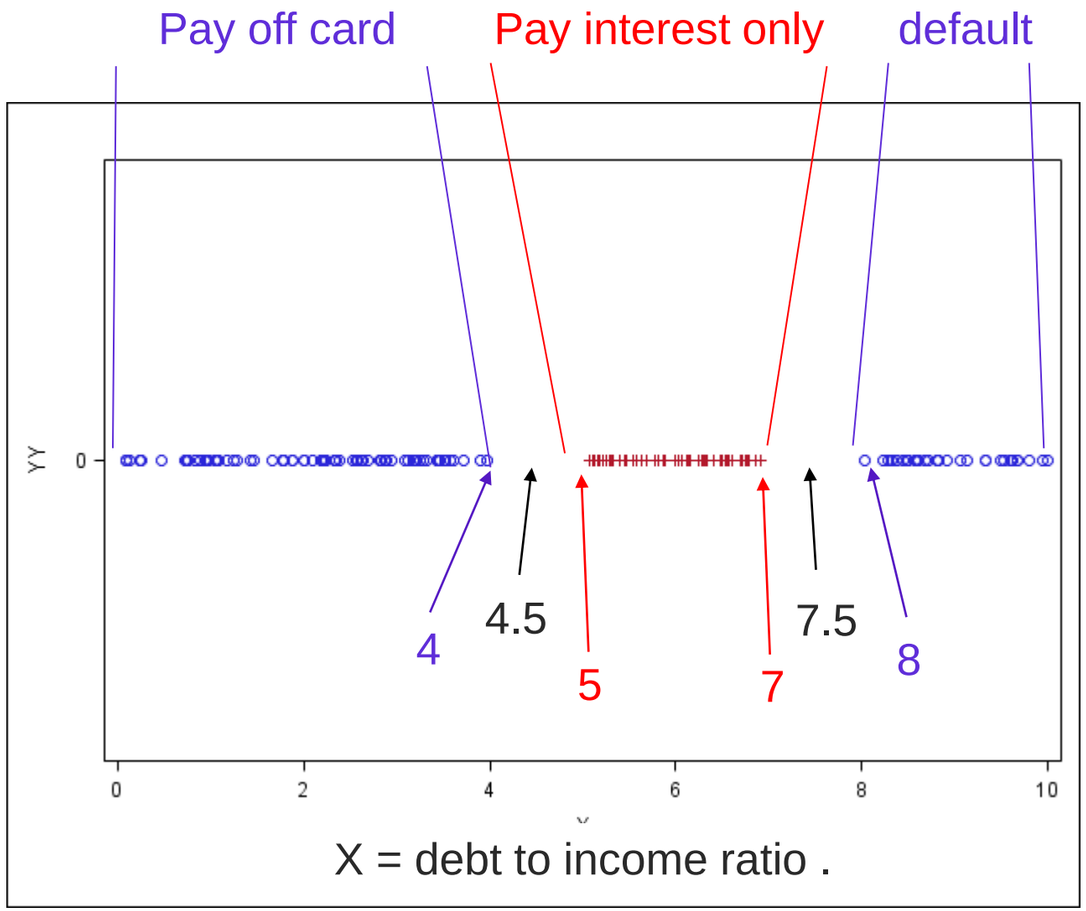

Minimize slope subject to all YZ>1

Y = -1 + 0.25*X1 – 0.20* X2




Maximum Margin Classifier
Label is Y = -1 + .25*X1 - .2*X2
4.4717
1
2.9214
2.6878
3.1116
2.8328
1.8064
1.6242
-1
-2.557
-2.52
1
30
25
20
15
X1
10
15
20
25
30
X2

Credit card payments versus
debt to income ratio .



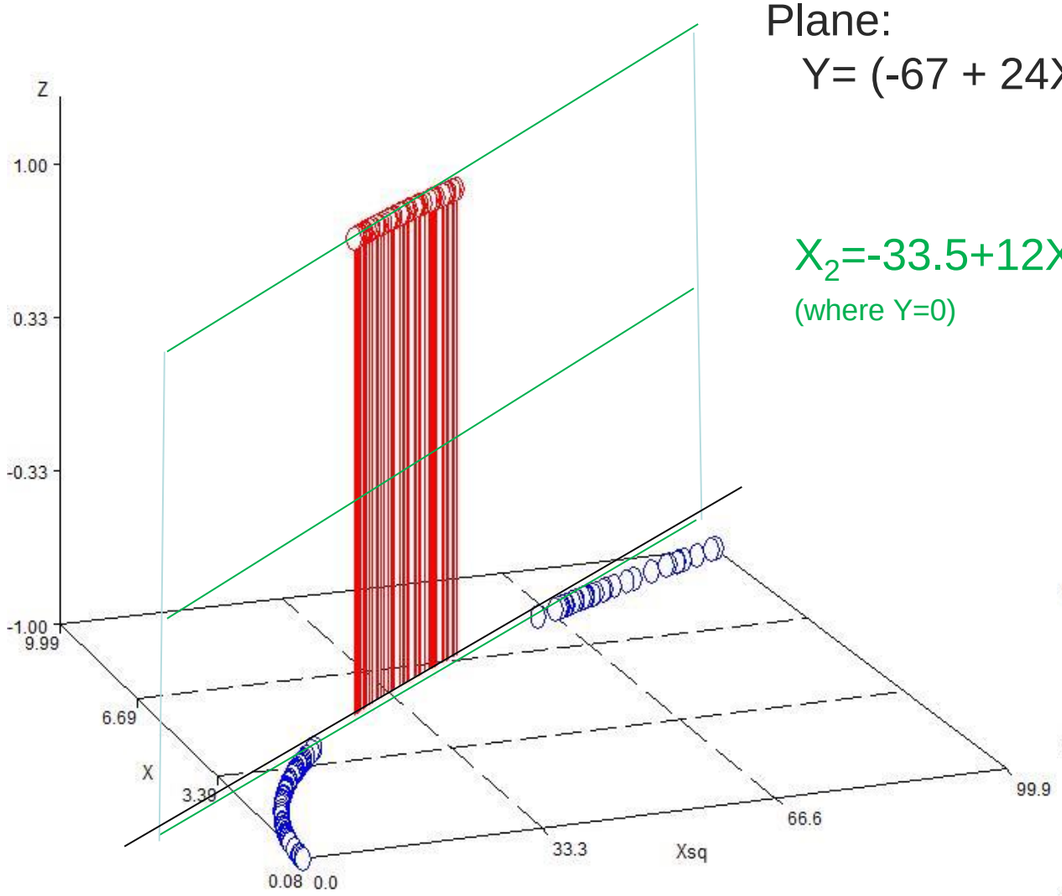
Plane:

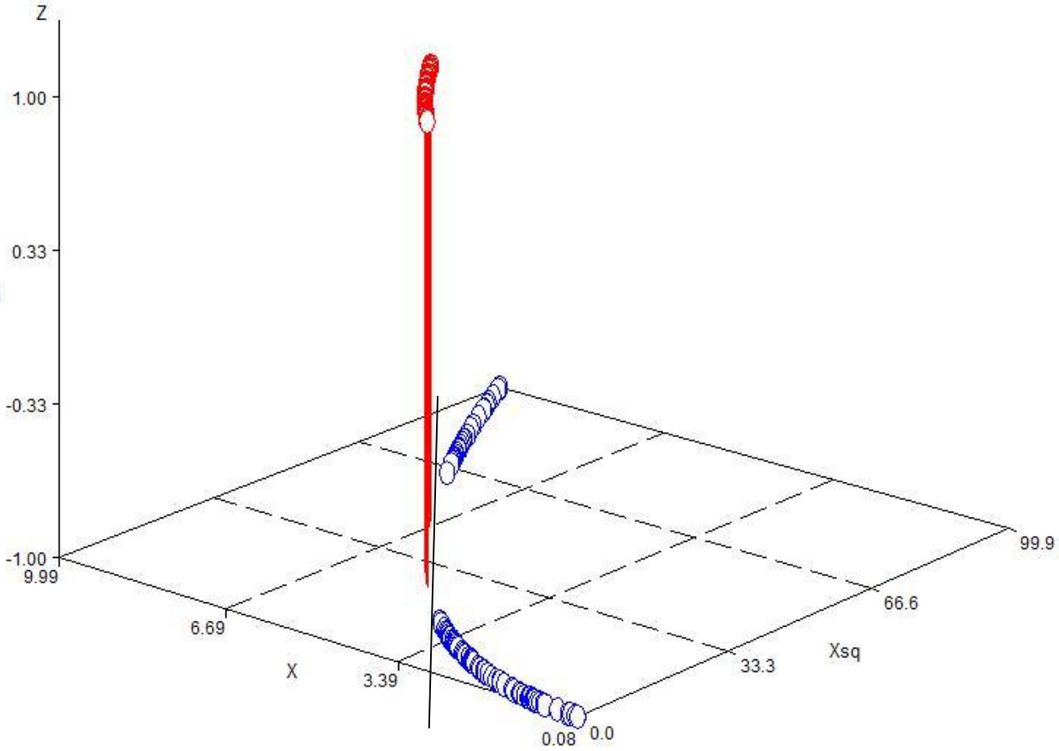
$Y= (-67 + 24X_1 – 2X_2 )/3$

$X_2=-33.5+12X_1$

(where Y=0)

Idea: plot Z against X1=X and X2 = $X^2$
Move to “higher dimension” to get linear separation (plane) $Y= (-67 + 24X_1 – 2X_2 )/3$



Look into the floor

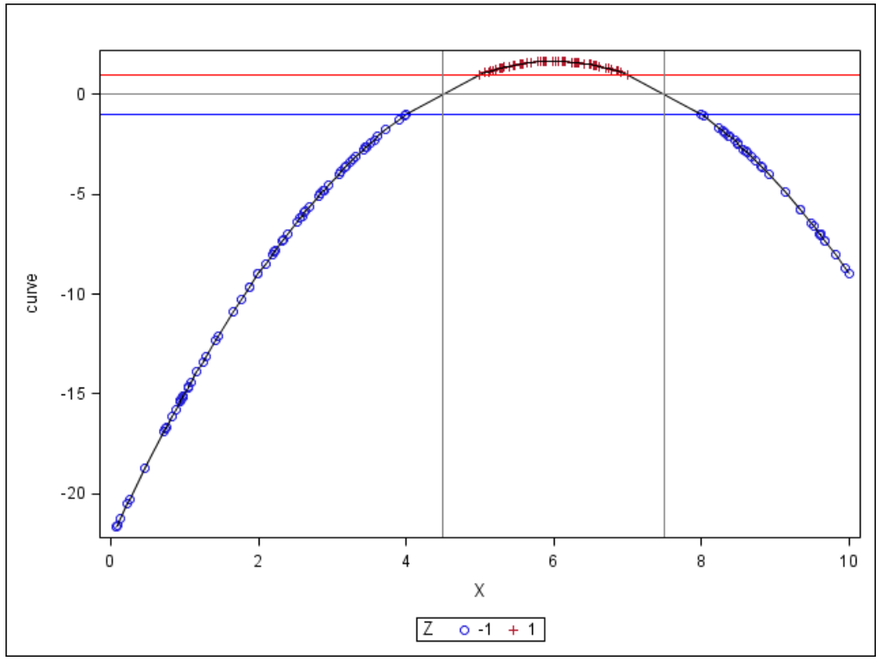
Plane hits Z=-1 at X=4 and 8, (b+4c= -1)

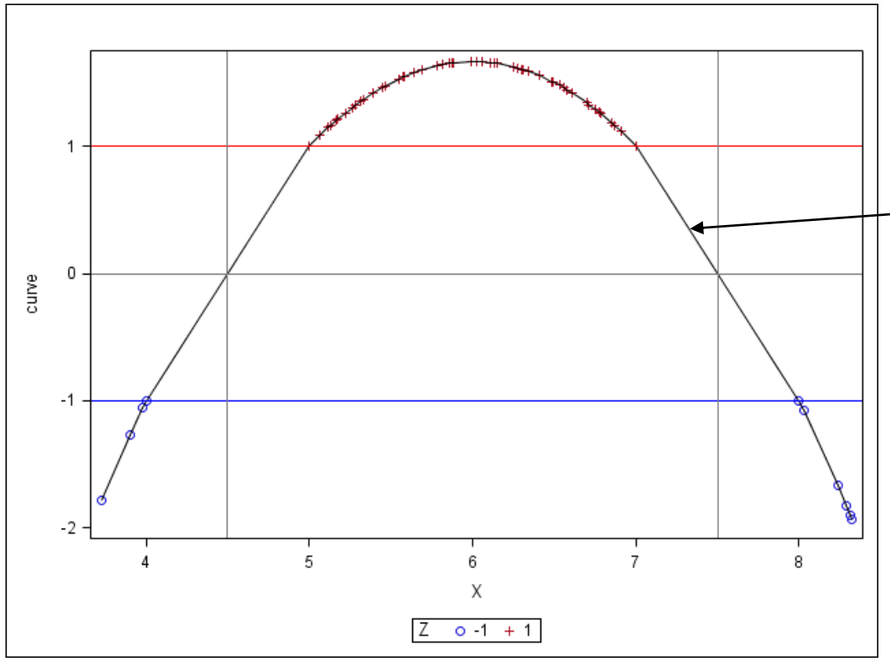
Z=1 at X=5 and 7 (b+c=1)

