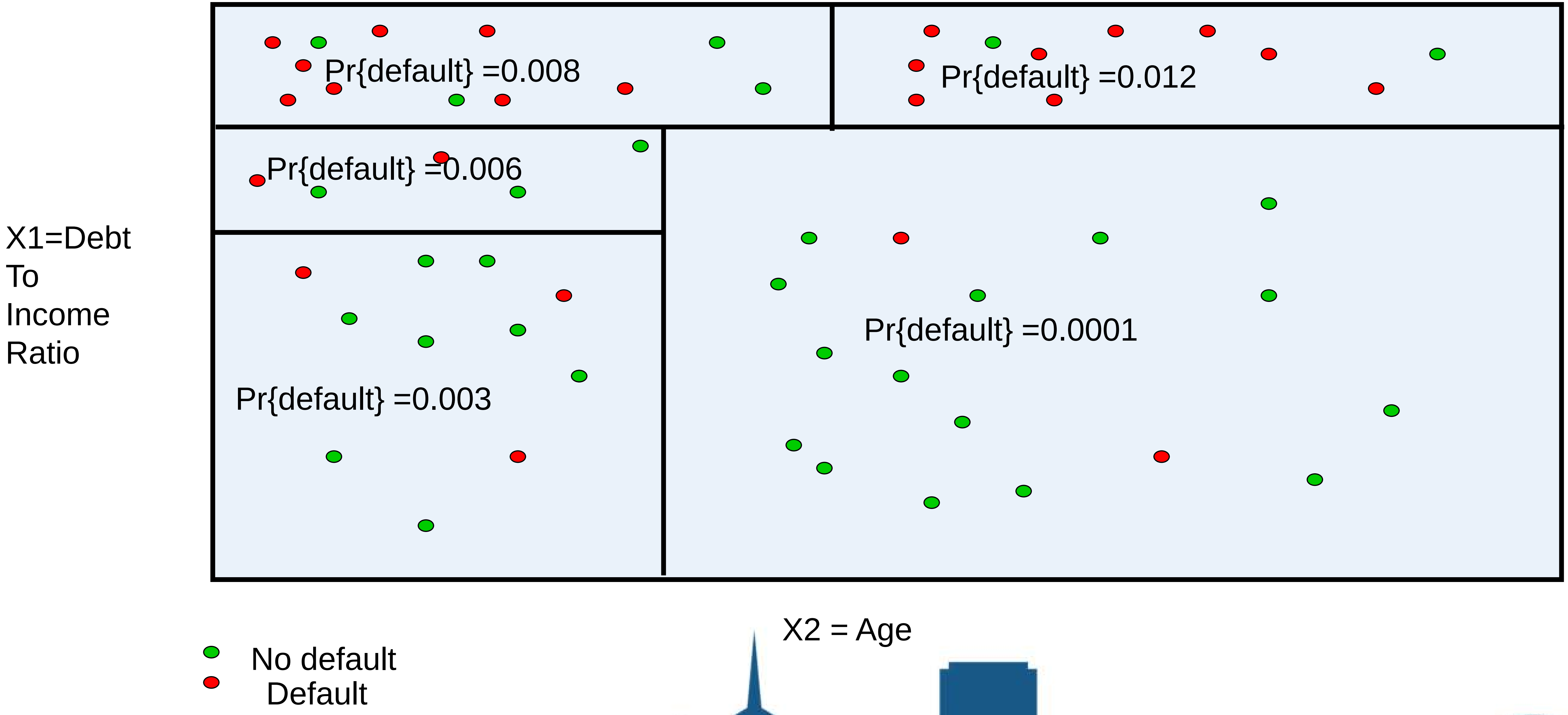
Finding the Gold in Your Data

- An introduction to Data Mining
- Originally presented @ SAS Global Forum

Decision Trees

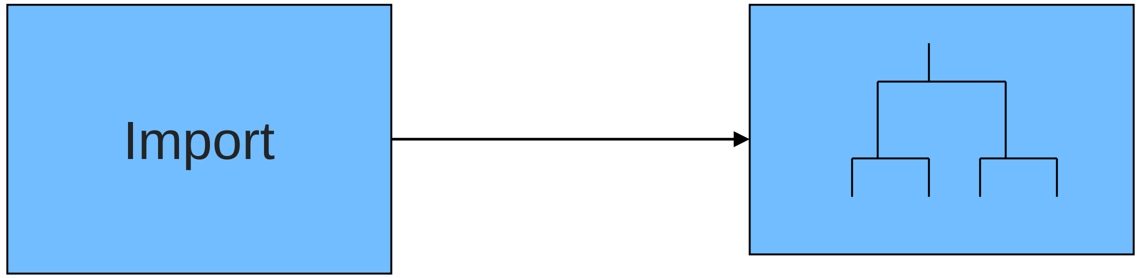
- A “divisive” method (splits)
- Start with “root node” – all in one group
- Get splitting rules
- Response often binary
- Result is a “tree”
- Example: Loan Defaults
- Example: Framingham Heart Study
- Example: Automobile Accidents

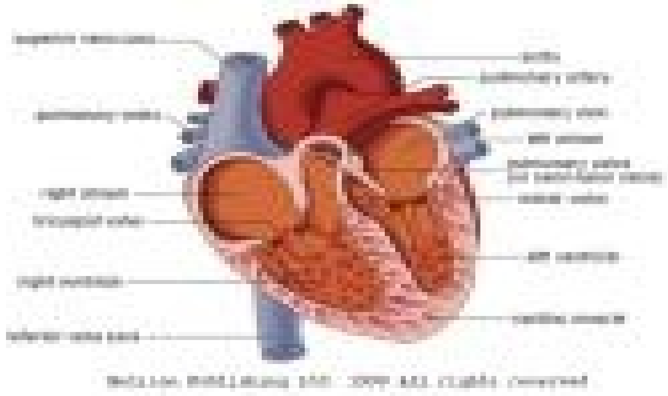
Recursive Splitting



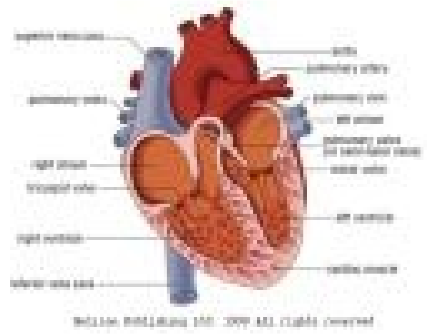
Some Actual Data

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

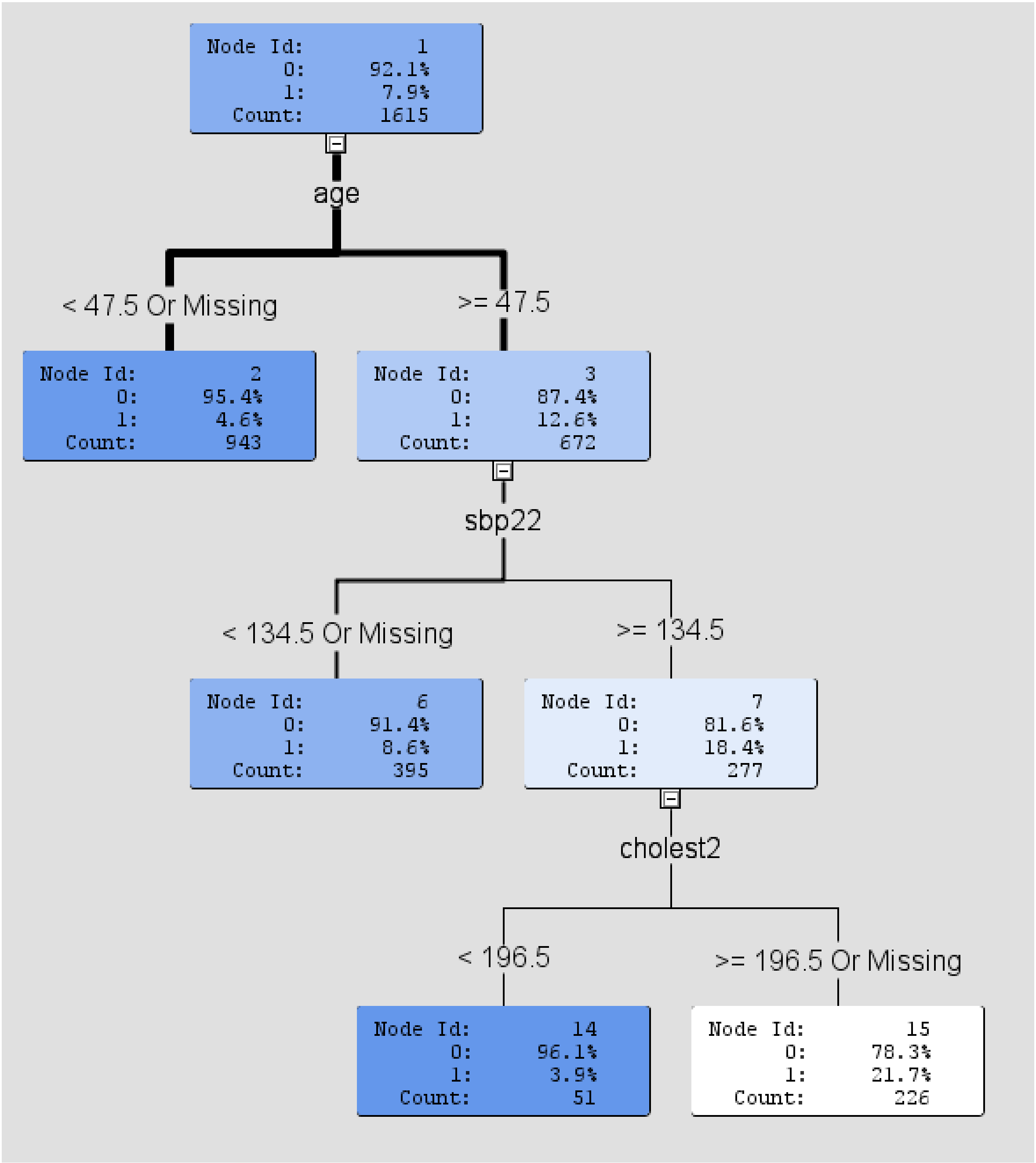




SAS GLOBAL FORUM



Example of a “tree” →



SAS GLOBAL FORUM

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)

SAS GLOBAL FORUM

How to make splits?