Corresponding quadratic: $f(X,X^2) = b + c(X-6)^2$ b+4c=1, b+c=-1, c=-2/3, b=-5/3

$$f(X,X^2) = -(2X^2-24X+67)/3$$

Pay off card Pay interest only default Pay off card Pay interest only default

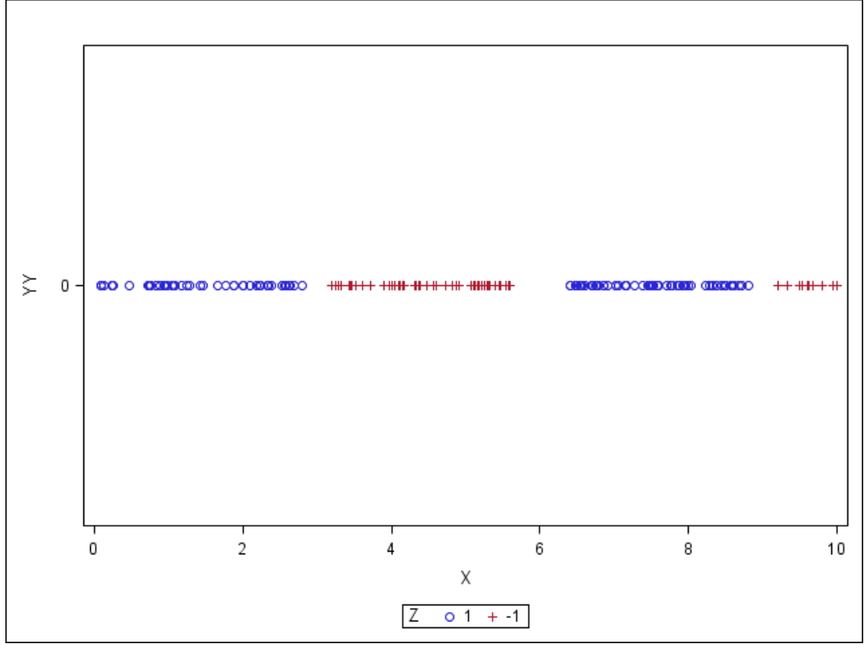


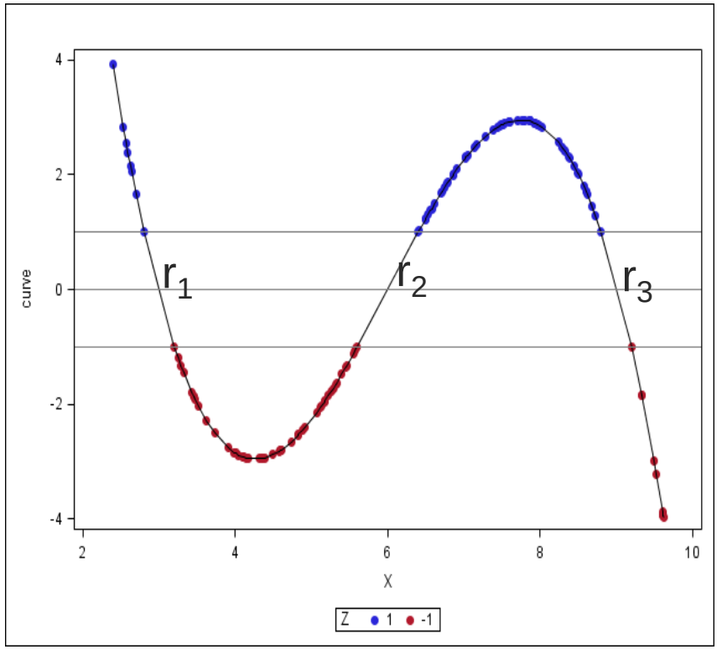




Line
Y = -2(X-7.5)
Not quadratic

Plane: Y=0 at X=7.5
Quadratic: Y=0 at X=7.58 (root)

Note: Quadratic $(2X^2-24X+67)/3$ and plane 2(X-7.5) hit 1 and -1 at the same X values



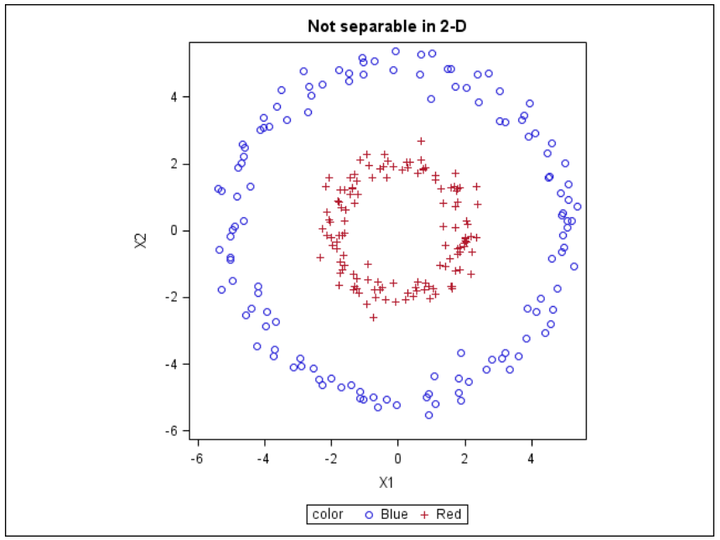

Existence:
Can always split with

$Y=c(X-r_1)(X-r_2)(X-r_3)$

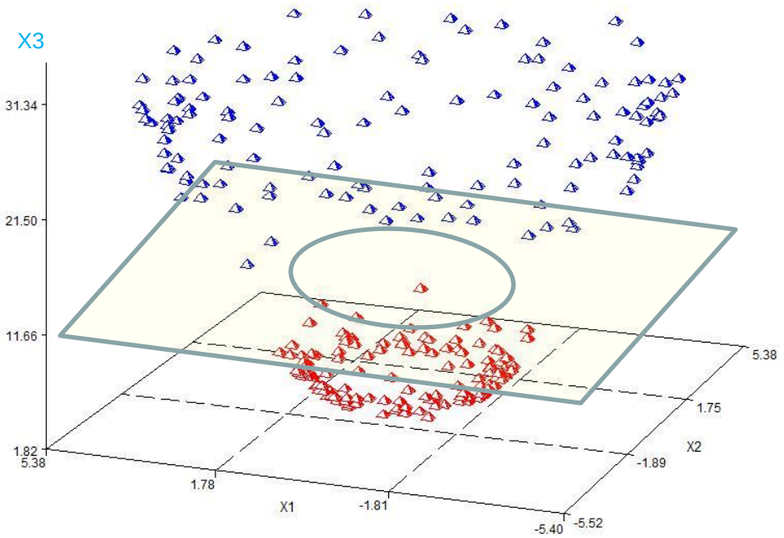
Best Y?
Maximizes margin?

Dimensions:
X1=X
$X2=X^2$
$X3=X^3$

# Example in 2-D




X3=X1*X1+X2*X2

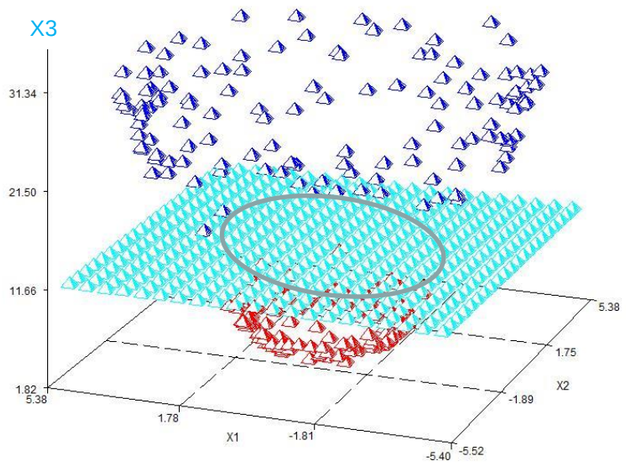


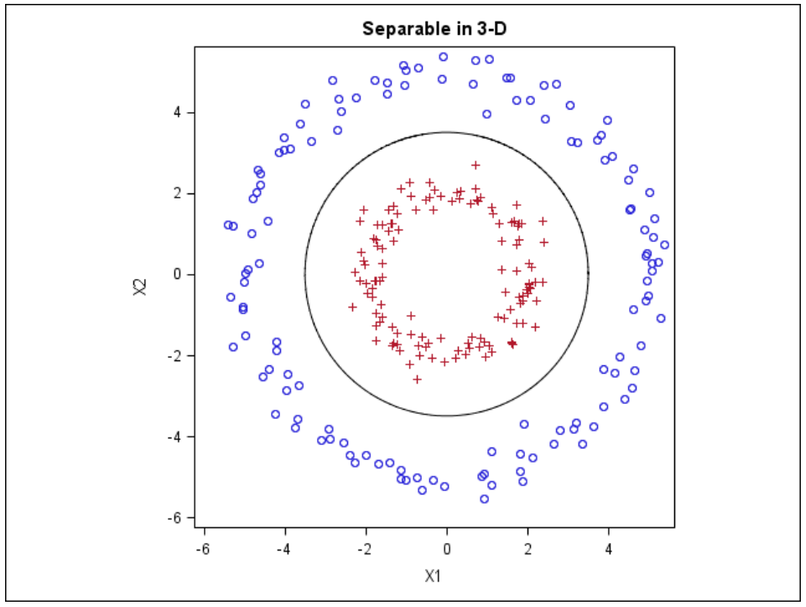

Dimensions:
(X1,X2,X3,Y)

X1, X2,
X3=X1*X1+X2*X2

Z=-1 or 1
(no Z axis shown)

X3=X1*X1+X2*X2



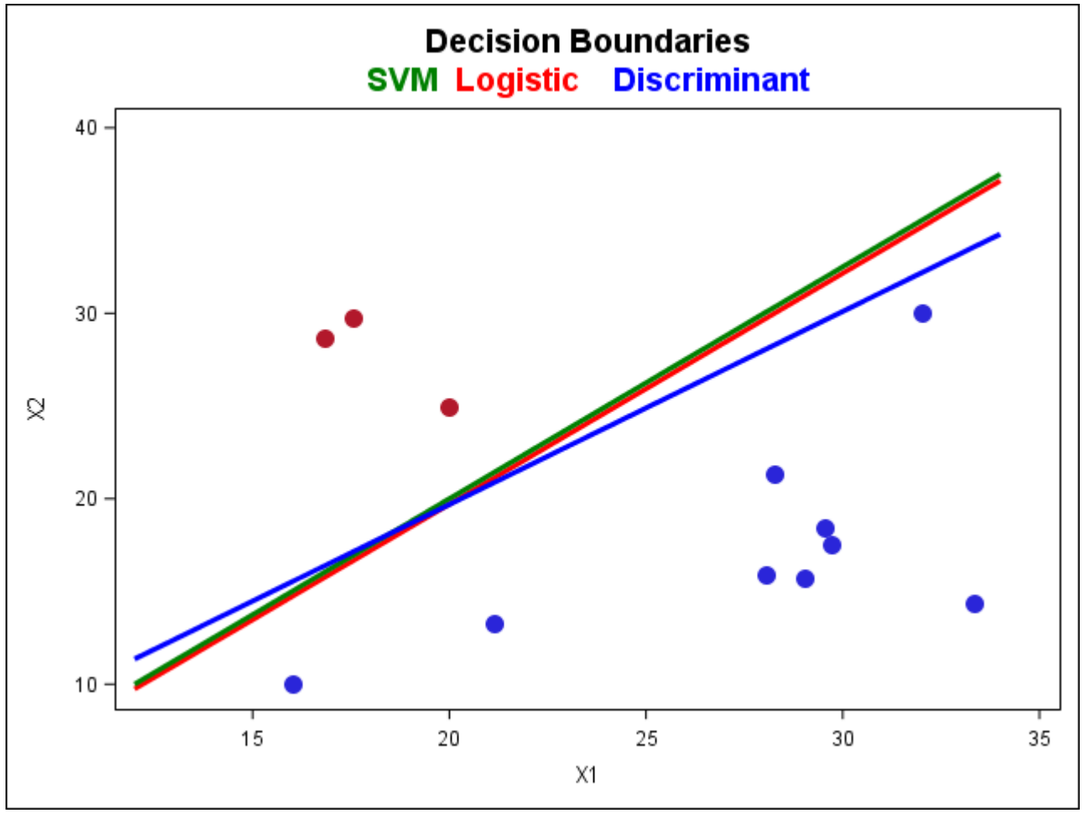



Reality:  Events and non-events typically mingled

Need to lighten up on ZY-1 >= 0 requirement !

This plus the move to higher dimension is full blown support vector technology.

# Comparing Decision Boundaries for SVM Data




Warning message from PROC LOGISTIC

Model Convergence Status

**Complete separation of data points detected**.

WARNING: The maximum likelihood estimate does not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

# Text Mining

Hypothetical collection of news releases (“corpus”) :

release 1: Did the NCAA investigate the basketball scores and vote for sanctions?
release 2: Republicans voted for and Democrats voted against it for the win.
(etc.)

Compute word counts:

| | NCAA | basketball | score | vote | Republican | Democrat | win |
|---|---|---|---|---|---|---|---|
| Release 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| Release 2 | 0 | 0 | 0 | 2 | 1 | 1 | 1 |

# Text Mining Mini-Example: Word counts in 14 e-mails

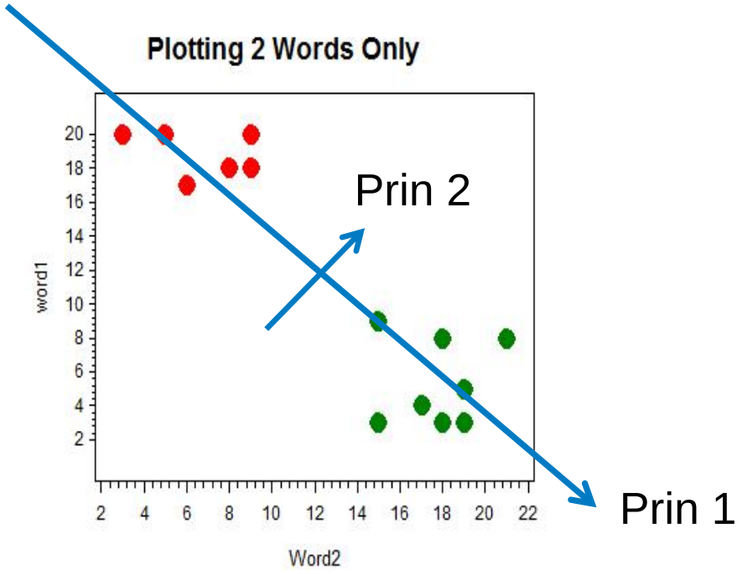
←----------words----------→

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R | B | | | | T | | | | | | | |
| | P | e | a | | | | o | | | | | | | |
| d | E | r | p | s | D | | | u | | | | | | |
| o | l | e | u | k | e | | | r | | S | S | | | |
| c | e | s | b | e | m | V | | n | S | | c | c | | |
| u | c | i | l | t | o | o | | a | p | | o | o | | |
| m | t | d | i | b | c | t | N | L | m | e | W | r | r | |
| e | i | e | c | a | r | e | C | i | e | e | i | e | e | |
| n | o | n | a | l | a | r | A | a | n | c | n | _ | _ | |
| t | n | t | n | l | t | s | A | r | t | h | s | V | N | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 8 | 10 | 12 | 6 | 0 | 1 | 5 | 3 | 8 | 18 | 15 | 21 | |
| 2 | 5 | 6 | 9 | 5 | 4 | 2 | 0 | 9 | 0 | 12 | 12 | 9 | 0 | |
| 3 | 0 | 2 | 0 | 14 | 0 | 2 | 12 | 0 | 16 | 4 | 24 | 19 | 30 | |
| 4 | 8 | 9 | 7 | 0 | 12 | 14 | 2 | 12 | 3 | 15 | 22 | 8 | 2 | |
| 5 | 0 | 0 | 4 | 16 | 0 | 0 | 15 | 2 | 17 | 3 | 9 | 0 | 1 | |
| 6 | 10 | 6 | 9 | 5 | 5 | 19 | 5 | 20 | 0 | 18 | 13 | 9 | 14 | |
| 7 | 2 | 3 | 1 | 13 | 0 | 1 | 12 | 13 | 20 | 0 | 0 | 1 | 6 | |
| 8 | 4 | 1 | 4 | 16 | 2 | 4 | 9 | 0 | 12 | 9 | 3 | 0 | 0 | |
| 9 | 26 | 13 | 9 | 2 | 16 | 20 | 6 | 24 | 4 | 30 | 9 | 10 | 14 | |
| 10 | 19 | 22 | 10 | 11 | 9 | 12 | 0 | 14 | 10 | 22 | 3 | 1 | 0 | |
| 11 | 2 | 0 | 0 | 14 | 1 | 3 | 12 | 0 | 16 | 12 | 17 | 23 | 8 | |
| 12 | 16 | 19 | 21 | 0 | 13 | 9 | 0 | 16 | 4 | 12 | 0 | 0 | 2 | |
| 13 | 14 | 17 | 12 | 0 | 20 | 19 | 0 | 12 | 5 | 9 | 6 | 1 | 4 | |
| 14 | 1 | 0 | 4 | 21 | 3 | 6 | 9 | 3 | 8 | 0 | 3 | 10 | 20 | |

Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 7.10954264 | 4.80499109 | 0.5469 | 0.5469 |
| 2 | 2.30455155 | 1.30162837 | 0.1773 | 0.7242 |
| 3 | 1.00292318 | 0.23404351 | 0.0771 | 0.8013 |
| 4 | 0.76887967 | 0.21070080 | 0.0591 | 0.8605 |
| 5 | 0.55817886 | 0.10084923 | 0.0429 | 0.9034 |
| | | (more) | | |
| 13 | 0.0008758 | | 0.0001 | 1.0000 |

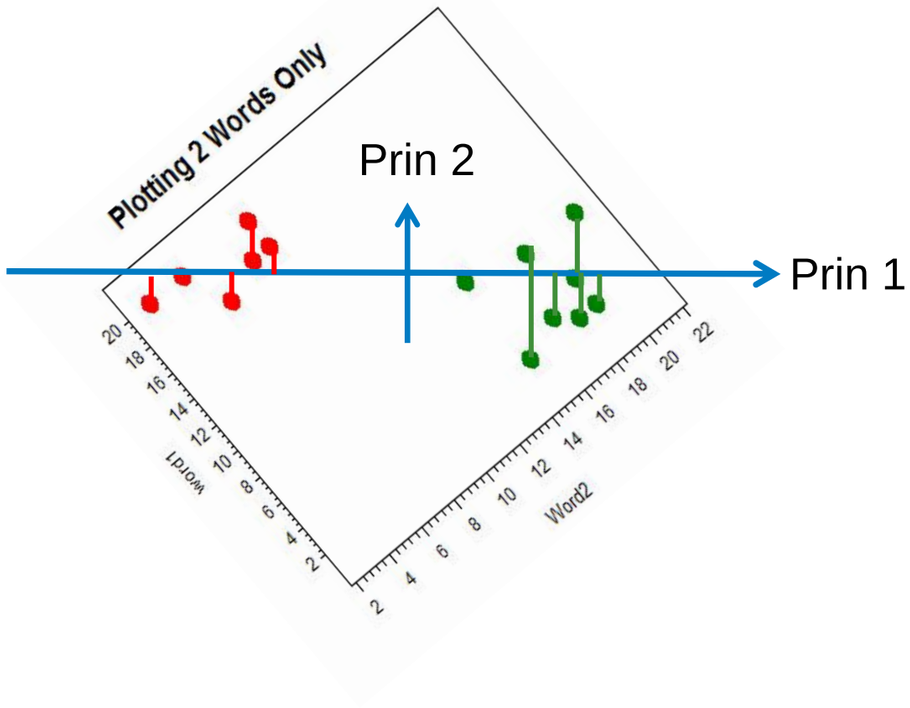



55% of the variation in these 13-dimensional vectors occurs in one dimension.

| Variable | Prin1 |
|---|---|
| Basketball | -.320074 |
| NCAA | -.314093 |
| Tournament | -.277484 |
| Score_V | -.134625 |
| Score_N | -.120083 |
| Wins | -.080110 |
| Speech | 0.273525 |
| Voters | 0.294129 |
| Liar | 0.309145 |
| Election | 0.315647 |
| Republican | 0.318973 |
| President | 0.333439 |
| Democrat | 0.336873 |

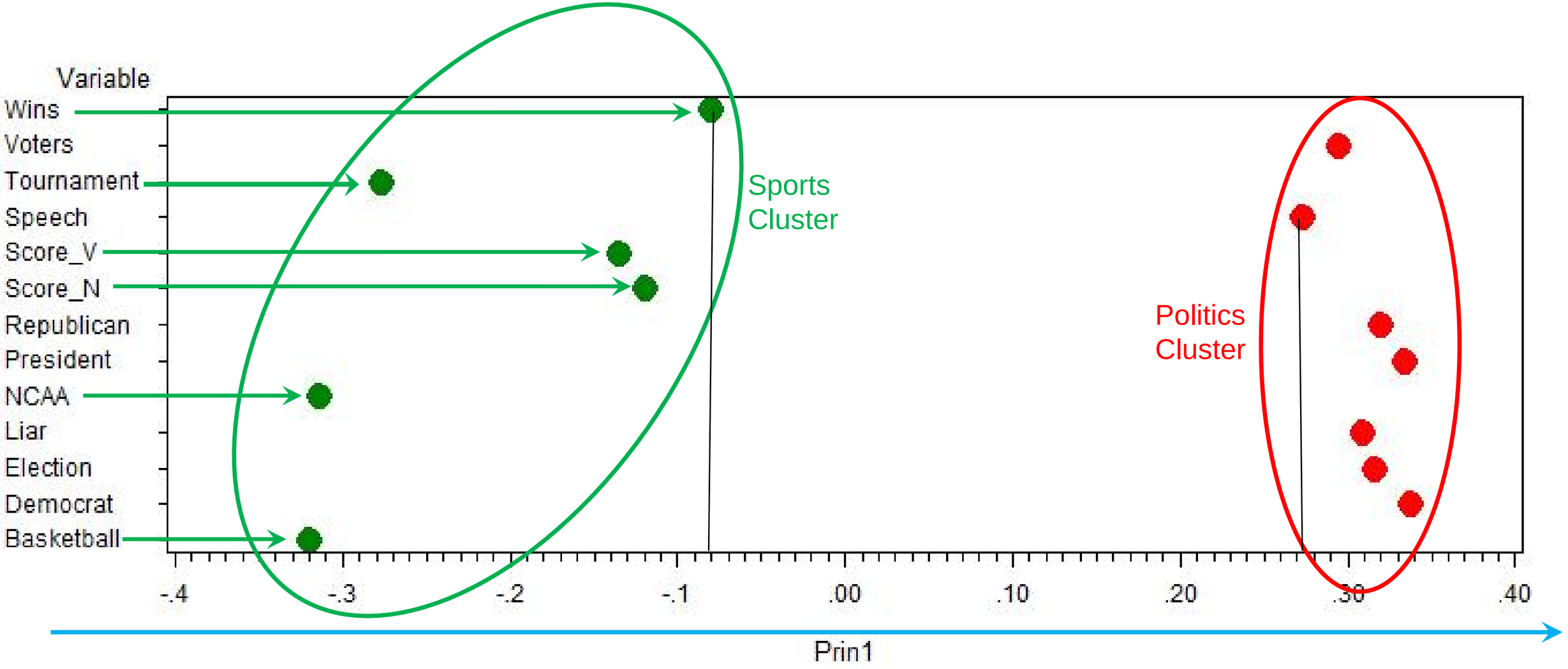
Eigenvalues of the Correlation Matrix

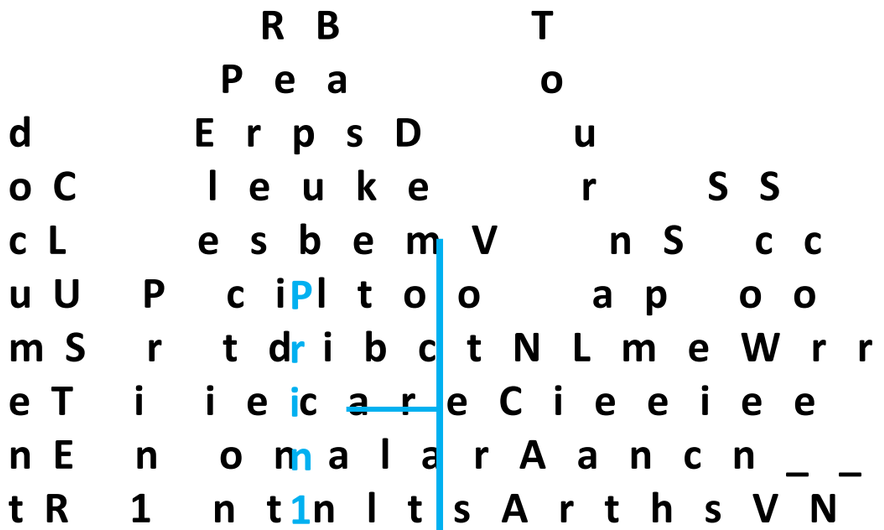
| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 7.10954264 | 4.80499109 | 0.5469 | 0.5469 |
| 2 | 2.30455155 | 1.30162837 | 0.1773 | 0.7242 |
| 3 | 1.00292318 | 0.23404351 | 0.0771 | 0.8013 |
| 4 | 0.76887967 | 0.21070080 | 0.0591 | 0.8605 |
| 5 | 0.55817886 | 0.10084923 | 0.0429 | 0.9034 |
| | | (more) | | |
| 13 | 0.0008758 | | 0.0001 | 1.0000 |




55% of the variation in these 13-dimensional vectors occurs in one dimension.

| Variable | Prin1 |
|---|---|
| Basketball | -.320074 |
| NCAA | -.314093 |
| Tournament | -.277484 |
| Score_V | -.134625 |
| Score_N | -.120083 |
| Wins | -.080110 |
| Speech | 0.273525 |
| Voters | 0.294129 |
| Liar | 0.309145 |
| Election | 0.315647 |
| Republican | 0.318973 |
| President | 0.333439 |
| Democrat | 0.336873 |


Variable
Wins
Voters
Tournament
Speech
Score_V
Score_N
Republican
President
NCAA
Liar
Election
Democrat
Basketball
Sports
Cluster
Politics
Cluster
-.4
-.3
-.2
-.1
.00
.10
.20
.30
.40
Prin1



Sports Documents

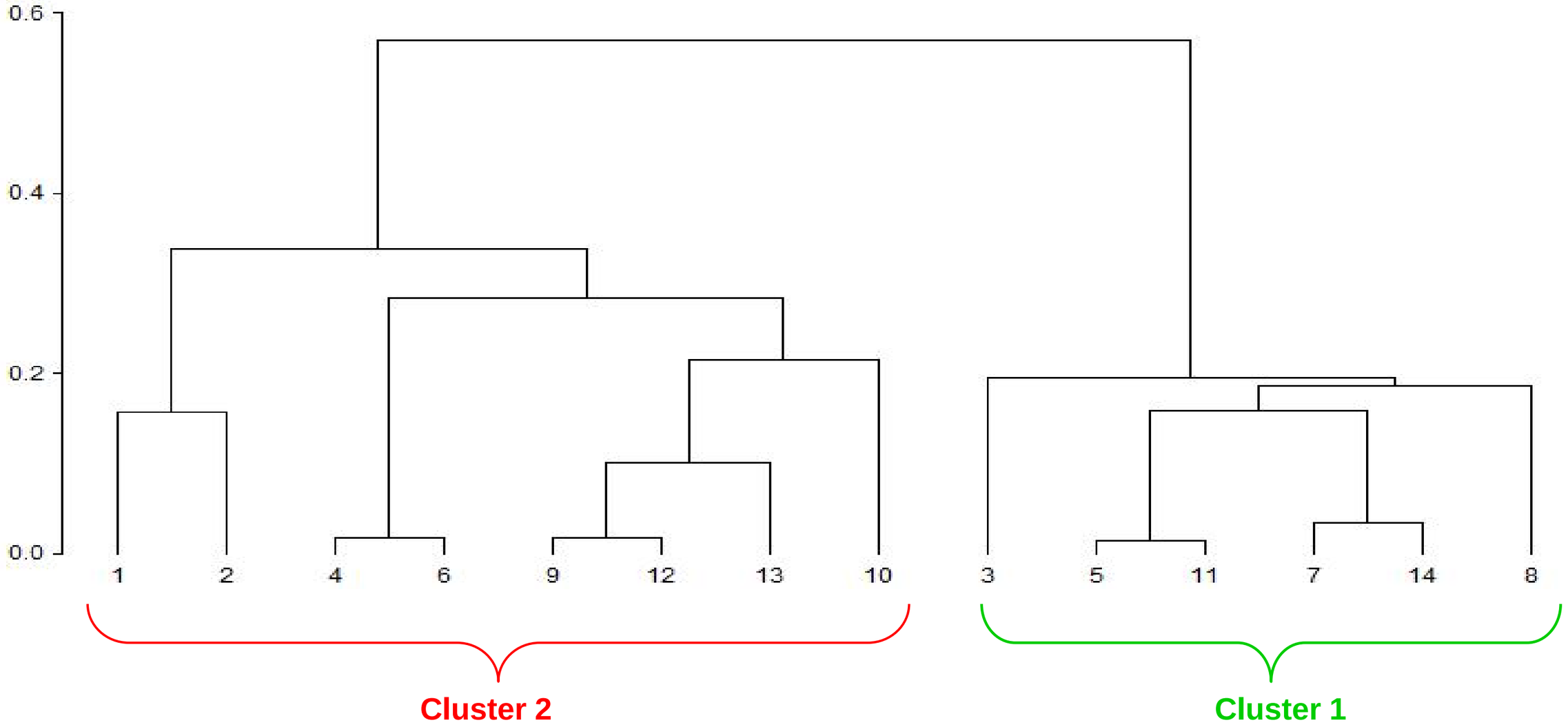
| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | -3.63815 | 0 | 2 | 0 | 14 | 0 | 2 | 12 | 0 | 16 | 4 | 24 | 19 | 30 |
| 11 | 1 | -3.02803 | 2 | 0 | 0 | 14 | 1 | 3 | 12 | 0 | 16 | 12 | 17 | 23 | 8 |
| 5 | 1 | -2.98347 | 0 | 0 | 4 | 16 | 0 | 0 | 15 | 2 | 17 | 3 | 9 | 0 | 1 |
| 14 | 1 | -2.48381 | 1 | 0 | 4 | 21 | 3 | 6 | 9 | 3 | 8 | 0 | 3 | 10 | 20 |
| 7 | 1 | -2.37638 | 2 | 3 | 1 | 13 | 0 | 1 | 12 | 13 | 20 | 0 | 0 | 1 | 6 |
| 8 | 1 | -1.79370 | 4 | 1 | 4 | 16 | 2 | 4 | 9 | 0 | 12 | 9 | 3 | 0 | 0 |