- Contingency tables

Heart Disease

		No	Yes	
180 ? 240?	Low BP	95 75	5 25	100
	High BP	55 75	45 25	100
		150	50	

DEPENDENT (effect)

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

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

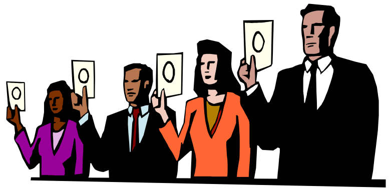
Compare to tables –
Significant!
(Why “Significant” ???)

SAS GLOBAL FORUM

H_0: 

H_1: 

H_0: Innocence
H_1: Guilt

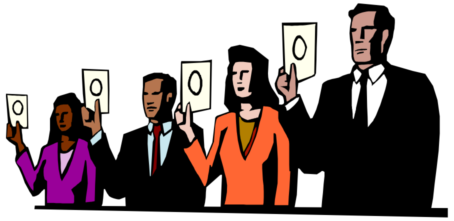


Beyond reasonable doubt
$P<0.05$

95 75	5 25
55 75	45 25

H_0: No association
H_1: BP and heart disease *are* associated

$P=0.00000000064$



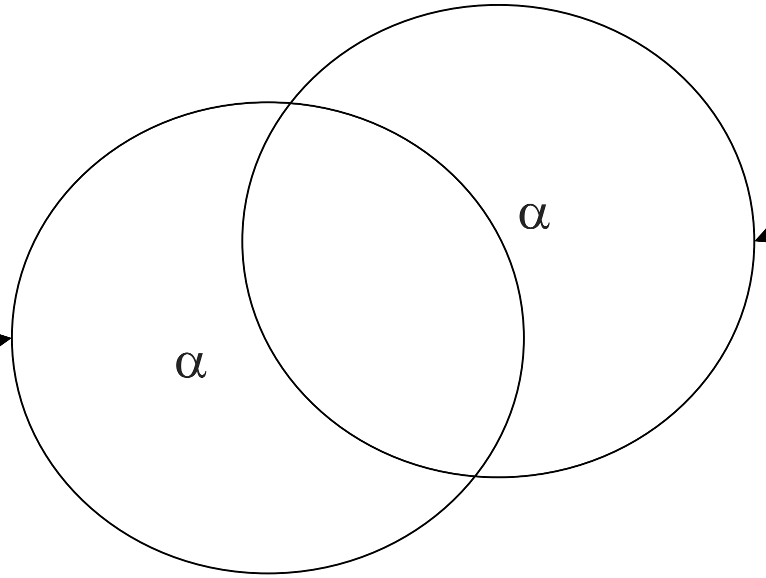
Framingham Conclusion: Sufficient evidence *against* the (null) hypothesis of no relationship.

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 χ^2
- Split point data-derived!
- What does "<u>significance</u>" mean now?

Multiple testing

α =
Pr{ falsely reject hypothesis 2}



α =
Pr{ falsely reject hypothesis 1}

Pr{ falsely reject one or the other} < 2α
Desired: 0.05 probability or less
Solution: Compare 2(p-value) to 0.05

Other Sp lit Criteria

- Gini Diversity Index
  - (1) { A A A A B A B B C B}
  - Pick 2, Pr{different} = 1-Pr{AA}-Pr{BB}-Pr{CC}
    - » 1-[10+6+0]/45=29/45=0.64
  - (2) { A A B C B A A B C C }
    - » 1-[6+3+3]/45 = 33/45 = 0.73 → (2) IS MORE DIVERSE, LESS PURE
- Shannon Entropy
  - Larger → more diverse (less pure)
  - $-\Sigma_i\ p_i \log_2(p_i)$

{0.5, 0.4, 0.1} → 1.36
{0.4, 0.2, 0.3} → 1.51 (more diverse)

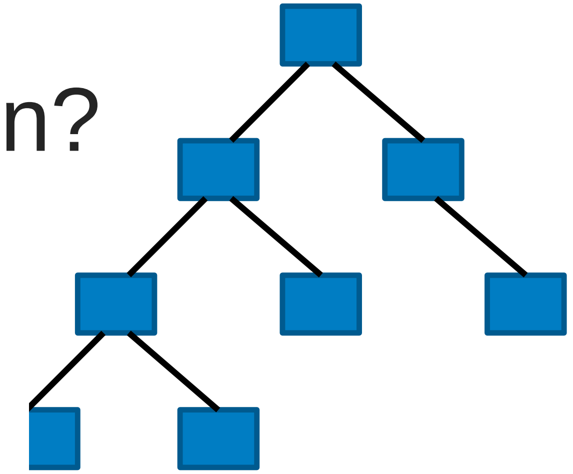
SAS GLOBAL FORUM

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.

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 if diversity in parent “node” > summed diversities in child nodes
- Prune to optimize
  - Estimates
  - Decisions
  - Ranking
- in validation data

SAS GLOBAL FORUM

Accounting for Costs

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

SAS GLOBAL FORUM

Including Probabilities

Leaf has Pr(**M**)=**.7**, Pr(**F**)=**.3**

You say:

True Gender	Sir	Ma'am
M	**0.7** (2)	**0.7** (-10)
F	**0.3** (-1)	**0.3** (5)
	+$1.10	-$5.50

Expected profit is
2(0.7)-1(0.3) = $1.10
if I say "sir"

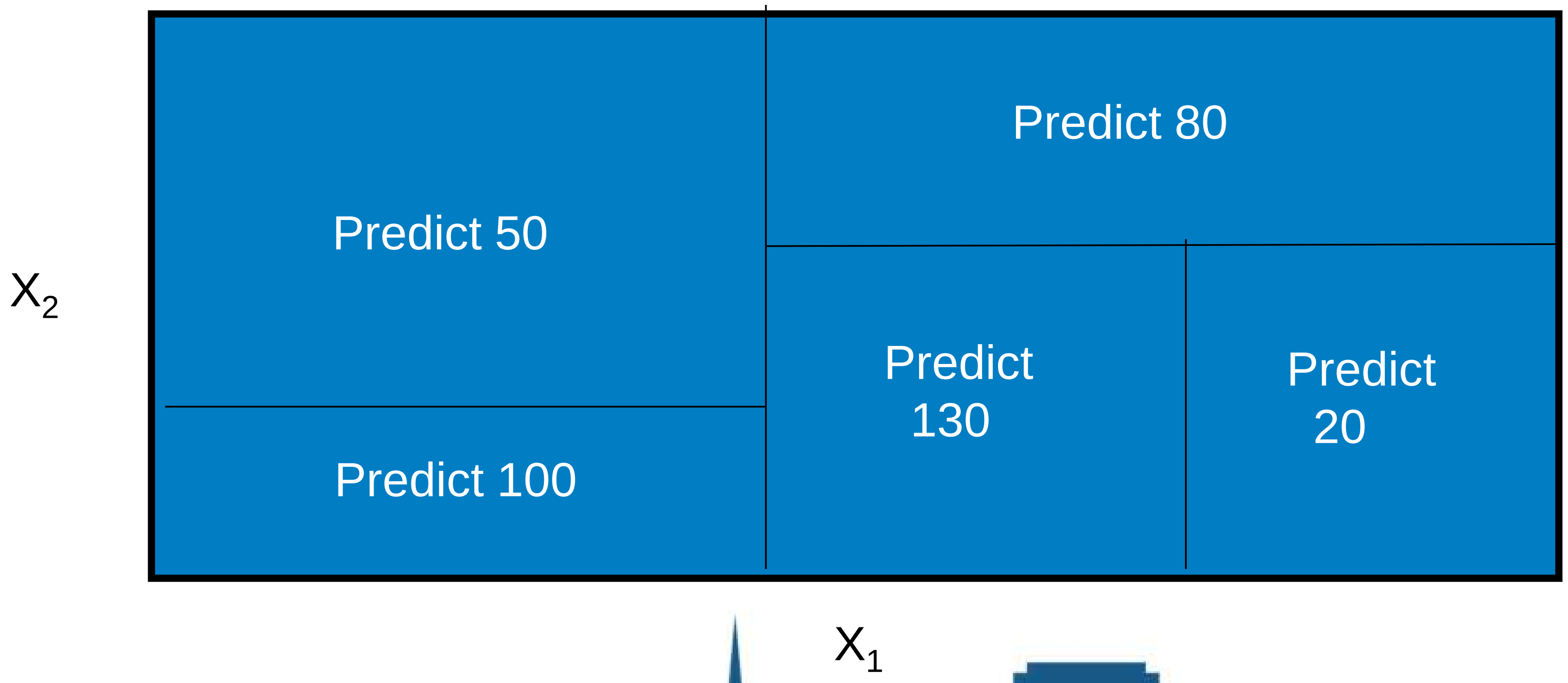
Expected profit is
-7+1.5 = -$5.50 (a loss)
if I say "Ma'am"

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

Prune (and split) to maximize profits.