(biggest gap)

Politics Documents

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | -0.00738 | 20 | 8 | 10 | 12 | 6 | 0 | 1 | 5 | 3 | 8 | 18 | 15 | 21 |
| 2 | 2 | 0.48514 | 5 | 6 | 9 | 5 | 4 | 2 | 0 | 9 | 0 | 12 | 12 | 9 | 0 |
| 6 | 2 | 1.54559 | 10 | 6 | 9 | 5 | 5 | 19 | 5 | 20 | 0 | 18 | 13 | 9 | 14 |
| 4 | 2 | 1.59833 | 8 | 9 | 7 | 0 | 12 | 14 | 2 | 12 | 3 | 15 | 22 | 8 | 2 |
| 10 | 2 | 2.49069 | 19 | 22 | 10 | 11 | 9 | 12 | 0 | 14 | 10 | 22 | 3 | 1 | 0 |
| 13 | 2 | 3.16620 | 14 | 17 | 12 | 0 | 20 | 19 | 0 | 12 | 5 | 9 | 6 | 1 | 4 |
| 12 | 2 | 3.48420 | 16 | 19 | 21 | 0 | 13 | 9 | 0 | 16 | 4 | 12 | 0 | 0 | 2 |
| 9 | 2 | 3.54077 | 26 | 13 | 9 | 2 | 16 | 20 | 6 | 24 | 4 | 30 | 9 | 10 | 14 |

# PROC CLUSTER (single linkage) agrees !



# Fisher's Linear Discriminant Analysis

## - an older method of classification

## (optional)

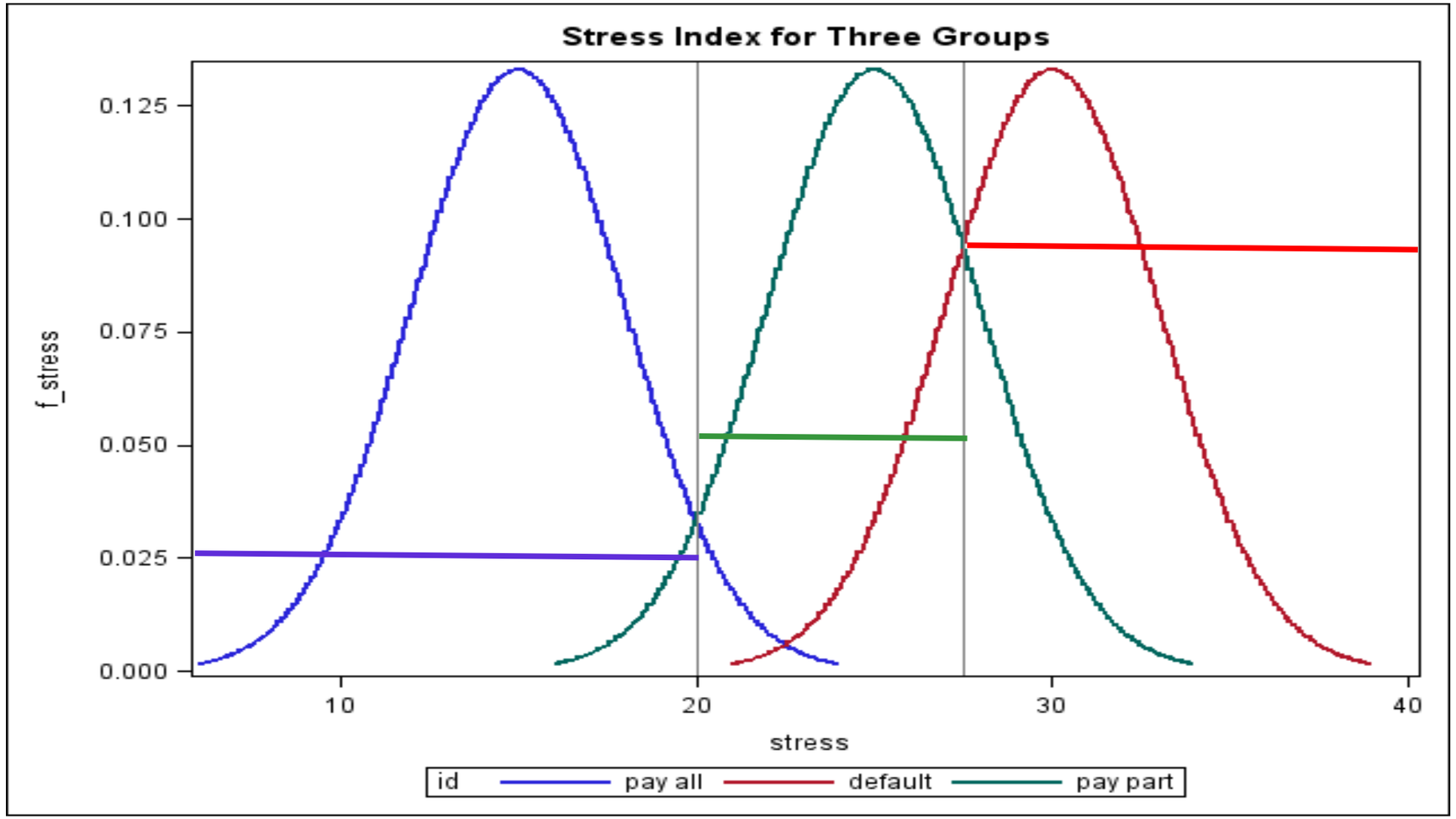
Assumes multivariate normal distribution of inputs (features)

Computes a "posterior probability" for each of several classes, given the observations.

Based on statistical theory

Example:
One input,
(financial
stress index)

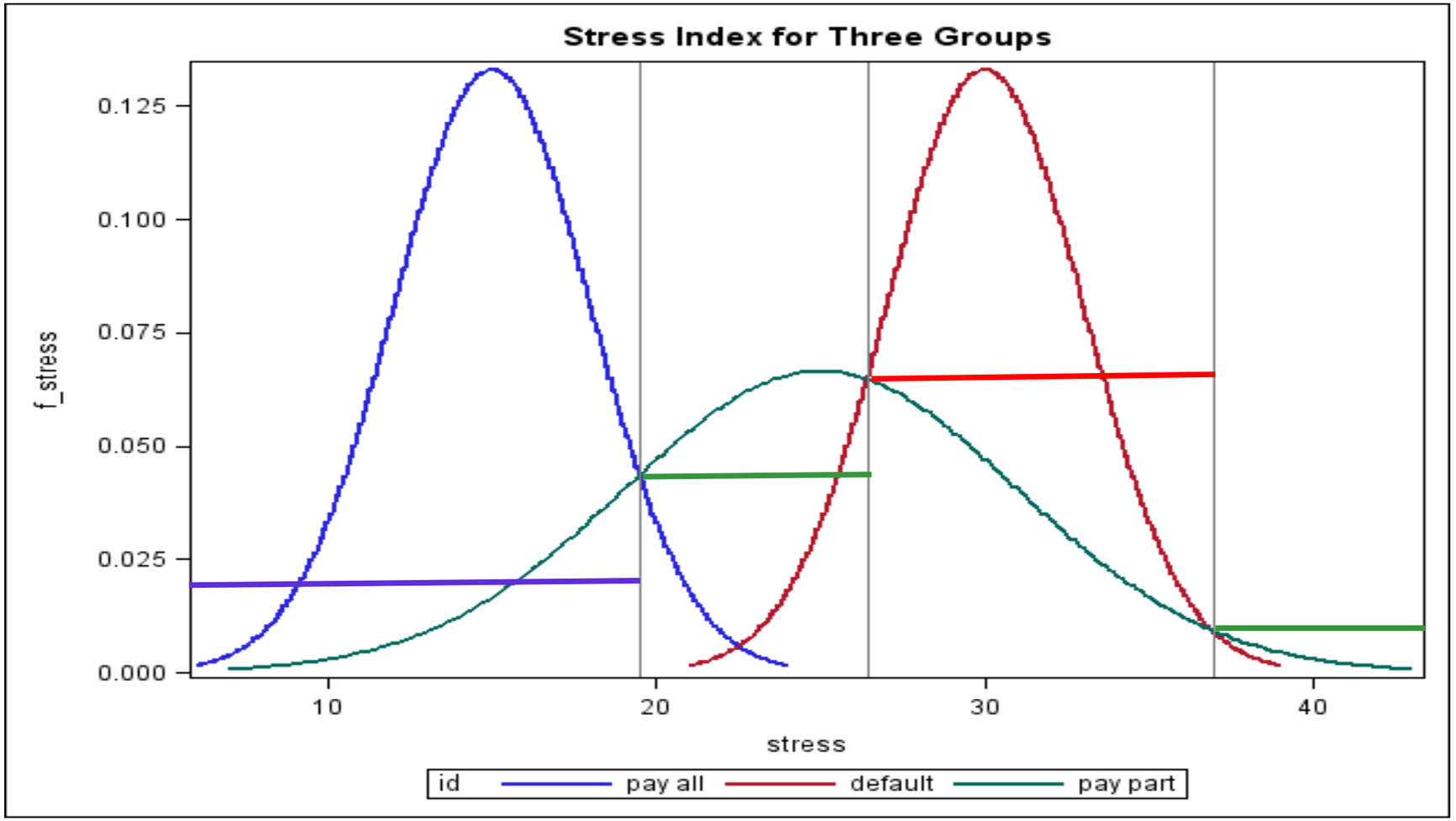
Three
Populations:
(pay all credit
card debt,
pay part,
default).




Normal distributions with same variance. Classify a credit card holder by financial stress index: pay all (stress<20), pay part (20<stress<27.5), default(stress>27.5)
Data are hypothetical. Means are 15, 25, 30.

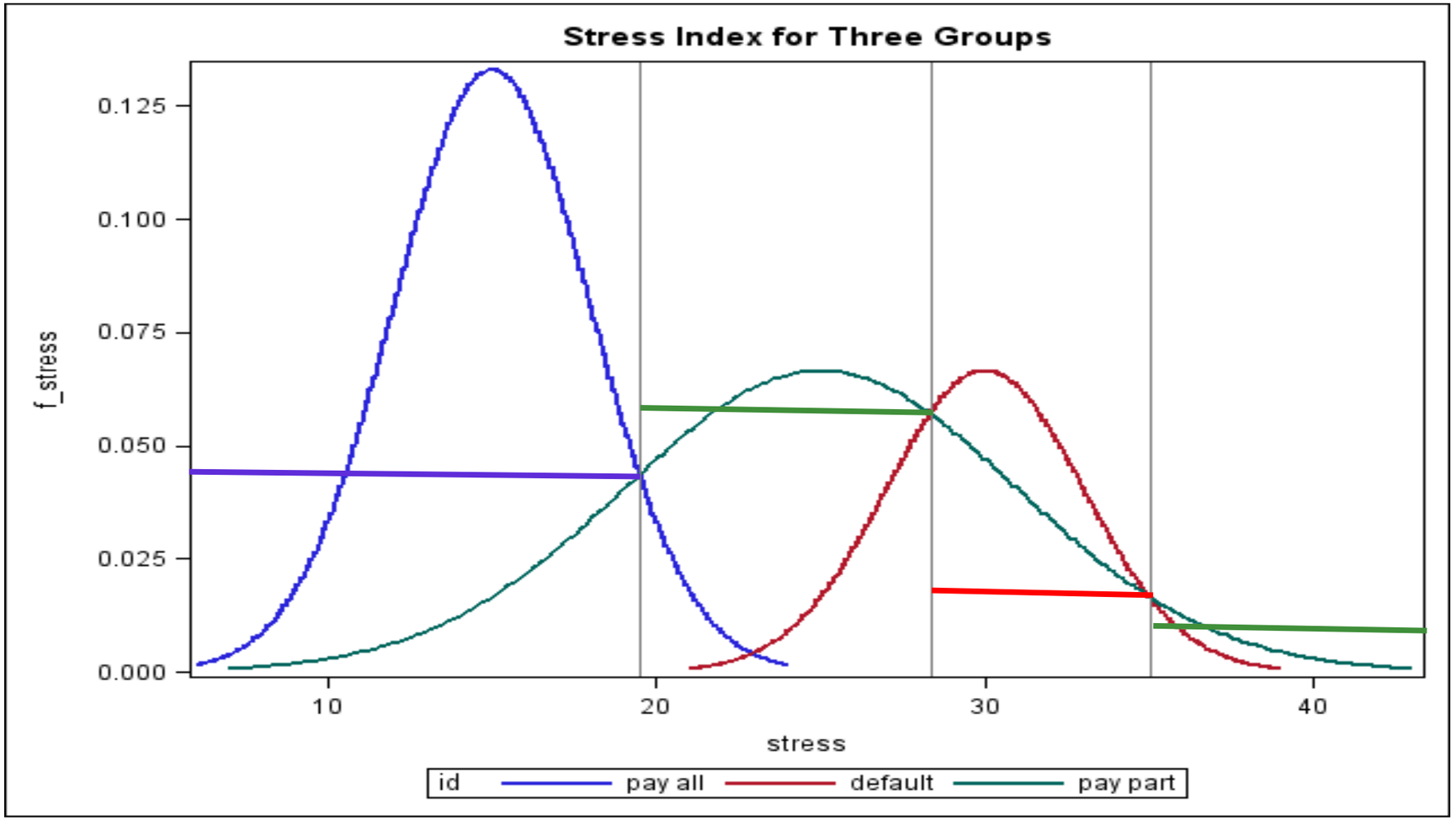
# Example: Differing variances.

Not just closest mean now.




Normal distributions with different variances. Classify a credit card holder by financial stress index: pay all (stress<19.48), pay part (19.48<stress<26.40), default(26.40<stress<36.93), pay part (stress>36.93)
Data are hypothetical. Means are 15, 25, 30. Standard Deviations 3,6,3.

Example:
20% defaulters,
40% pay part
40% pay all

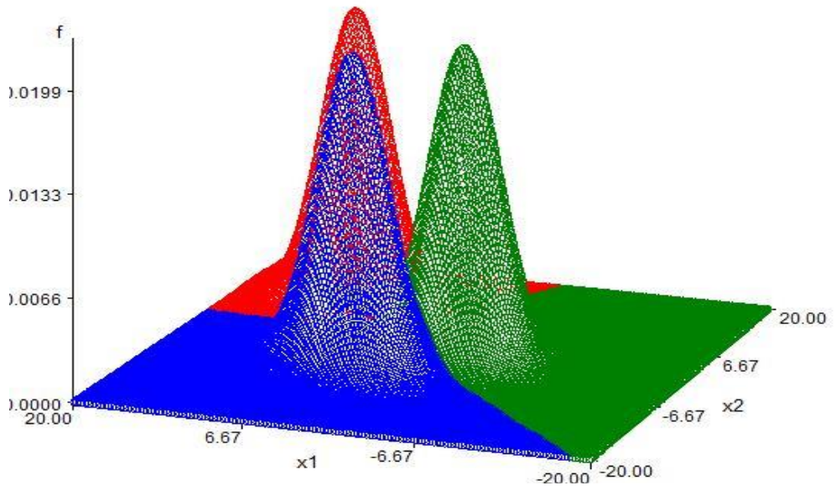



Normal distributions with same variance and “priors.” Classify a credit card holder by financial stress index: pay all (stress<19.48), pay part (19.48<stress<28.33), default(28.33<stress<35.00), pay part (stress>35.00) Data are hypothetical. Means are 15, 25, 30. Standard Deviations 3,6,3. Population size ratios 2:2:1

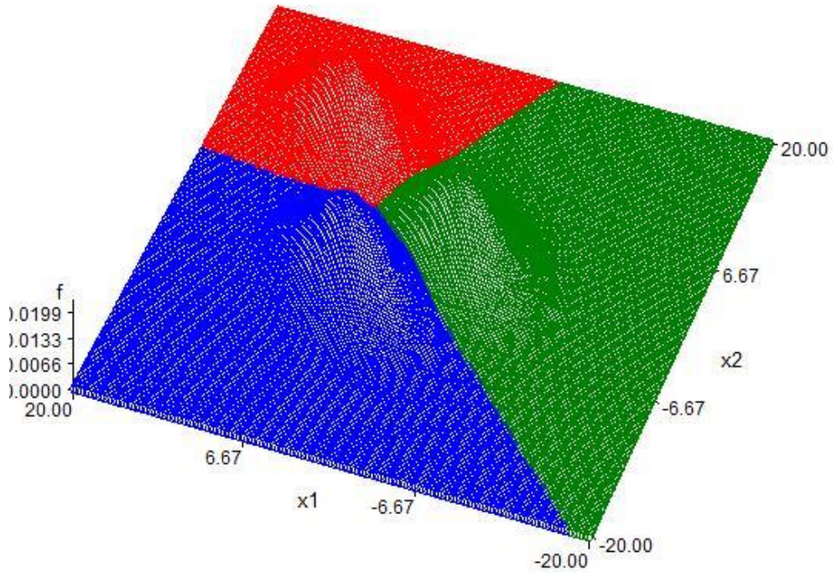
Example:
Two inputs, three populations

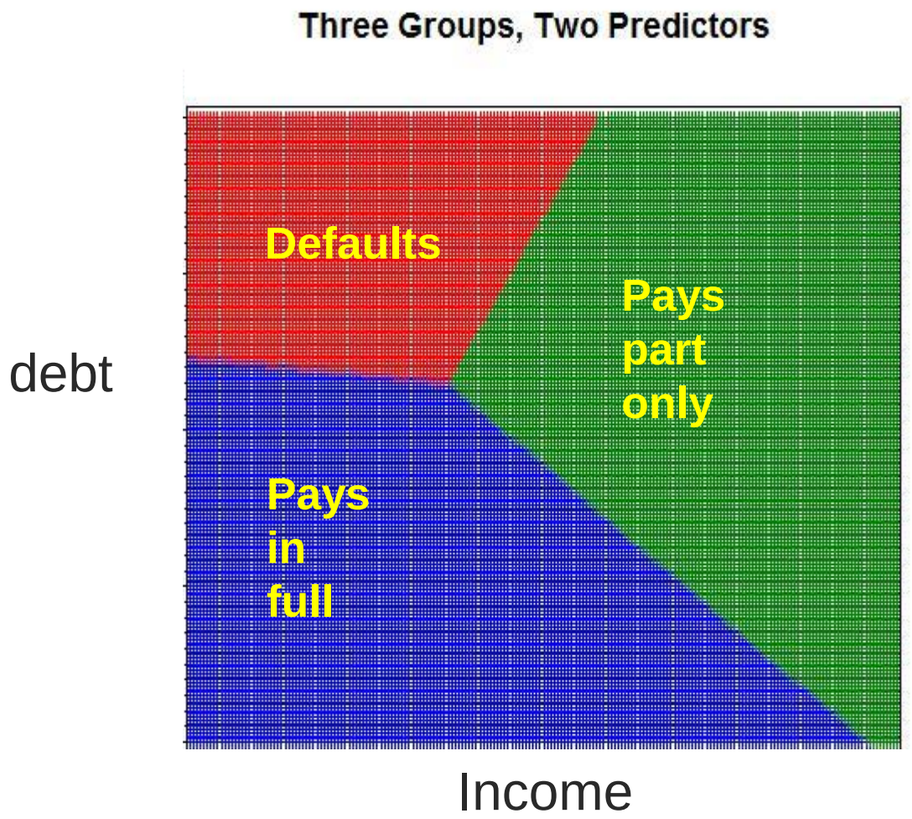
Multivariate normals, same 2x2 covariance matrix.

Viewed from above, boundaries appear to be linear.