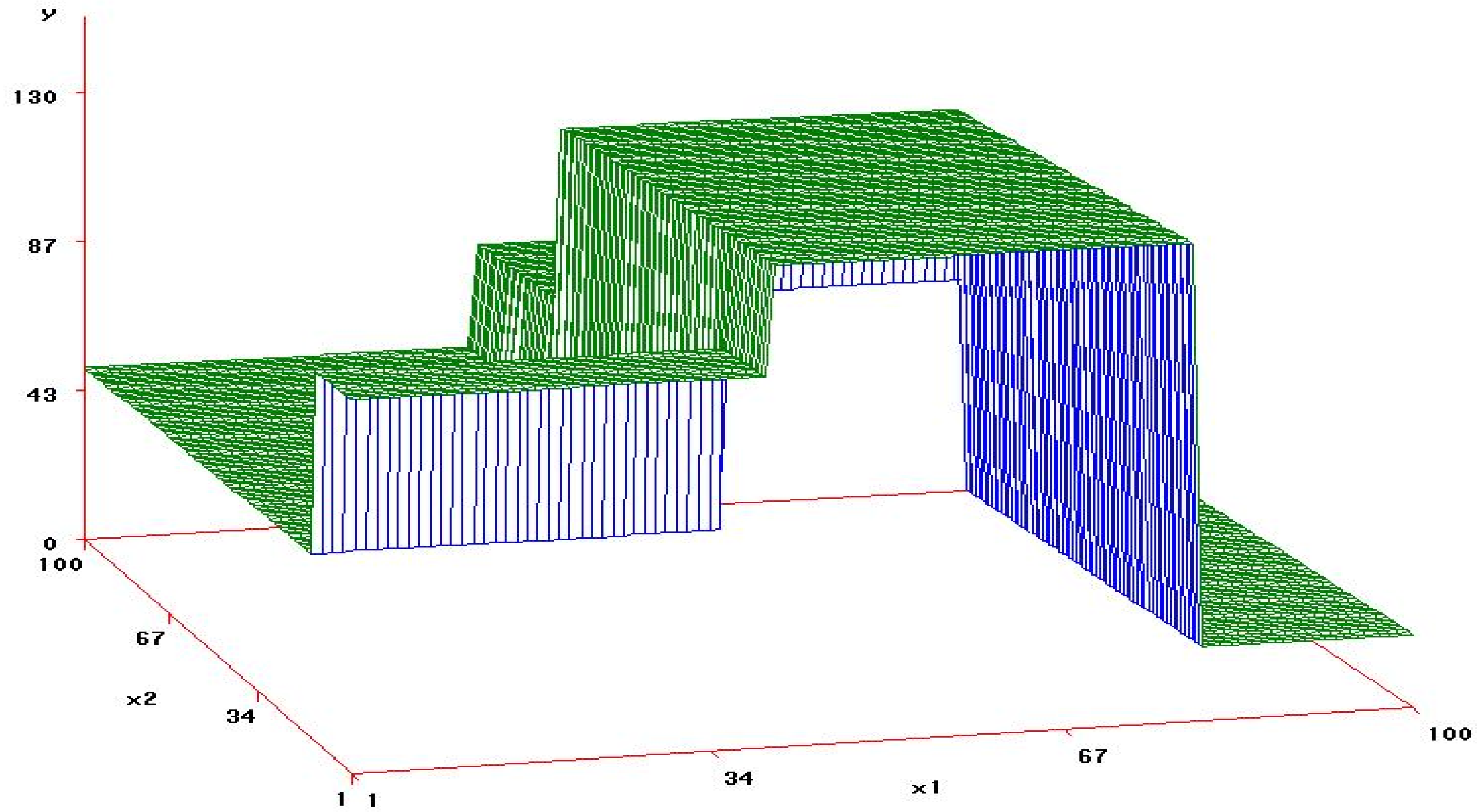
Regression Trees

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



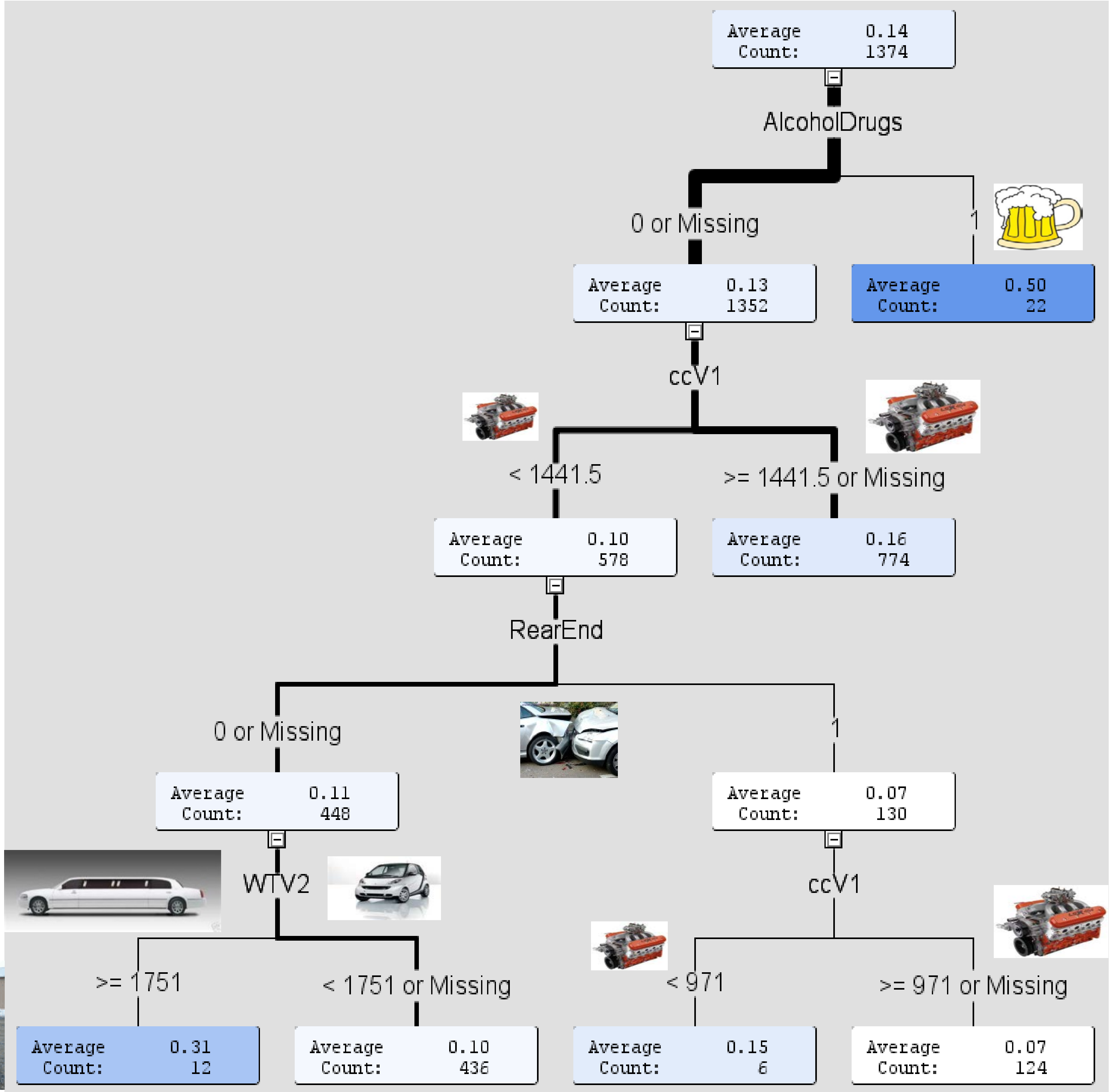
Regression Trees

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



Real data example: Traffic accidents in Portugal*

Y = injury induced “cost to society”



Help - I ran Into a “tree”

Help - I ran Into a “tree”