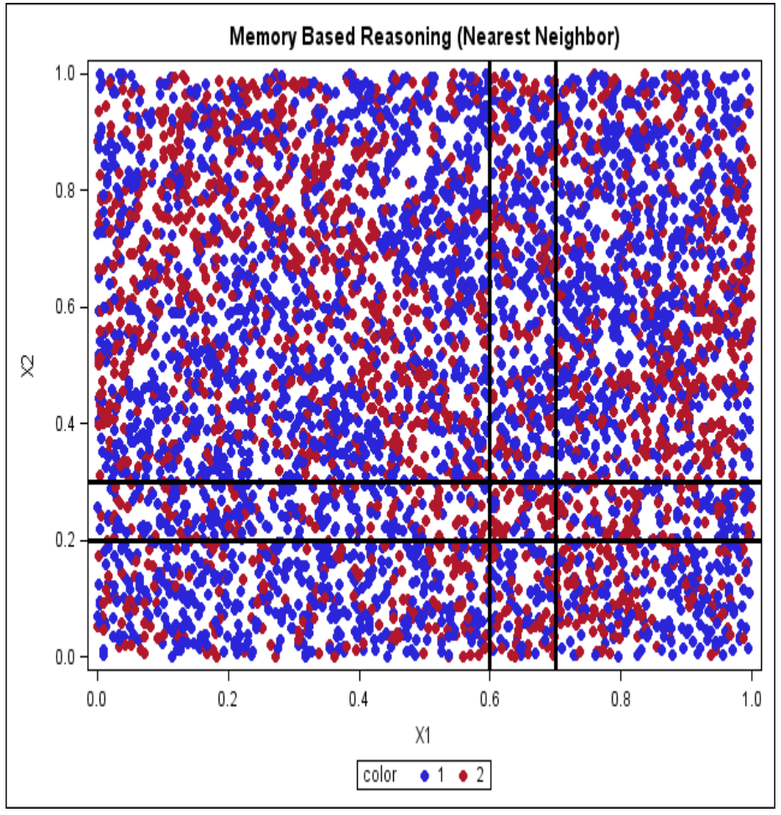


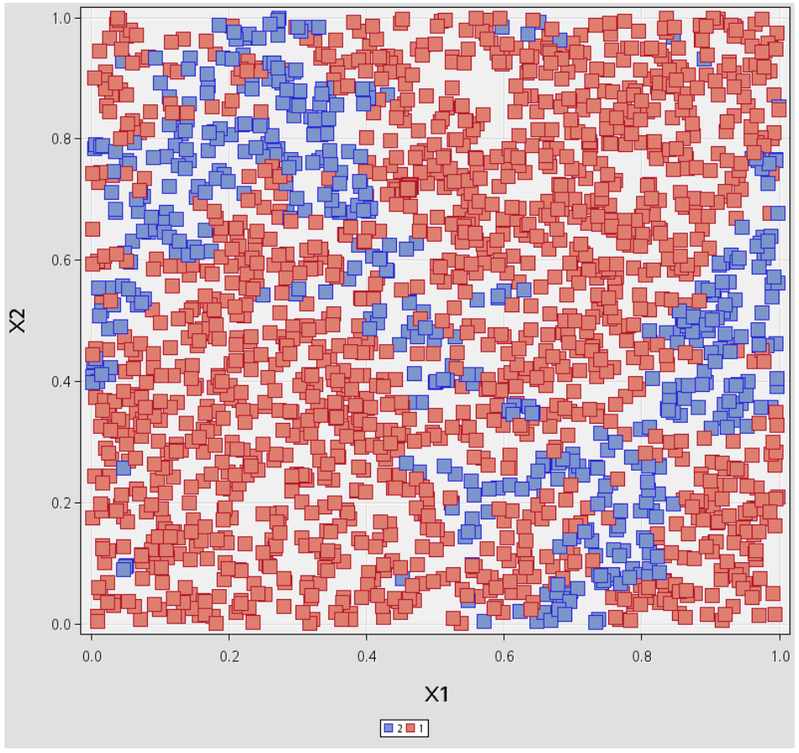

Three Groups, Two Predictors




Three Groups, Two Predictors
debt
Defaults
Pays part only
Pays in full
Income

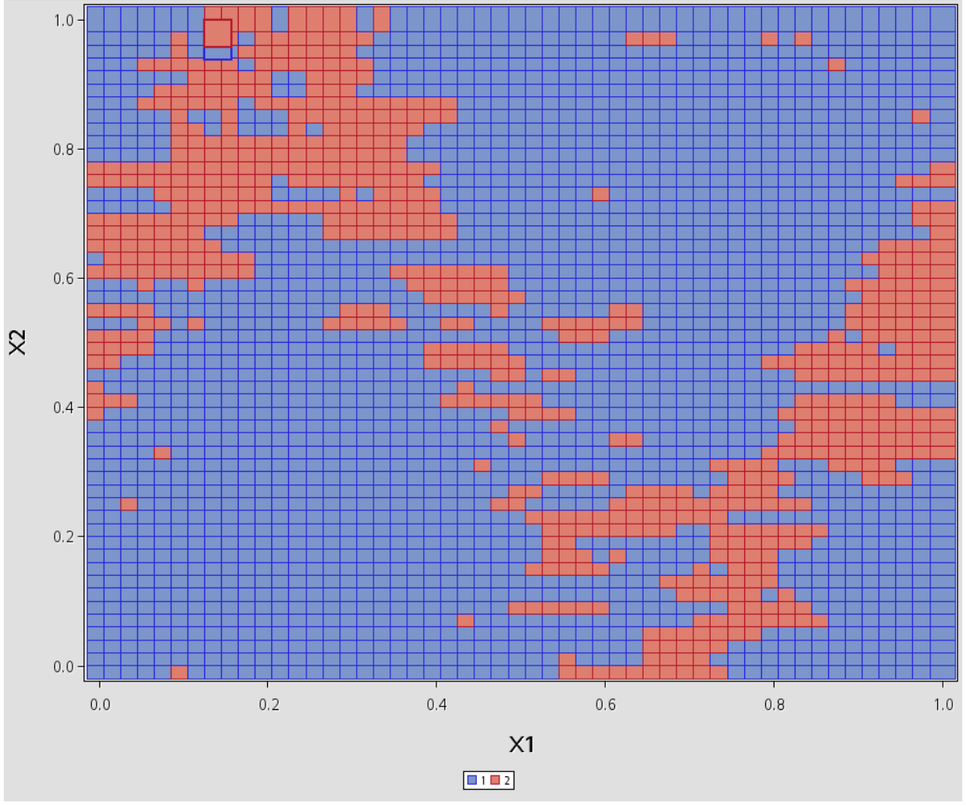
Original Data



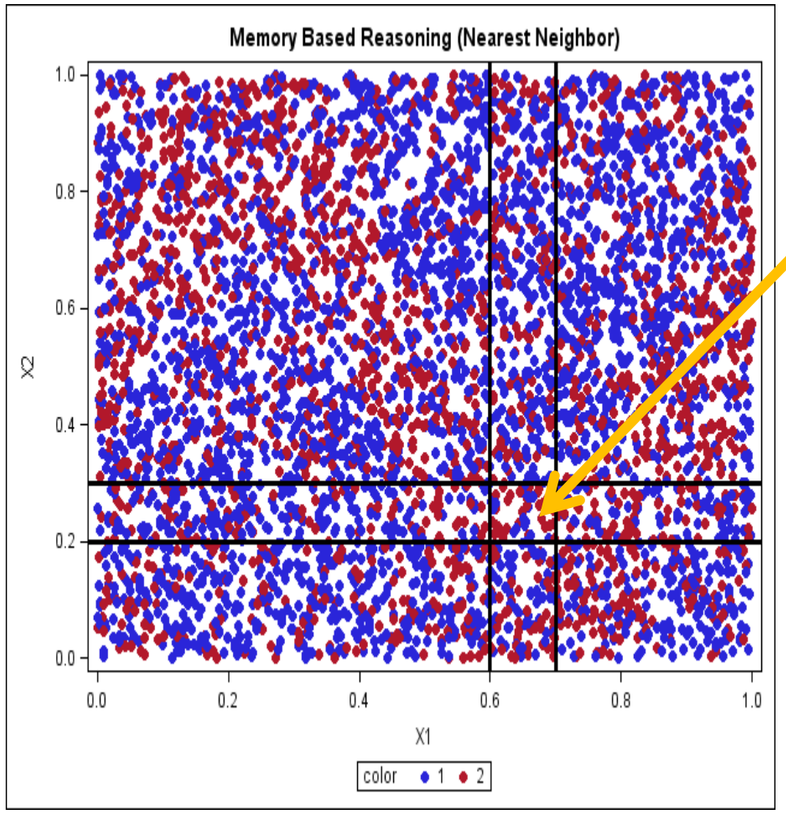




Original Data scored by Enterprise Miner Nearest Neighbor method

Score Data (grid) scored by Enterprise Miner Nearest Neighbor method



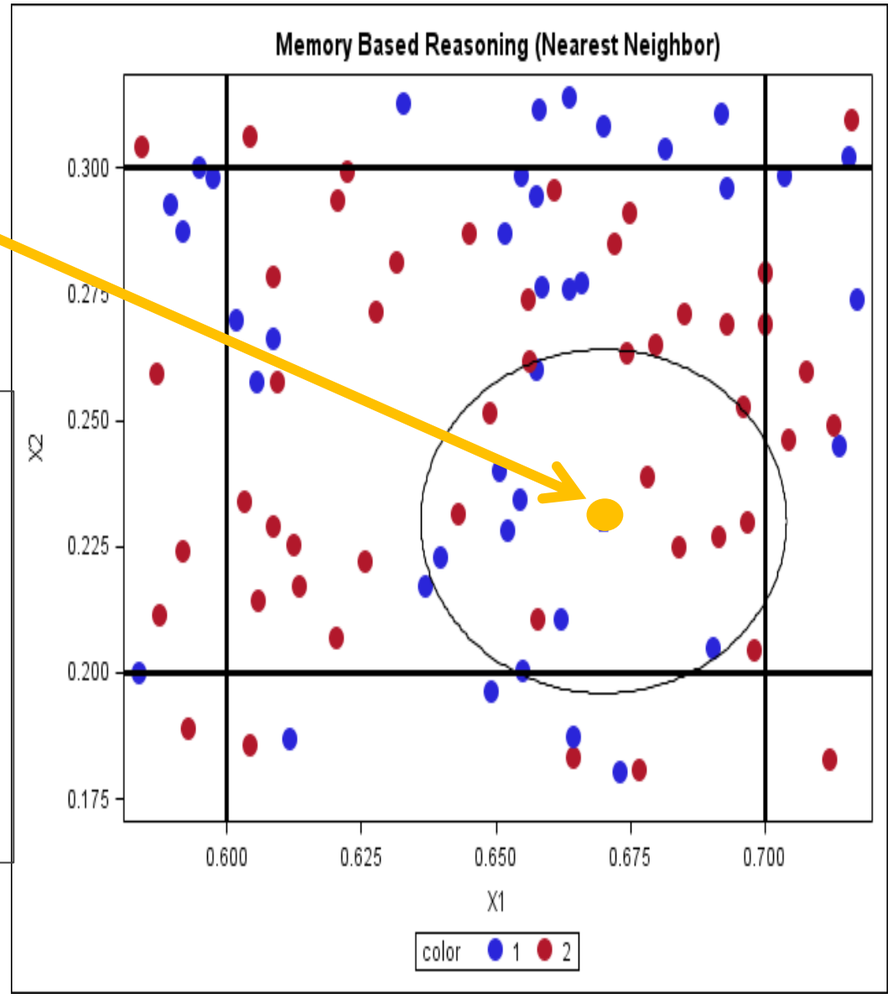
# Memory Based Reasoning
# (optional)
# Usual name: Nearest Neighbor Analysis

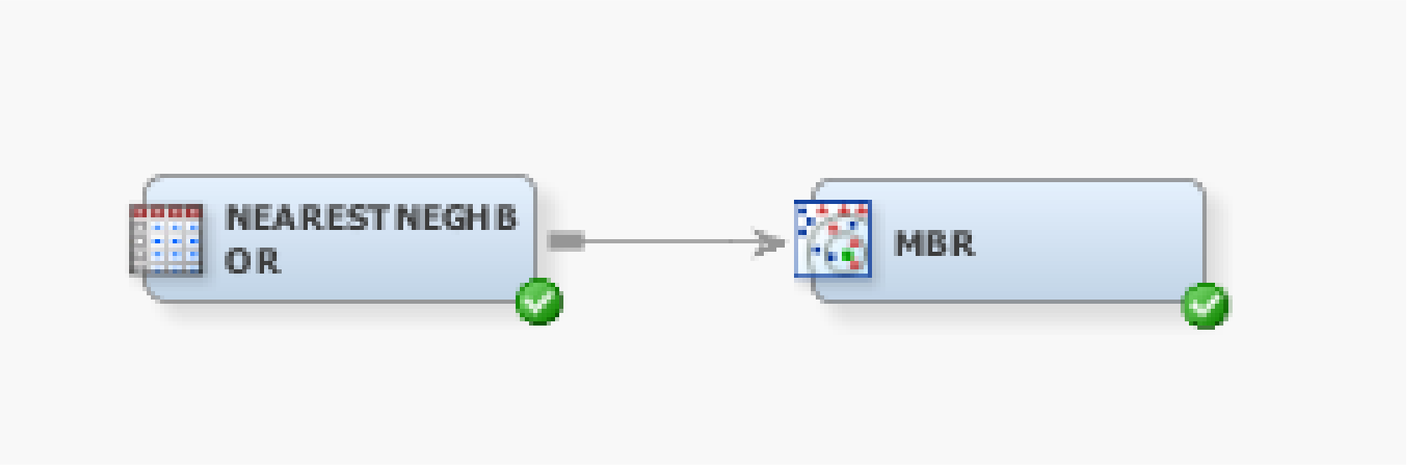


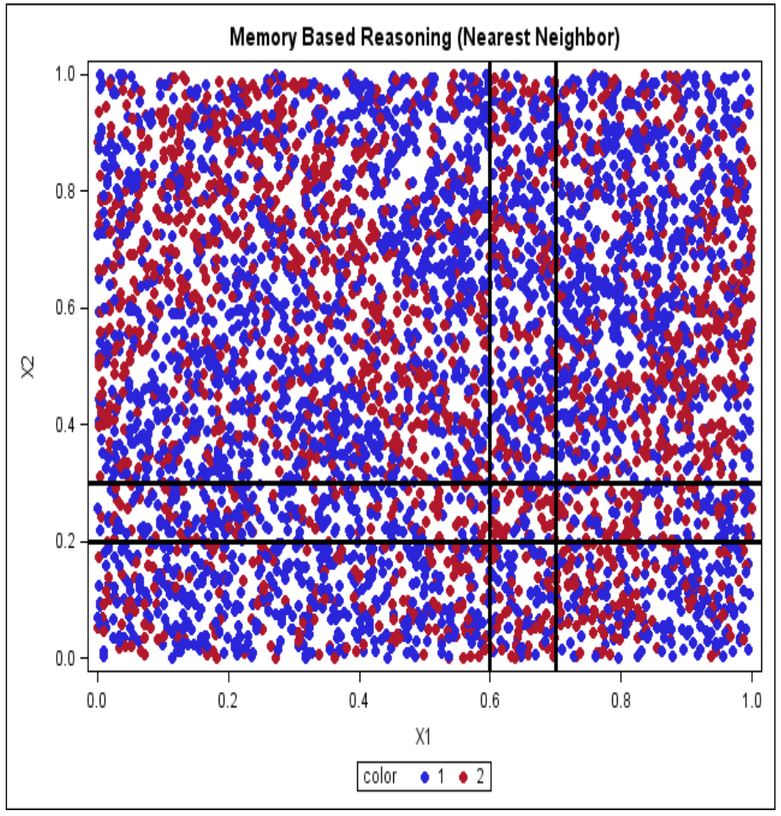

Probe Point

9 blue
7red

Classify as
BLUE
Estimate Pr{Blue}
as 9/16

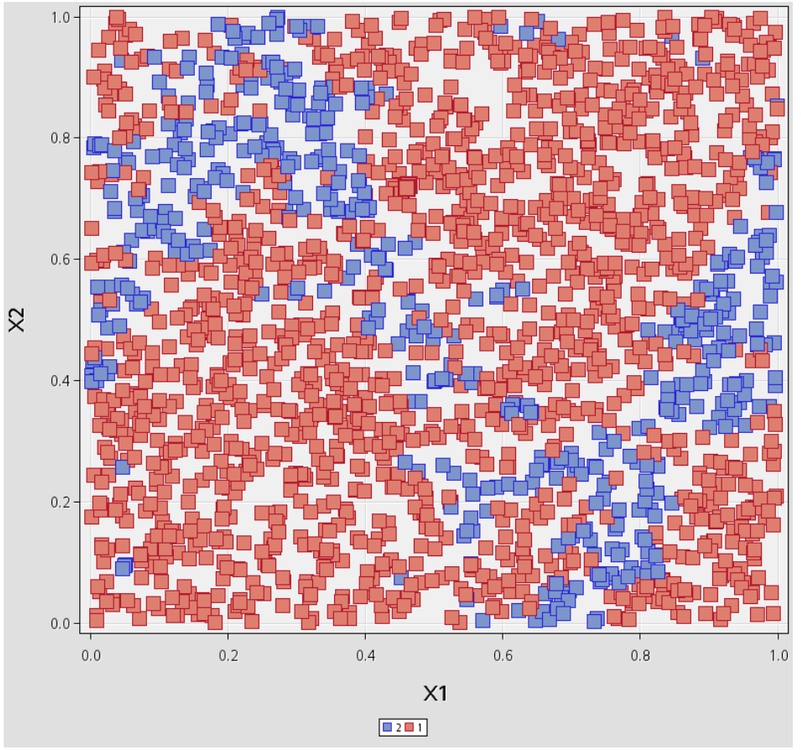




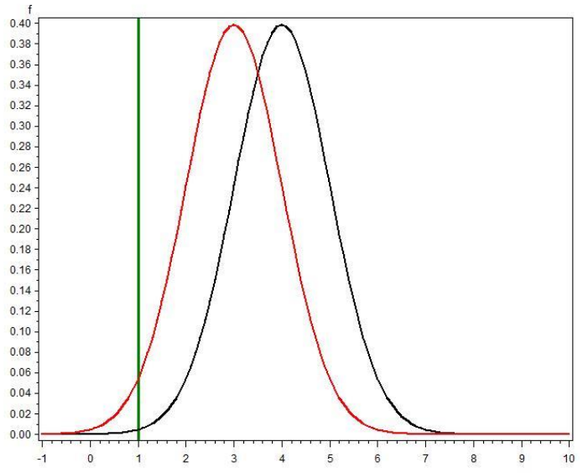

Note: Enterprise Miner Exported Data Explorer Chooses colors

→



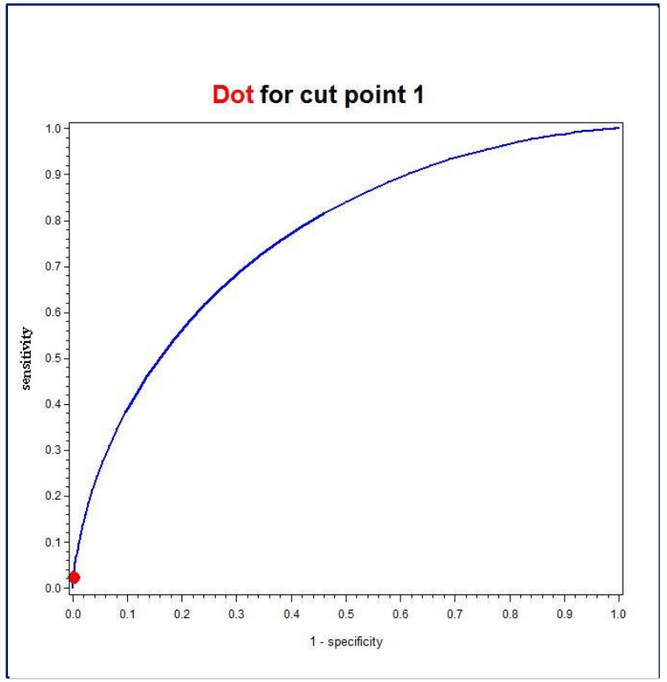
# Receiver Operating Characteristic Curve

Cut point 1




Logits of 1s red  Logits of 0s black

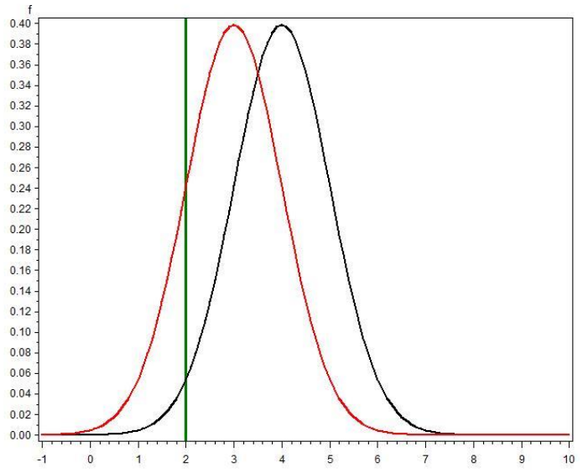
(or any model predictor from most to least likely to respond)




Select most likely $p_1$% according to model.
Call these 1, the rest 0. (call almost everything 0)
Y=proportion of all 1's correctly identified.(Y near 0)
X=proportion of all 0's incorrectly called 1's (X near 0)