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

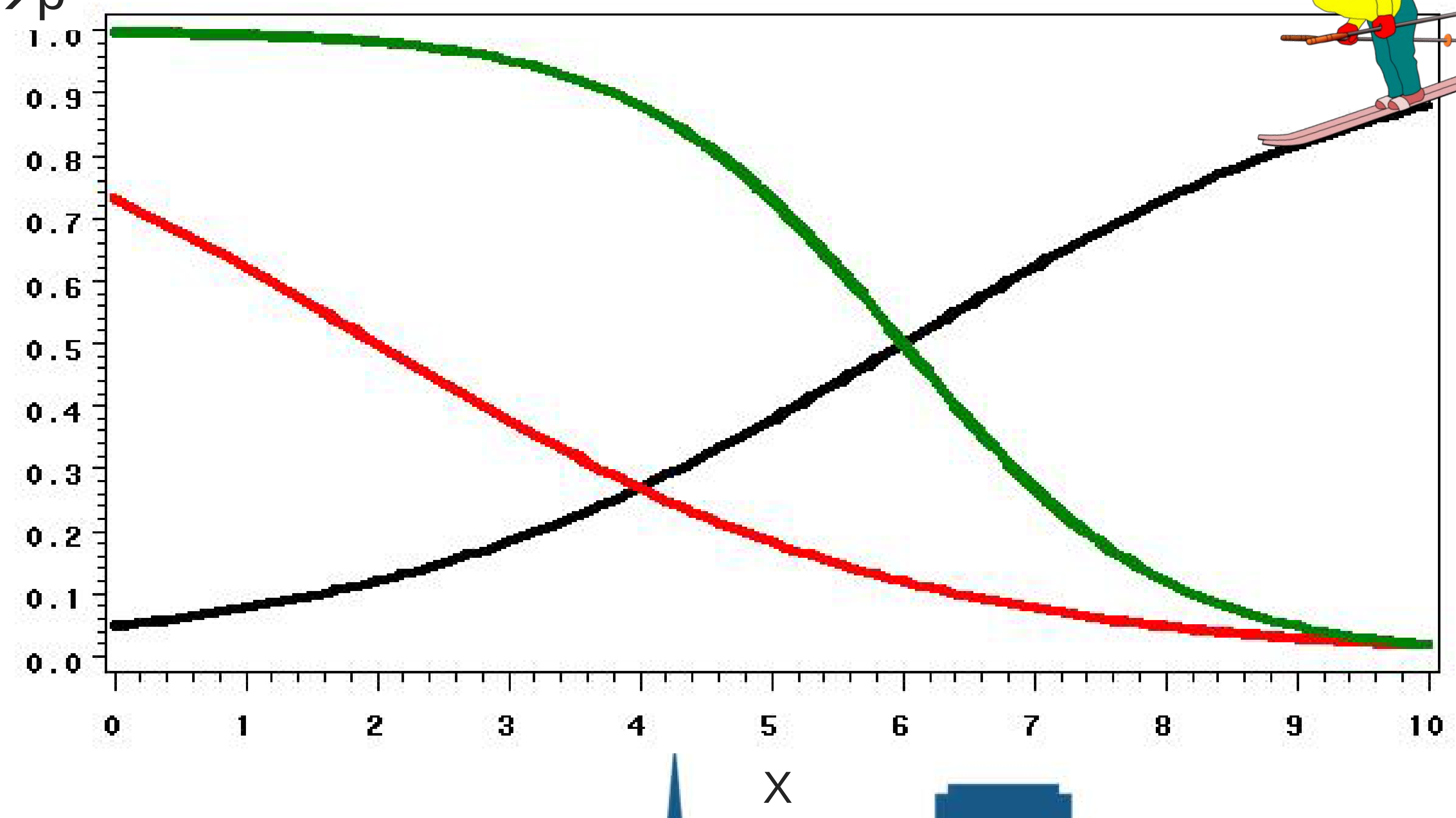
Logistic Regression

- 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

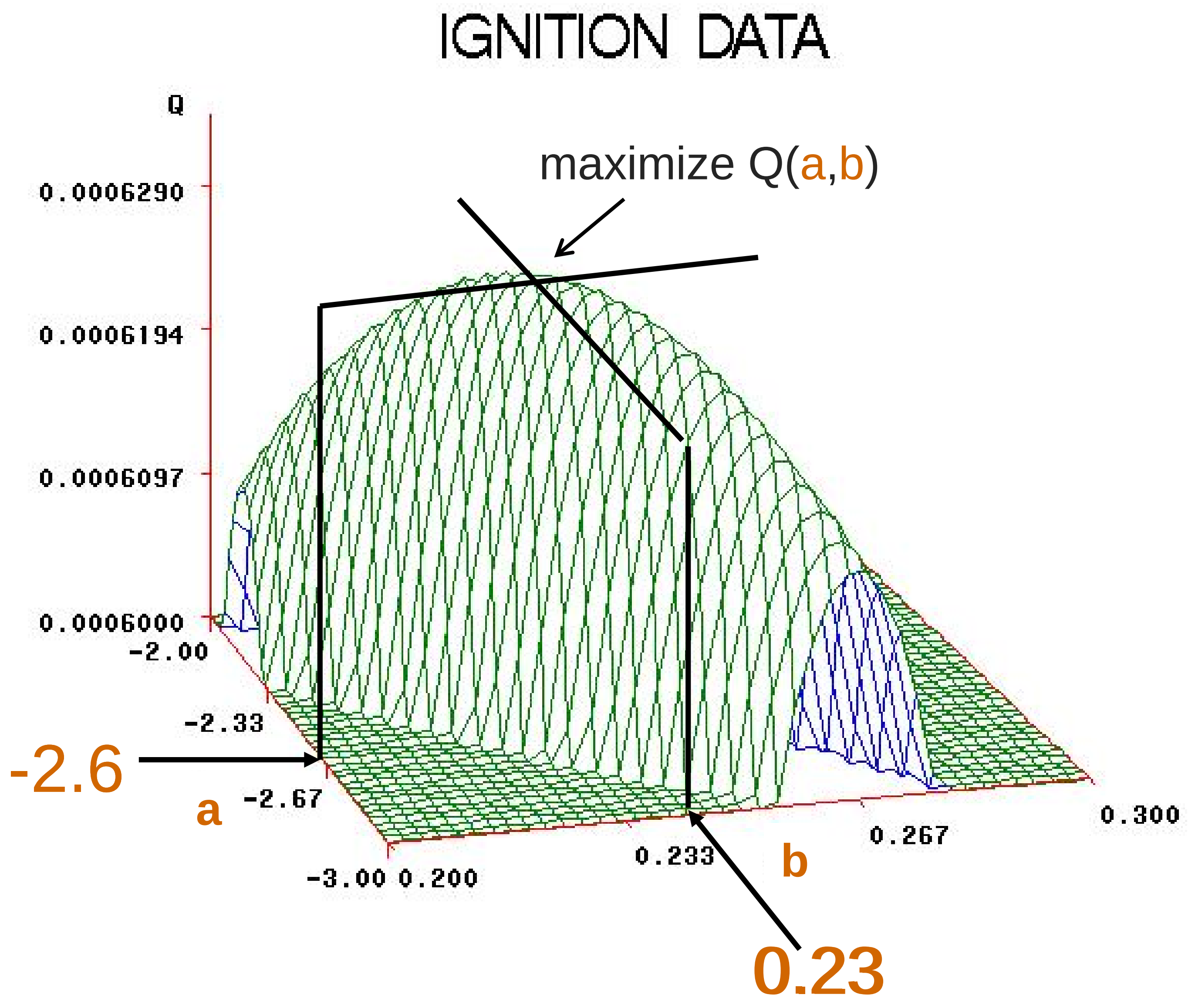


Example: Seat Fabric Ignition

- Flame exposure time = X
- Y=1 → ignited, Y=0→ did not ignite
  - Y=0, X= 3, 5, 9 10 , 13, 16
  - Y=1, X = 11, 12 14, 15, 17, 25, 30
- $Q=(1-p_1)(1-p_2)(1-p_3)(1-p_4)p_5p_6(1-p_7)p_8p_9(1-p_{10})p_{11}p_{12}p_{13}$
- p's all different $p_i = f(a+bX_i) = e^{a+bX_i}/(1+e^{a+bX_i})$
- Find a,b to maximize Q(a,b)

SAS GLOBAL FORUM

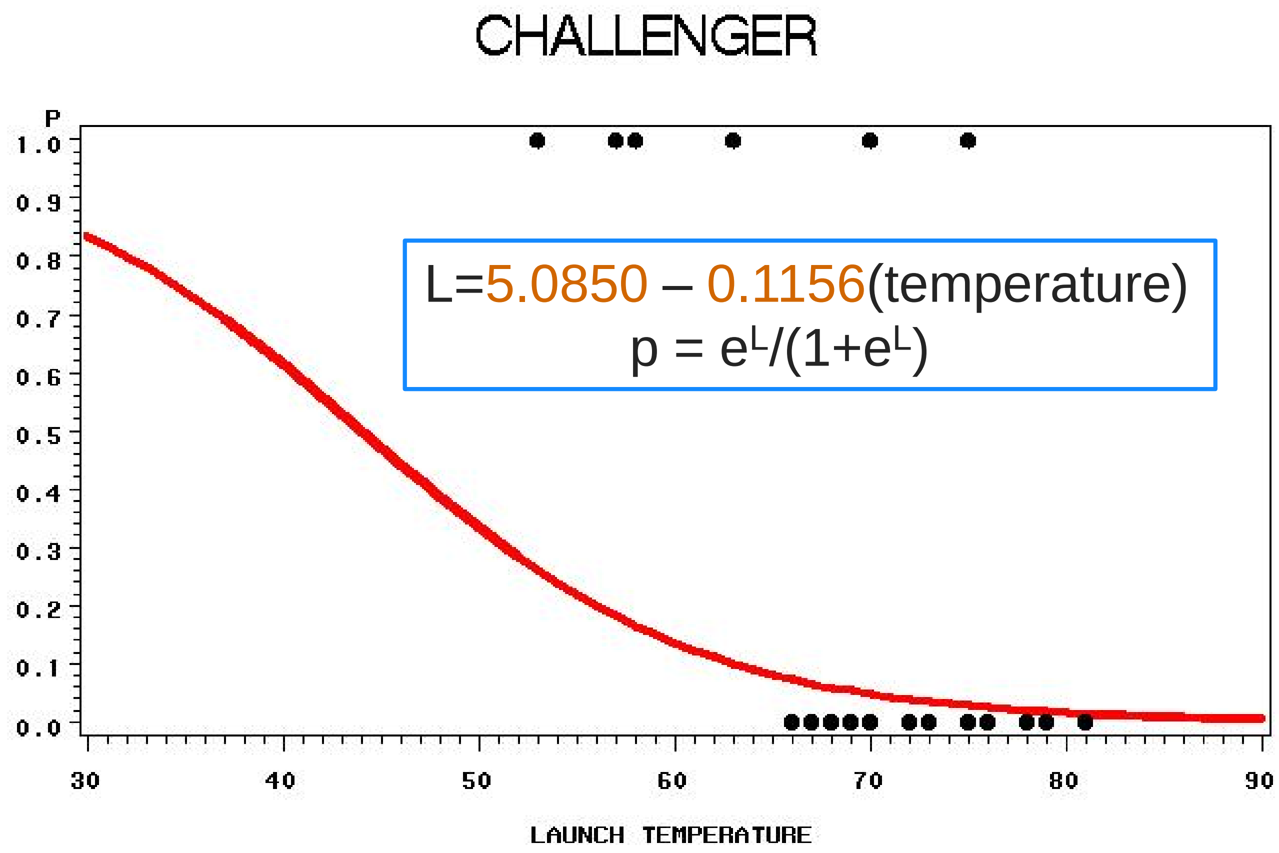
- Logistic idea:
- Given temperature X, compute $L(x)=a+bX$ then $p = e^L/(1+e^L)$
- $p(i) = e^{a+bX_i}/(1+e^{a+bX_i})$
- Write p(i) if response, 1-p(i) if not
- Multiply all n of these together, find a,b to maximize this “likelihood”
- Estimated L = ____ + _____ X



SAS GLOBAL FORUM

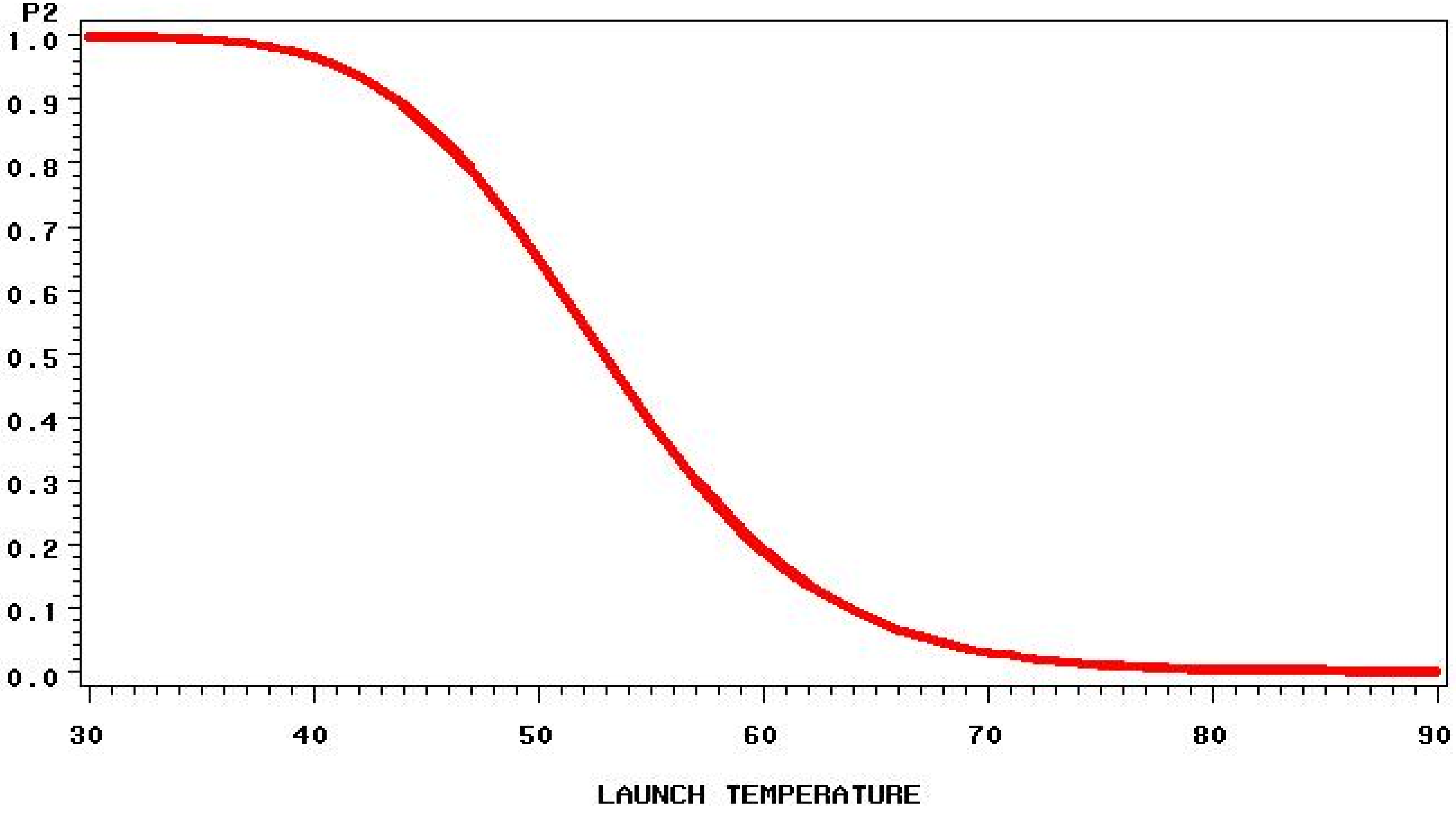
Example: Shuttle Missions

- 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)



SAS GLOBAL FORUM

Pr{2 OR MORE FAILURES}

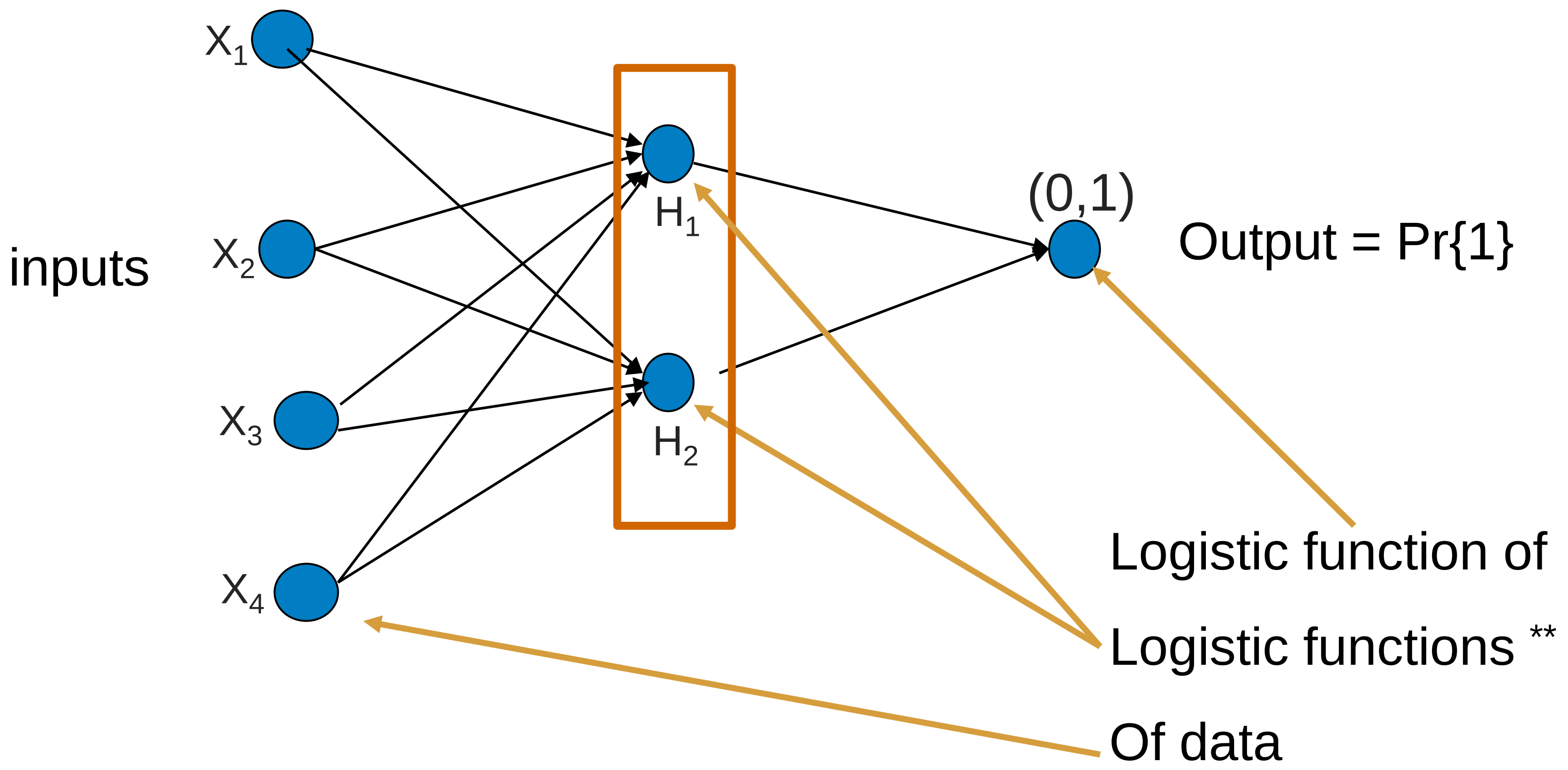


$$\Pr\{2\ \ or\ \ more\} = 1 - p_X{}^0 (1 - p_X)^6 - 6 p_X (1 - p_X)^5$$

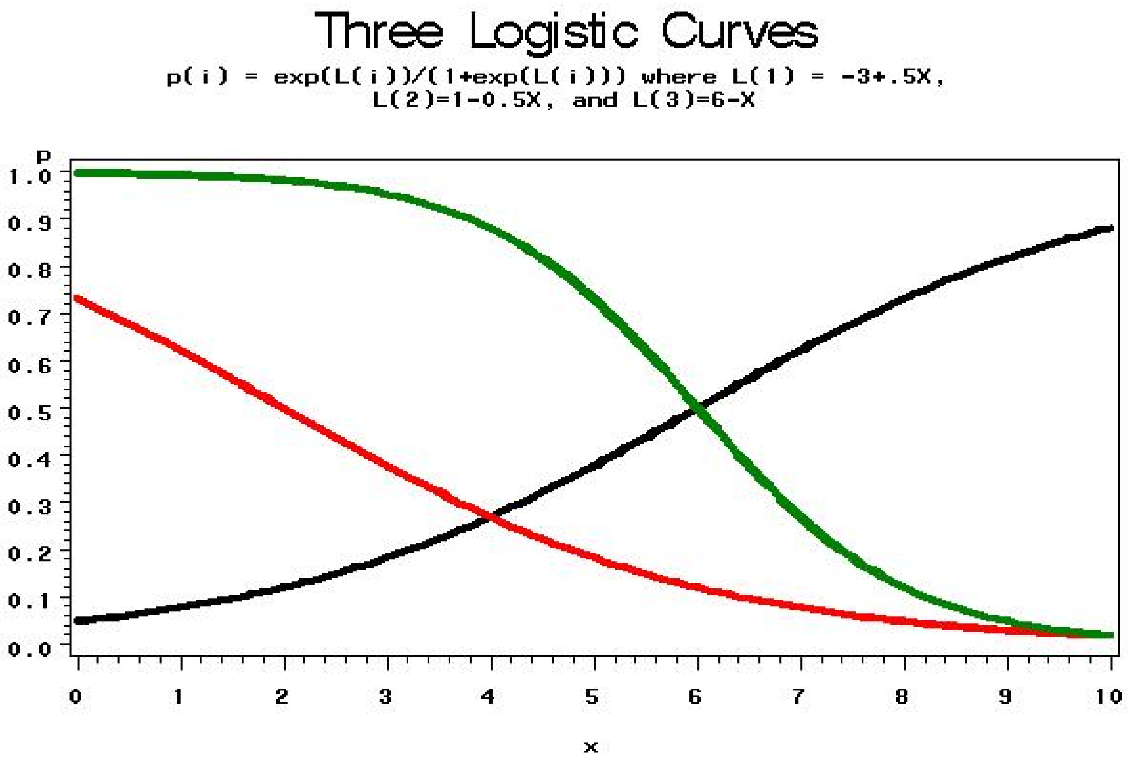
SAS GLOBAL FORUM

Neural Networks

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



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



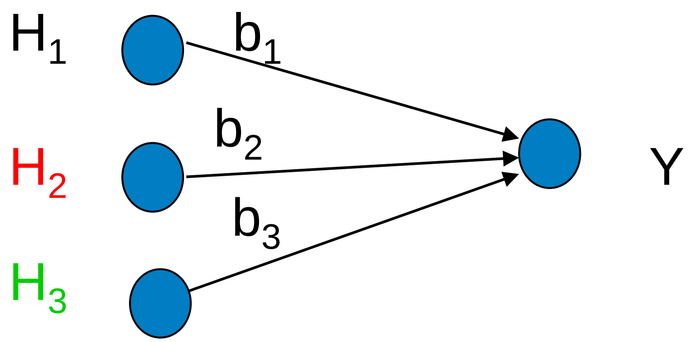
Example:

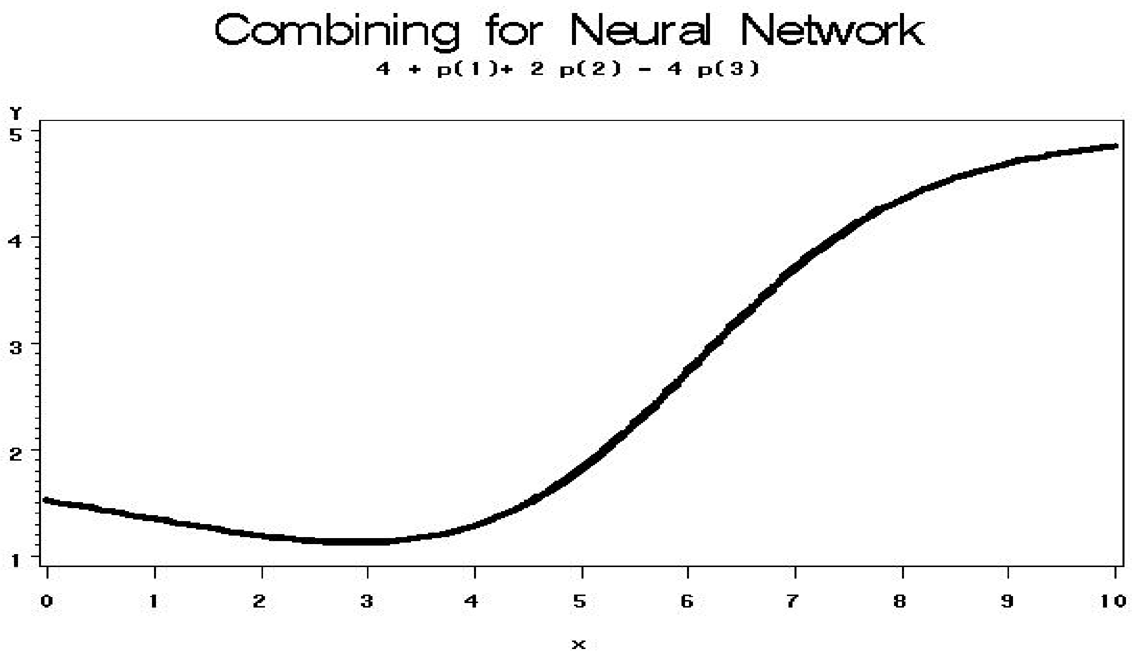
Y = a + b1 H1 + b2 H2 + b3 H3

Y = 4 + 1 H1 + 2 H2 - 4 H3

"bias"