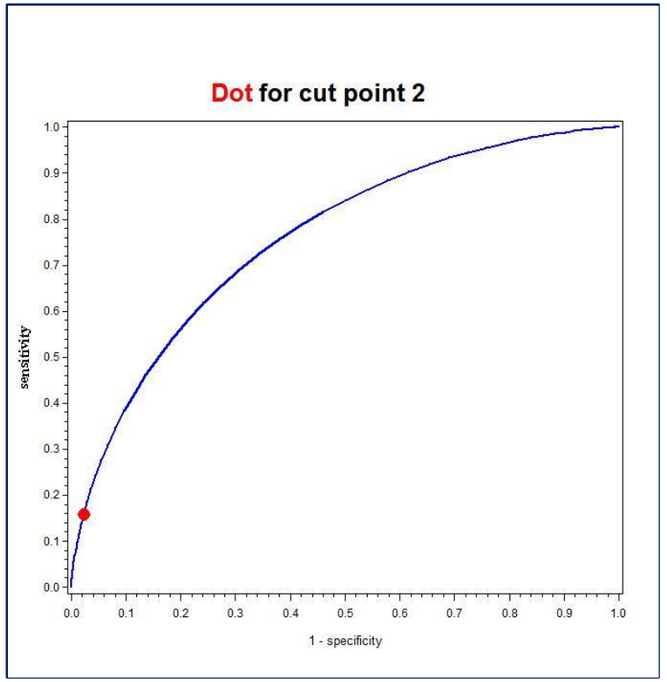
# Receiver Operating Characteristic Curve

Cut point 2




Logits of 1s red

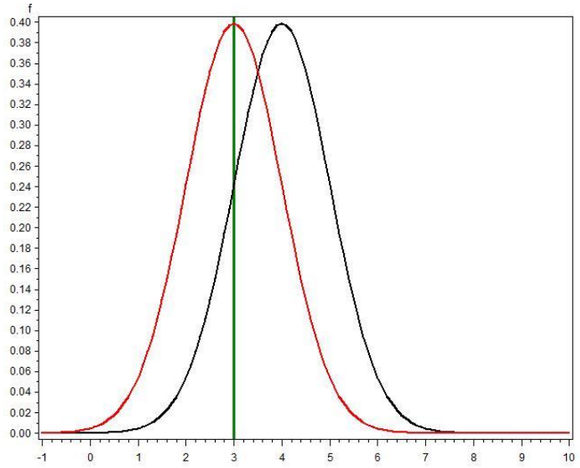
Logits of 0s black




Select most likely $p_2$% according to model.
Call these 1, the rest 0.
Y=proportion of all 1's correctly identified.
X=proportion of all 0's incorrectly called 1's

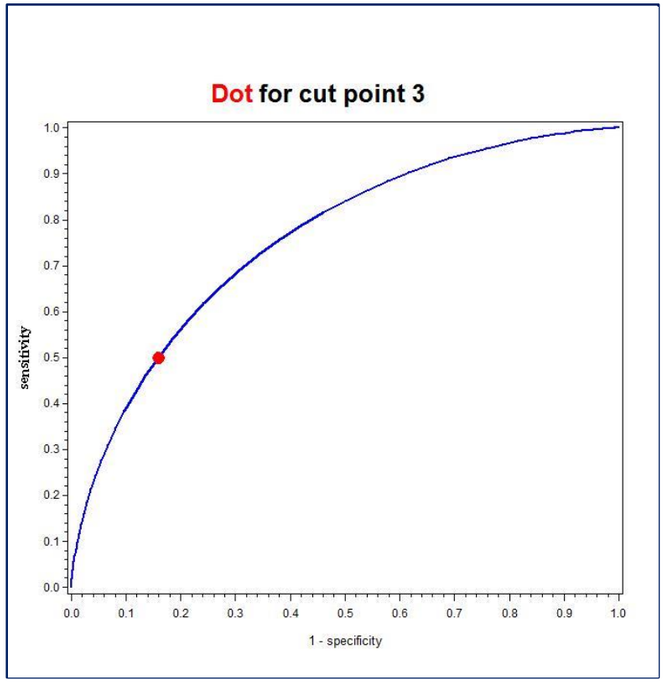
# Receiver Operating Characteristic Curve

Cut point 3




Logits of 1s red

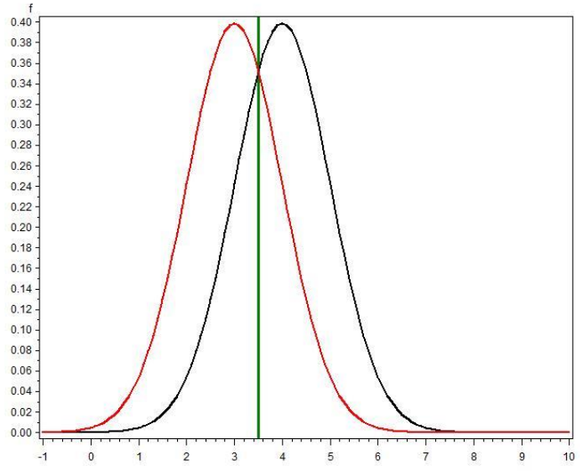
Logits of 0s black




Select most likely $p_3$% according to model.
Call these 1, the rest 0.
Y=proportion of all 1's correctly identified.
X=proportion of all 0's incorrectly called 1's

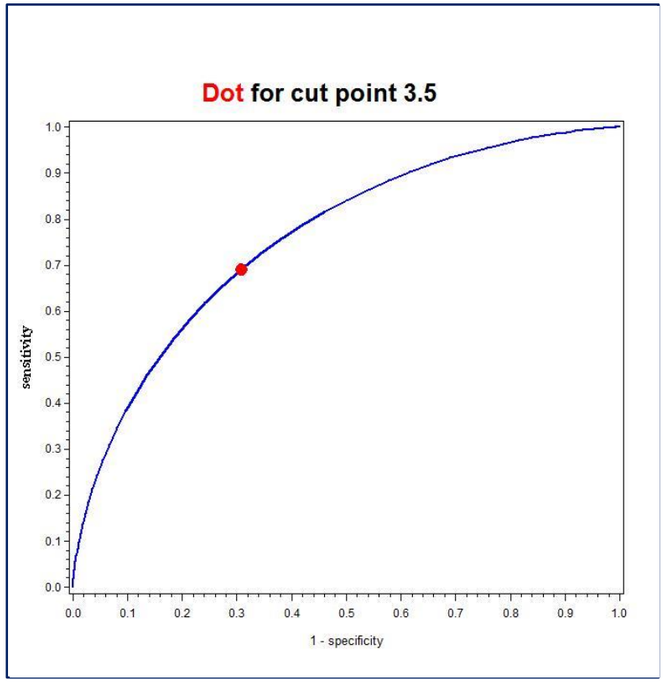
# Receiver Operating Characteristic Curve

Cut point 3.5




Logits of 1s red

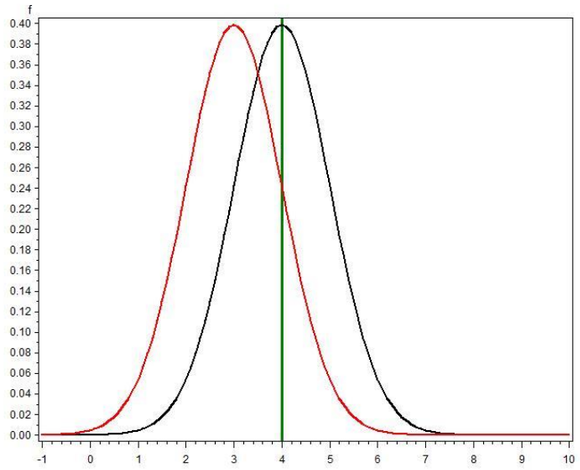
Logits of 0s black




Select most likely 50% according to model.
Call these 1, the rest 0.
Y=proportion of all 1's correctly identified.
X=proportion of all 0's incorrectly called 1's

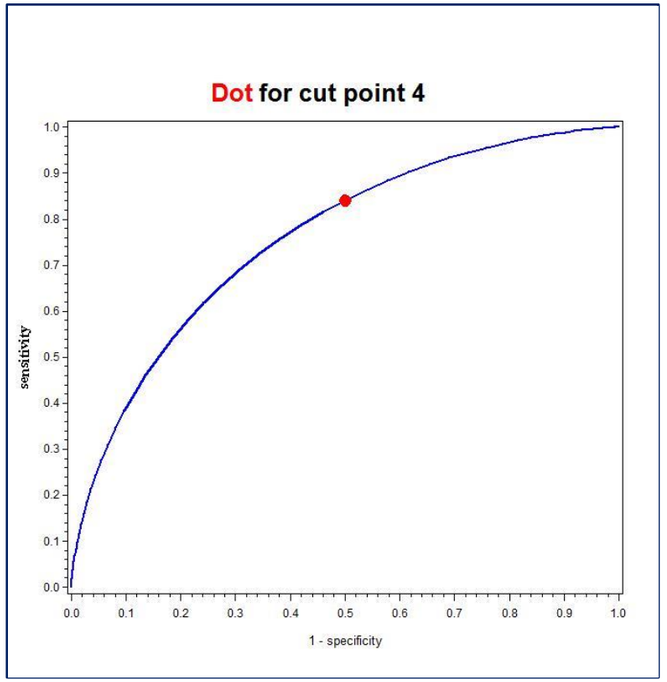
# Receiver Operating Characteristic Curve

Cut point 4




Logits of 1s red

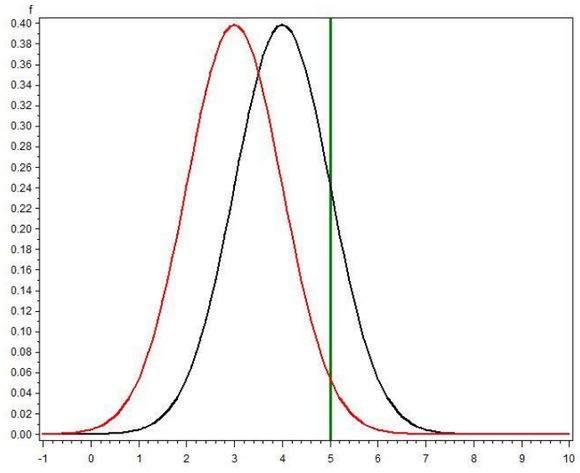
Logits of 0s black




Select most likely $p_5$% according to model.
Call these 1, the rest 0.
Y=proportion of all 1's correctly identified.
X=proportion of all 0's incorrectly called 1's

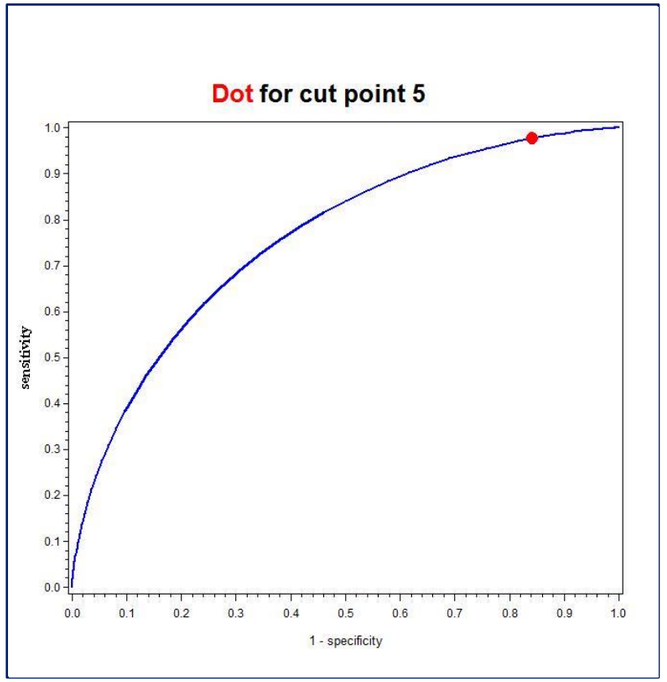
# Receiver Operating Characteristic Curve

Cut point 5




Logits of 1s red

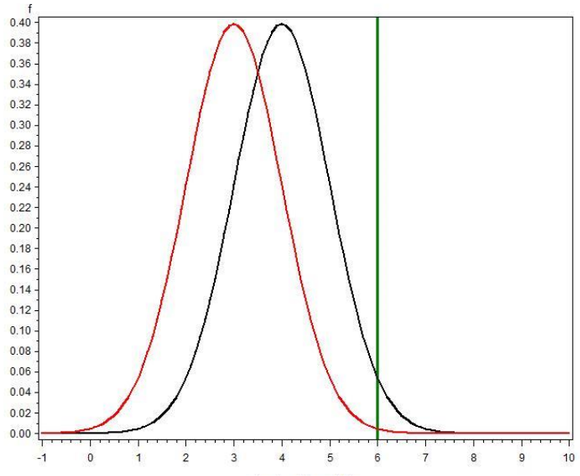
Logits of 0s black




Select most likely $p_6$% according to model.
Call these 1, the rest 0.
Y=proportion of all 1's correctly identified.
X=proportion of all 0's incorrectly called 1's

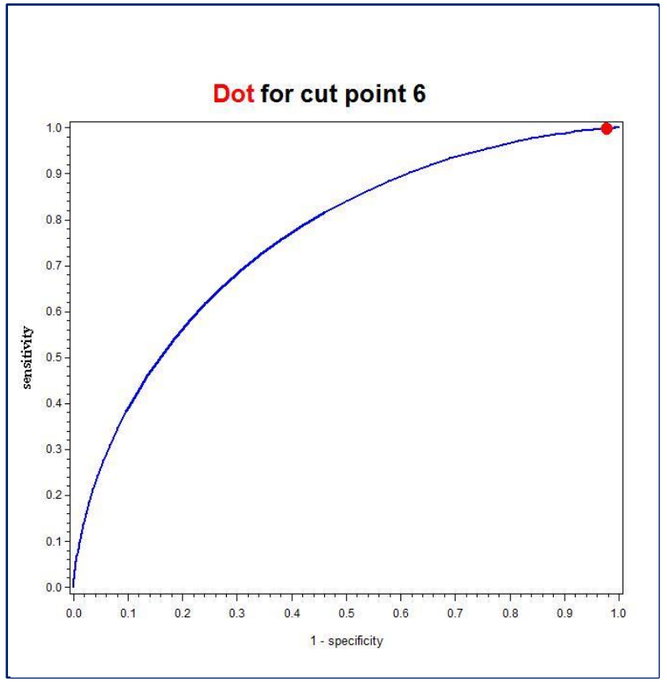
# Receiver Operating Characteristic Curve