"weights"

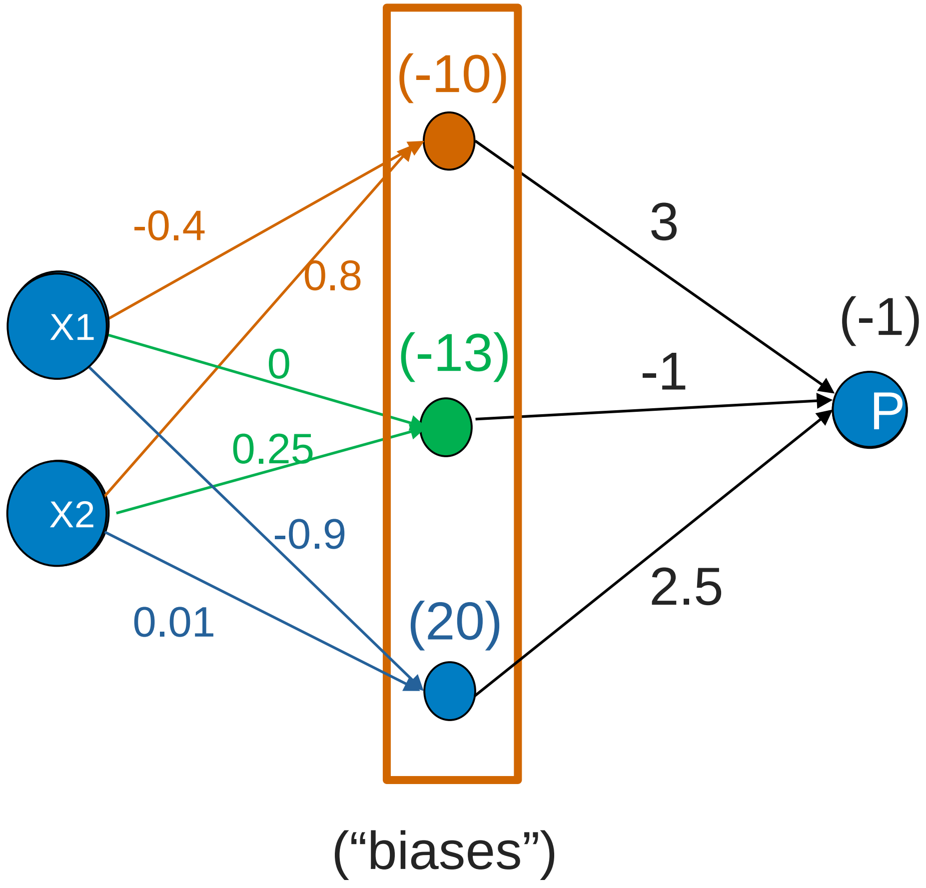


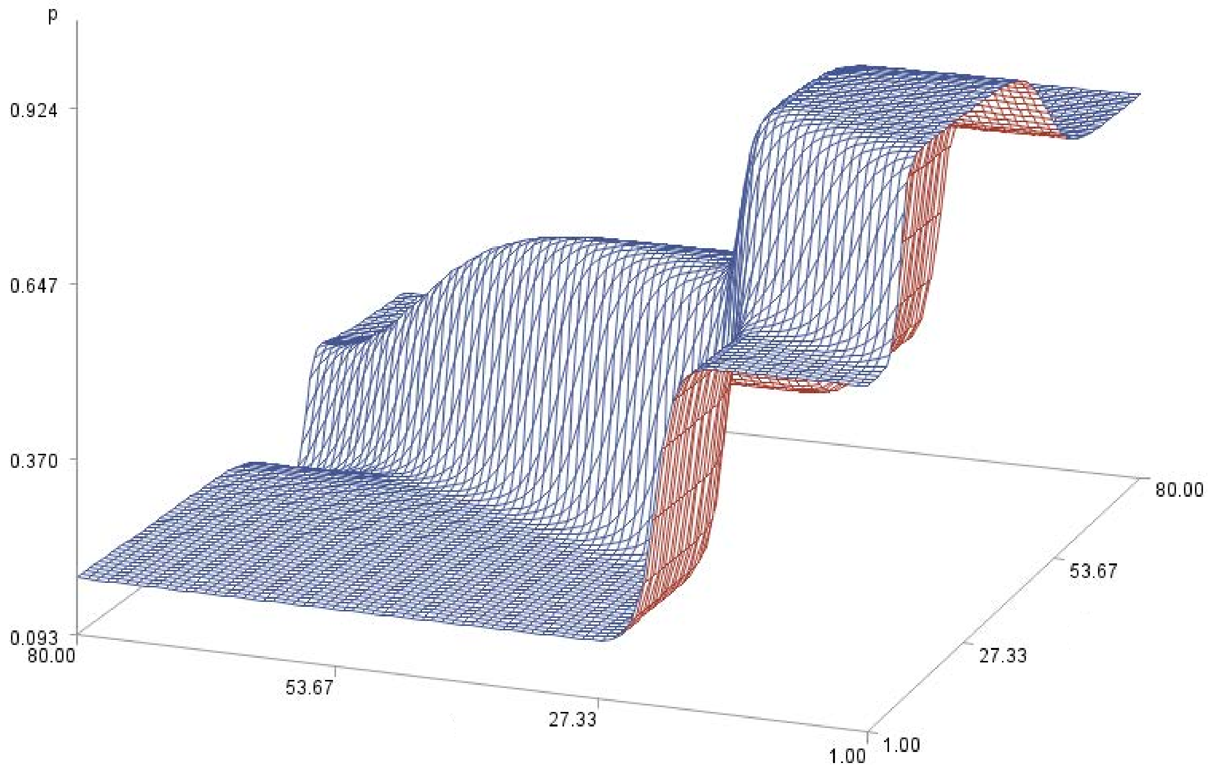


Arrows represent linear combinations of "basis functions," e.g. logistic curves (hyperbolic tangents)

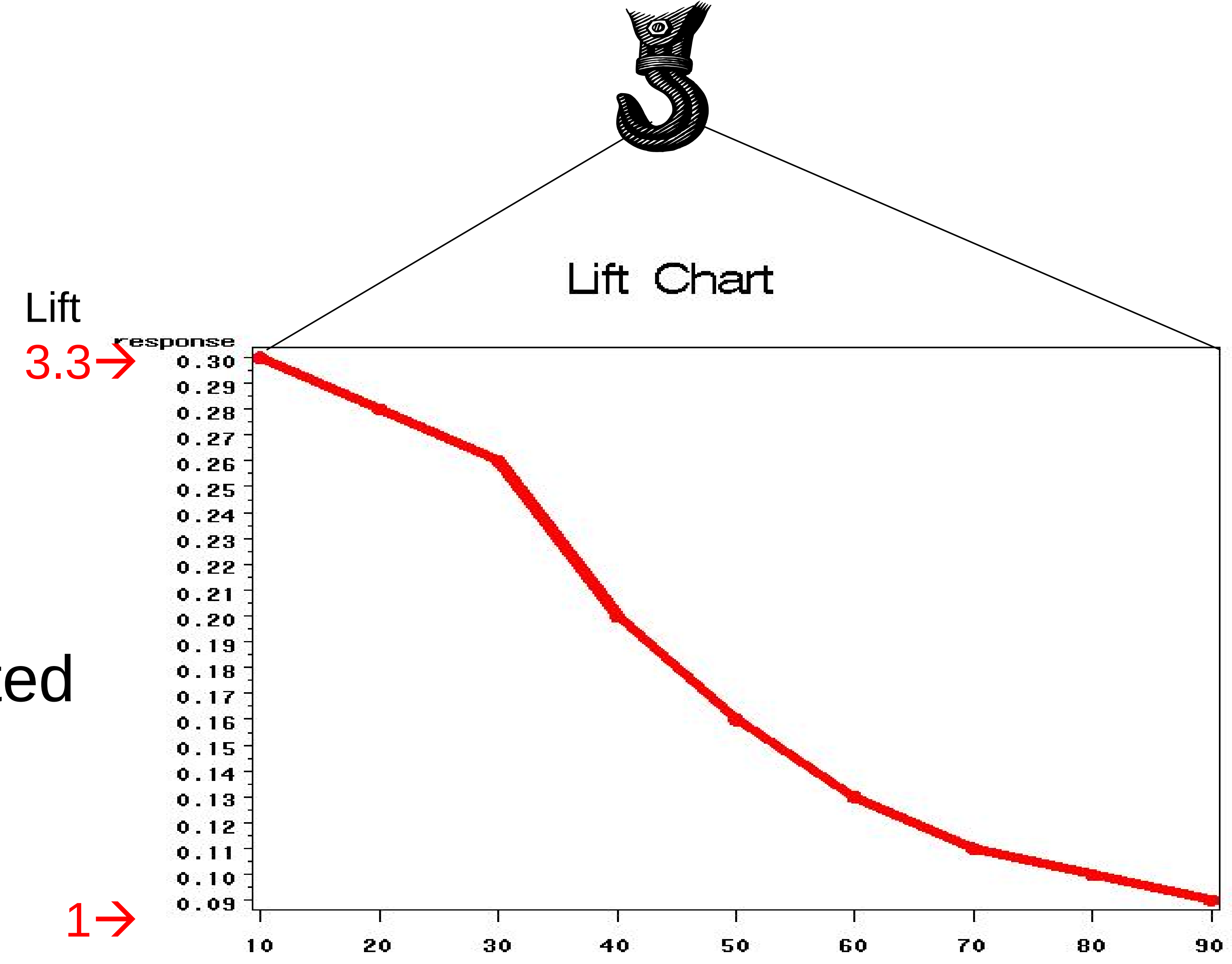
SAS GLOBAL FORUM

A Complex Neural Network Surface





* Cumulative Lift Chart
    - Go from leaf of most to least predicted response.
    - Lift is
      proportion responding in first p%
      overall population response rate