Cut point 6




Logits of 1s red

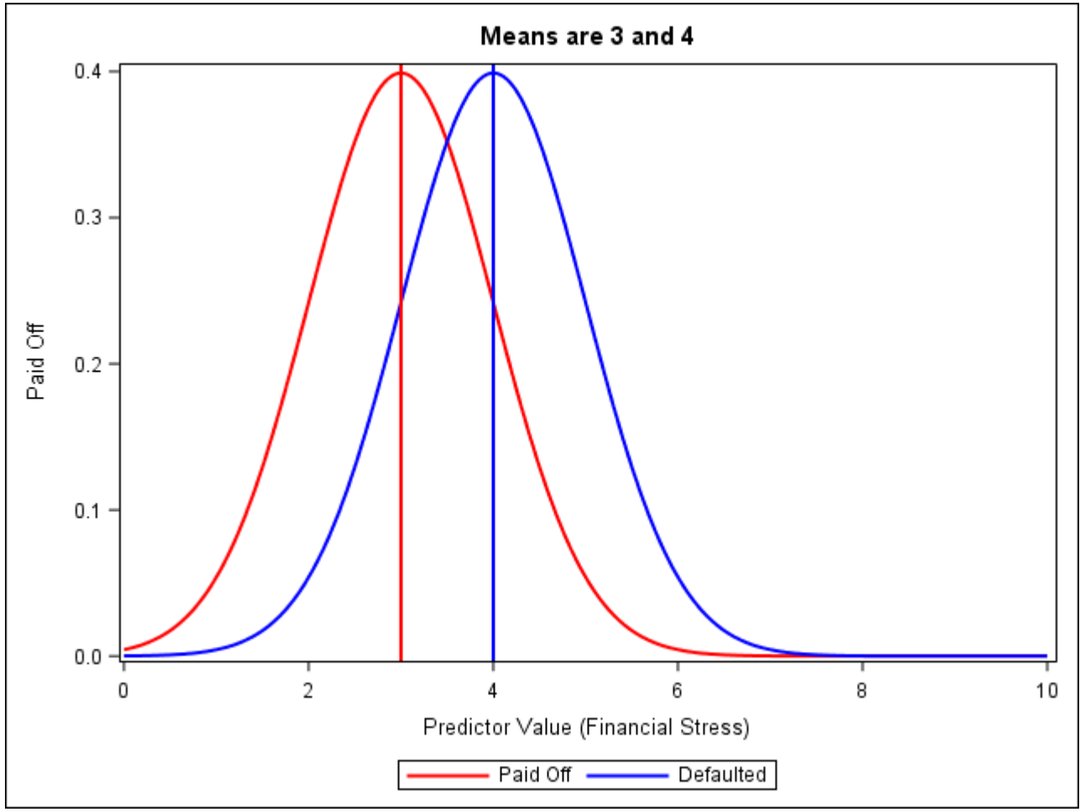
Logits of 0s black



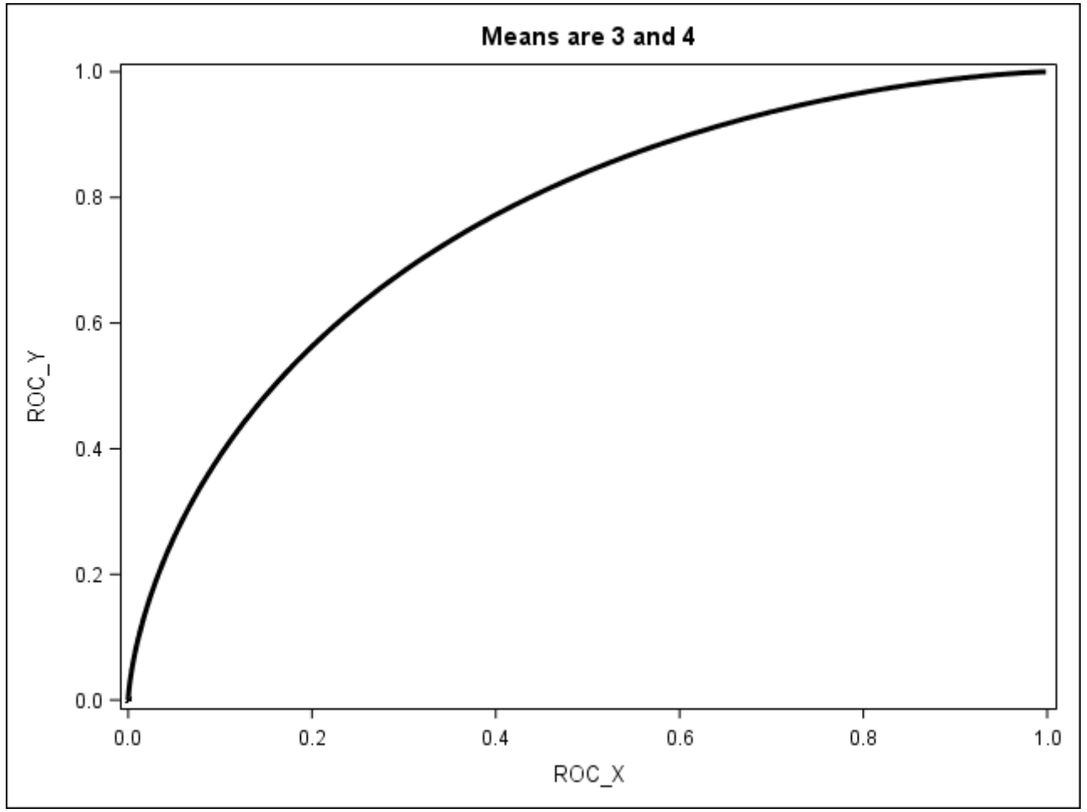

Select most likely $p_7$% according to model.
Call these 1, the rest 0.(call almost everything 1)
Y=proportion of all 1's correctly identified. (Y near 1)
X=proportion of all 0's incorrectly called 1's (X near 1)

Good Loan Bad Loan versus Financial Stress 1

Means: 3 and 4

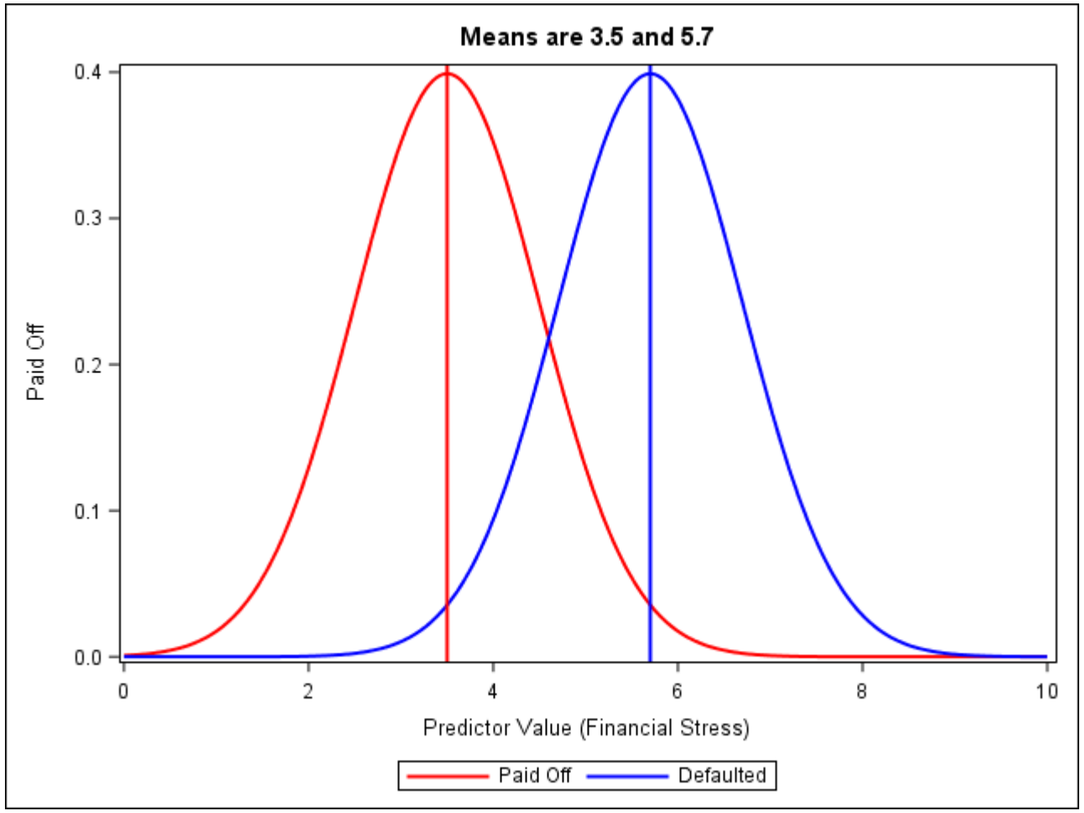



ROC Curve

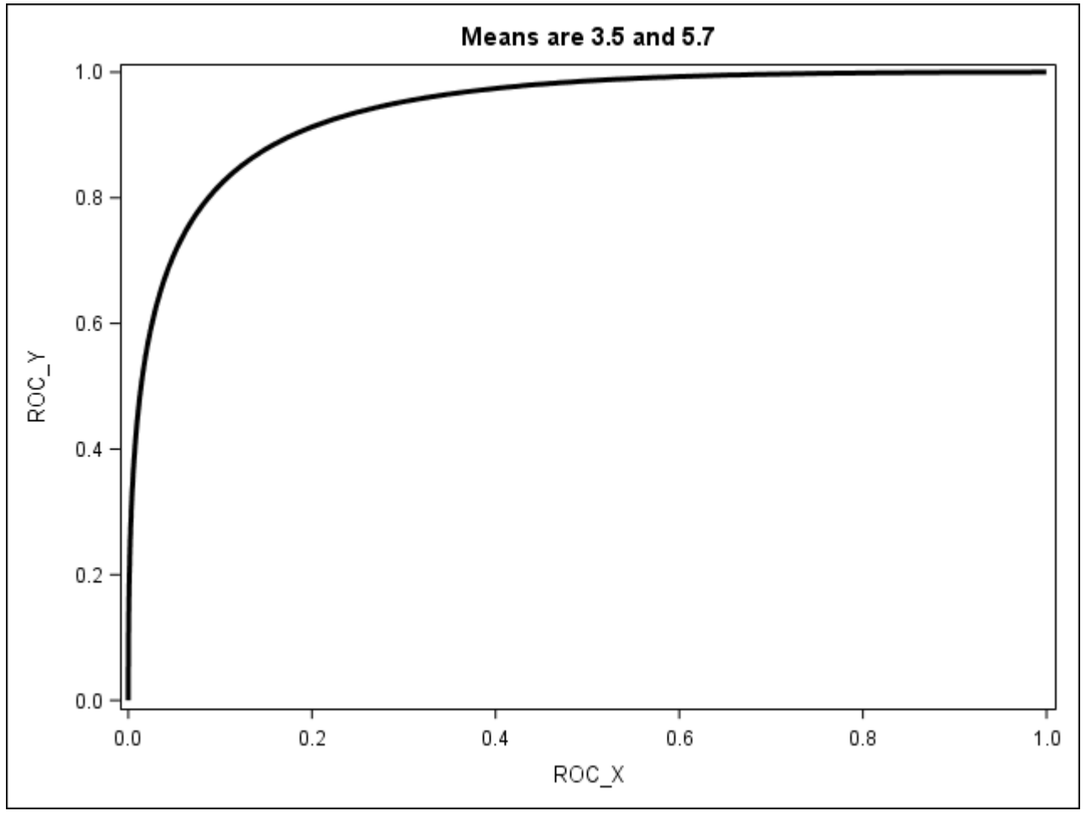


Good Loan Bad Loan versus Financial Stress 1

Means: 3.5 and 5.7

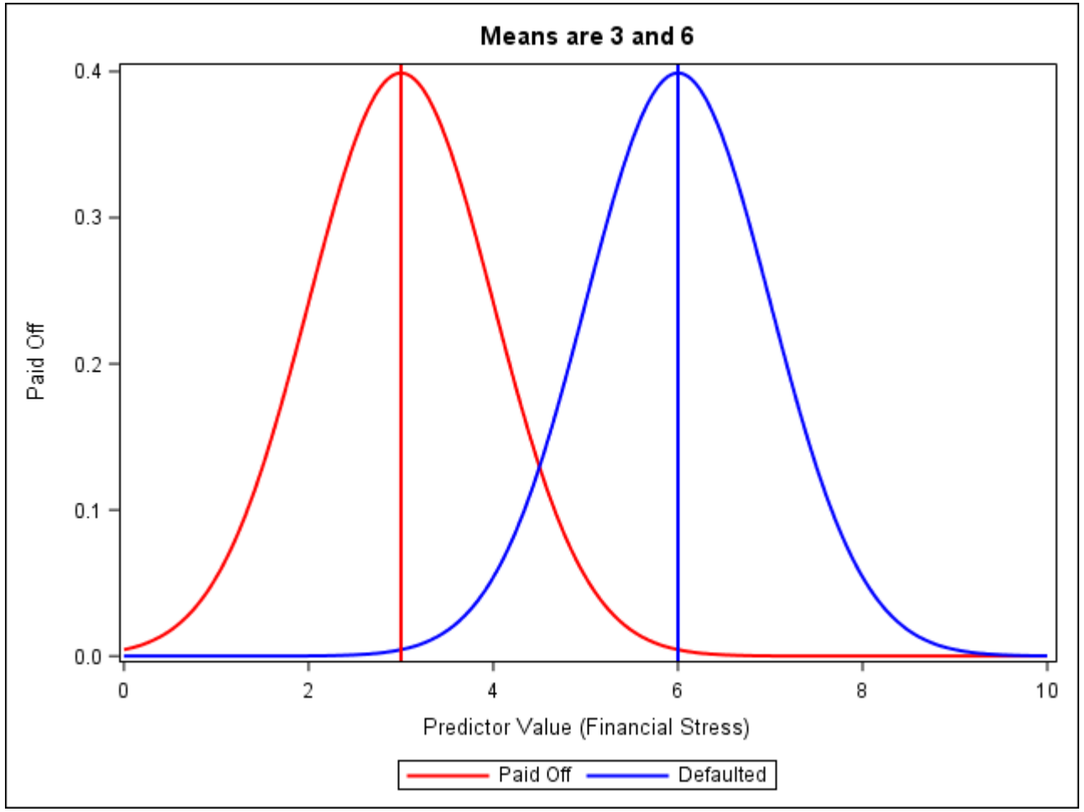



ROC Curve



Good Loan Bad Loan versus Financial Stress 2

Means: 3 and 6




ROC Curve