Predicted pct response

high ←-----------------------------------→ low

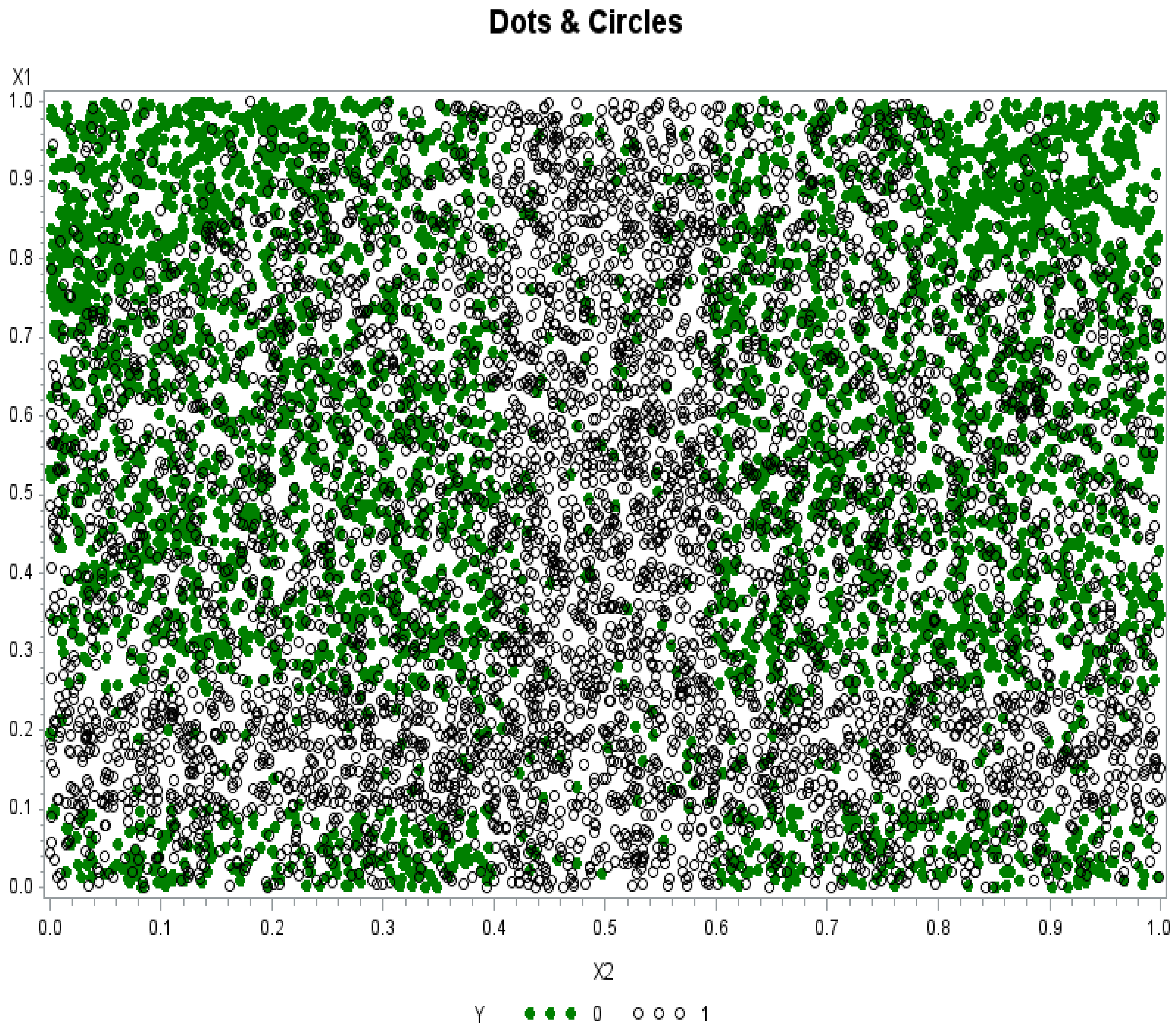
SAS GLOBAL FORUM

A Combined Example

Cell Phone Texting Locations

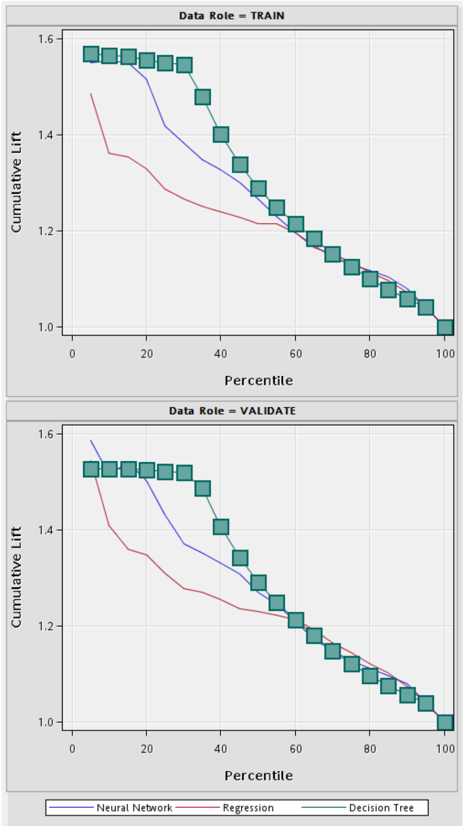
Black circle: ○
Phone moved > 50 feet in first two minutes of texting.

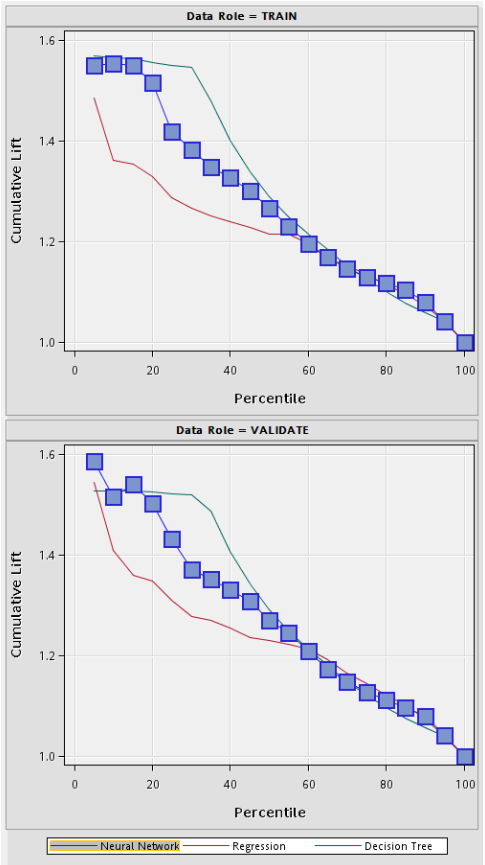
Green dot: ●
Phone moved < 50 feet. .

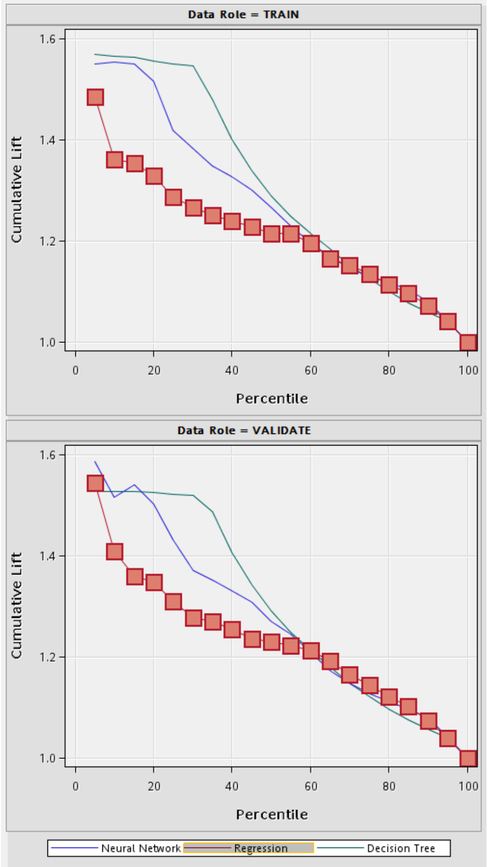


SAS GLOBAL FORUM

Tree | Neural Net | Logistic Regression ← Three Models



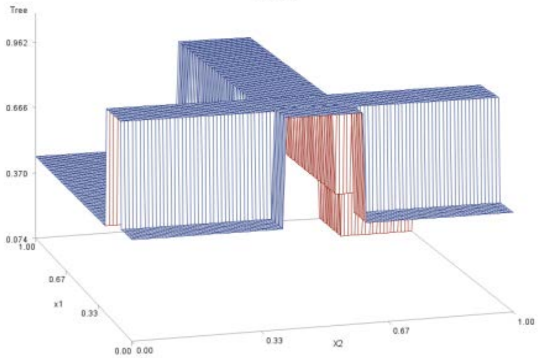




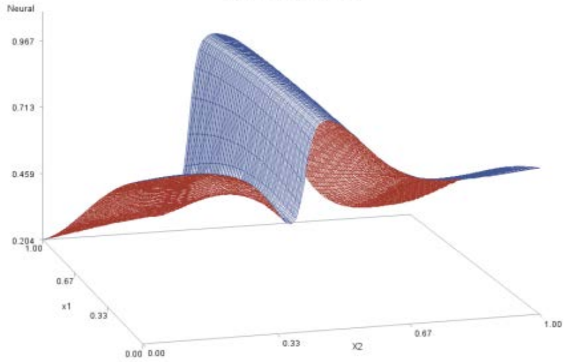
← Training Data Lift Charts

← Validation Data Lift Charts

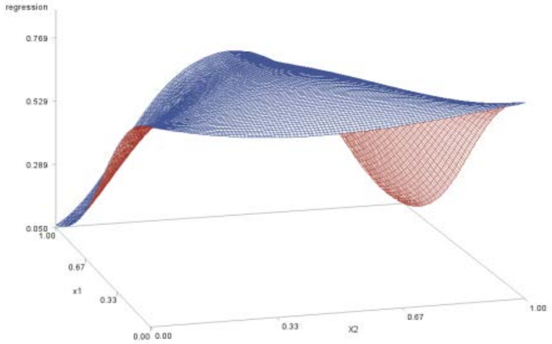
Tree



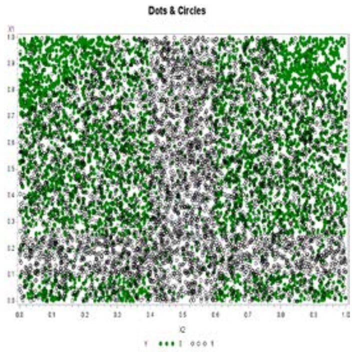
Neural Net



Logistic Regression

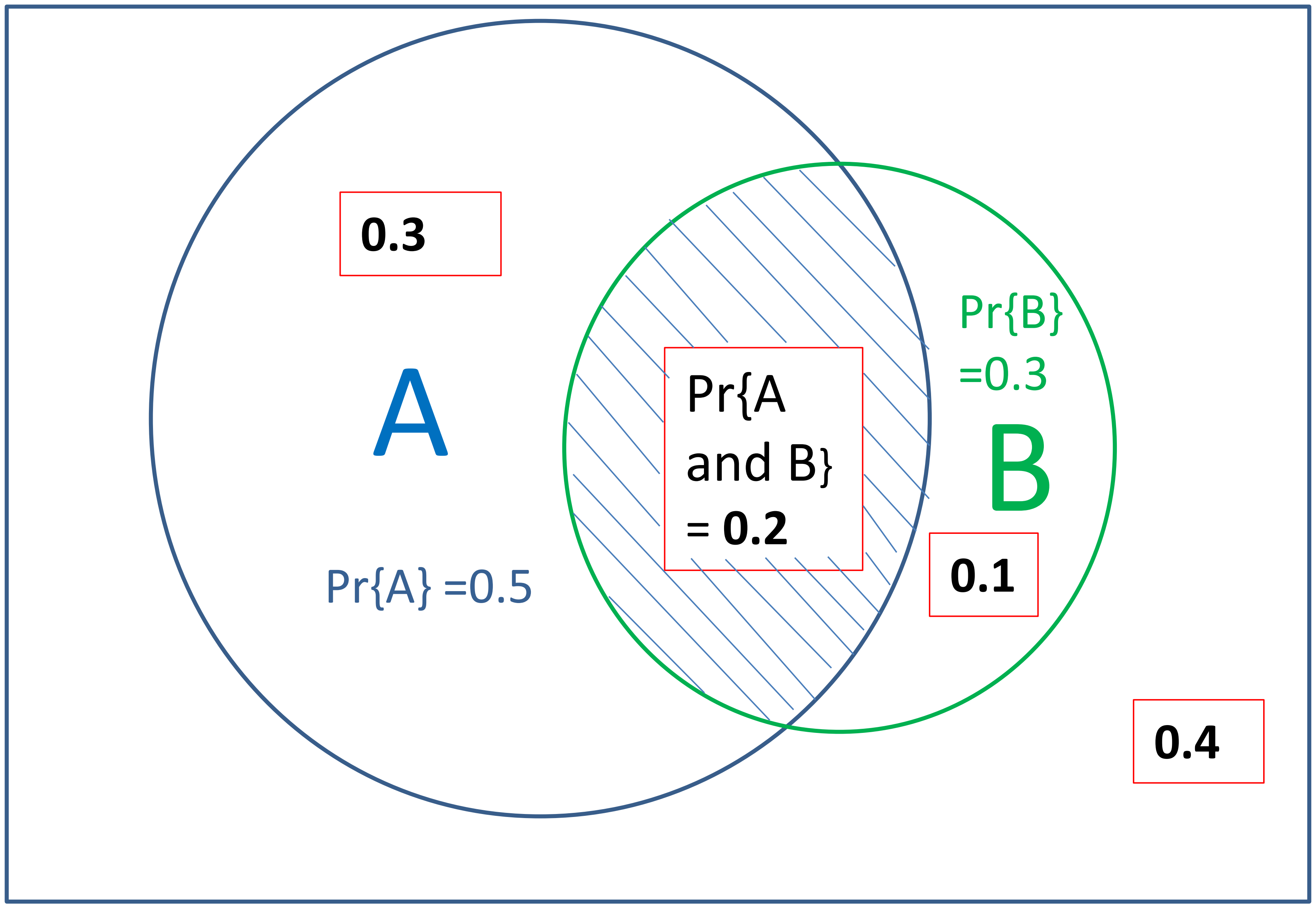


← Resulting Surfaces



SAS GLOBAL FORUM

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:

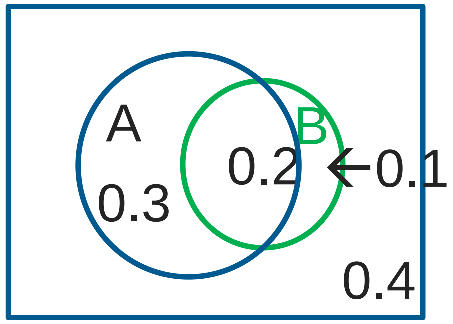
Confidence = Pr{A|B}=Pr{A and B}/Pr{B}=**2/3**

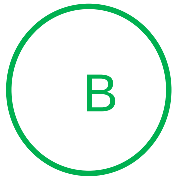
??- Is the confidence in B=>A the same as the confidence in A=>B?? (yes, no)

Lift:

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

Gain = 33%





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.

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
- EM
  - Large number of clusters with k-means (k clusters)
  - Ward’s method to combine (less clusters)
  - One more k means

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 16 e-mails

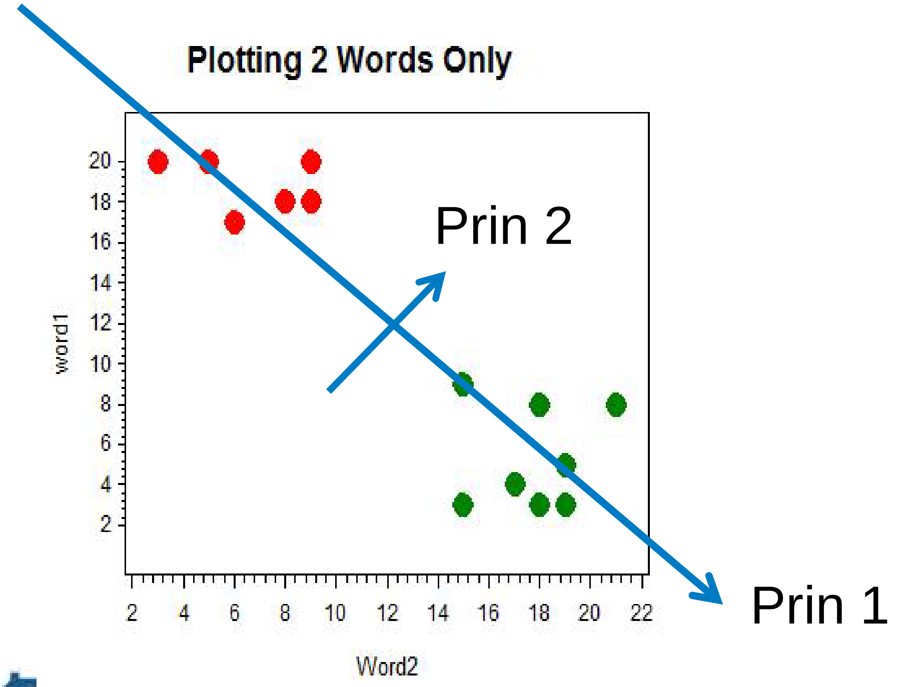
←--------------------------------words---→

document	Election	President	Republican	Basketball	Democrat	Voters	NCAA	Liar	Tournament	Speech	Wins	Score_V	Score_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

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

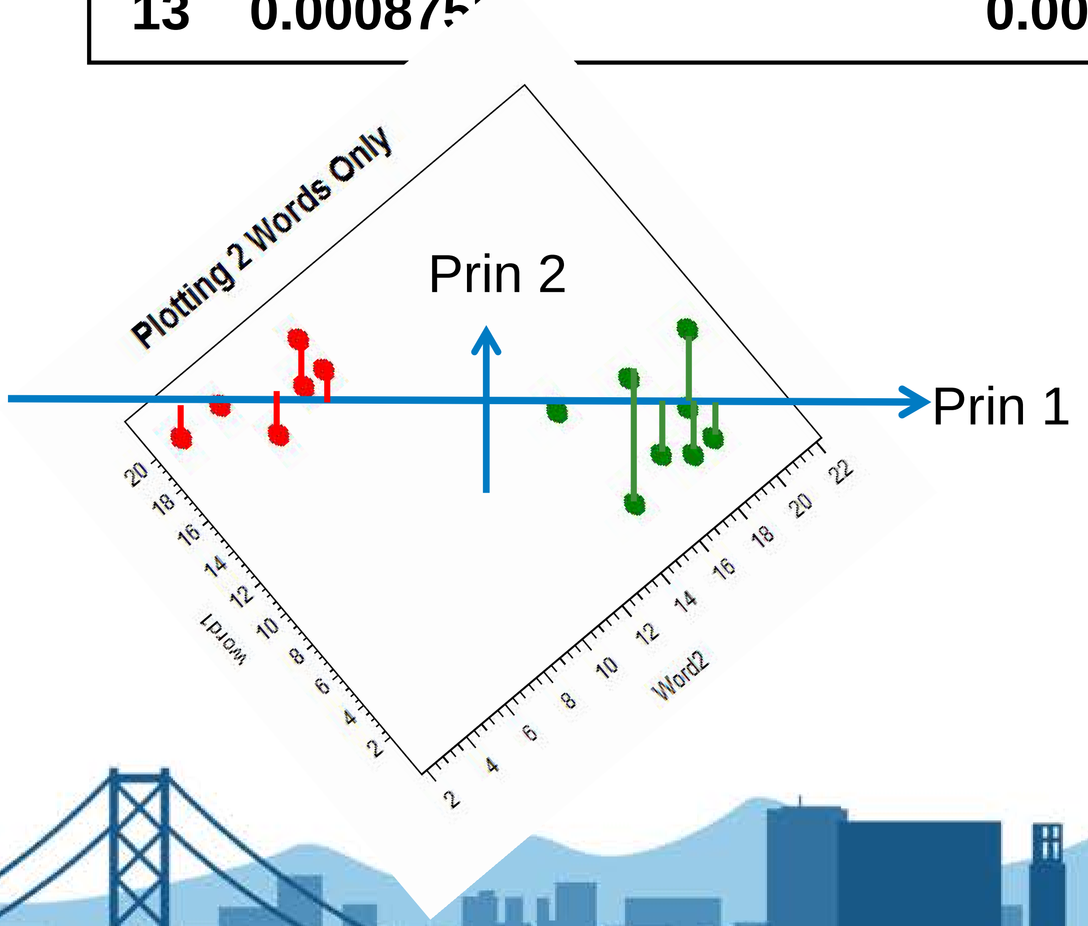


55% of the variation in these 13-dimensional vectors occurs in one dimension.

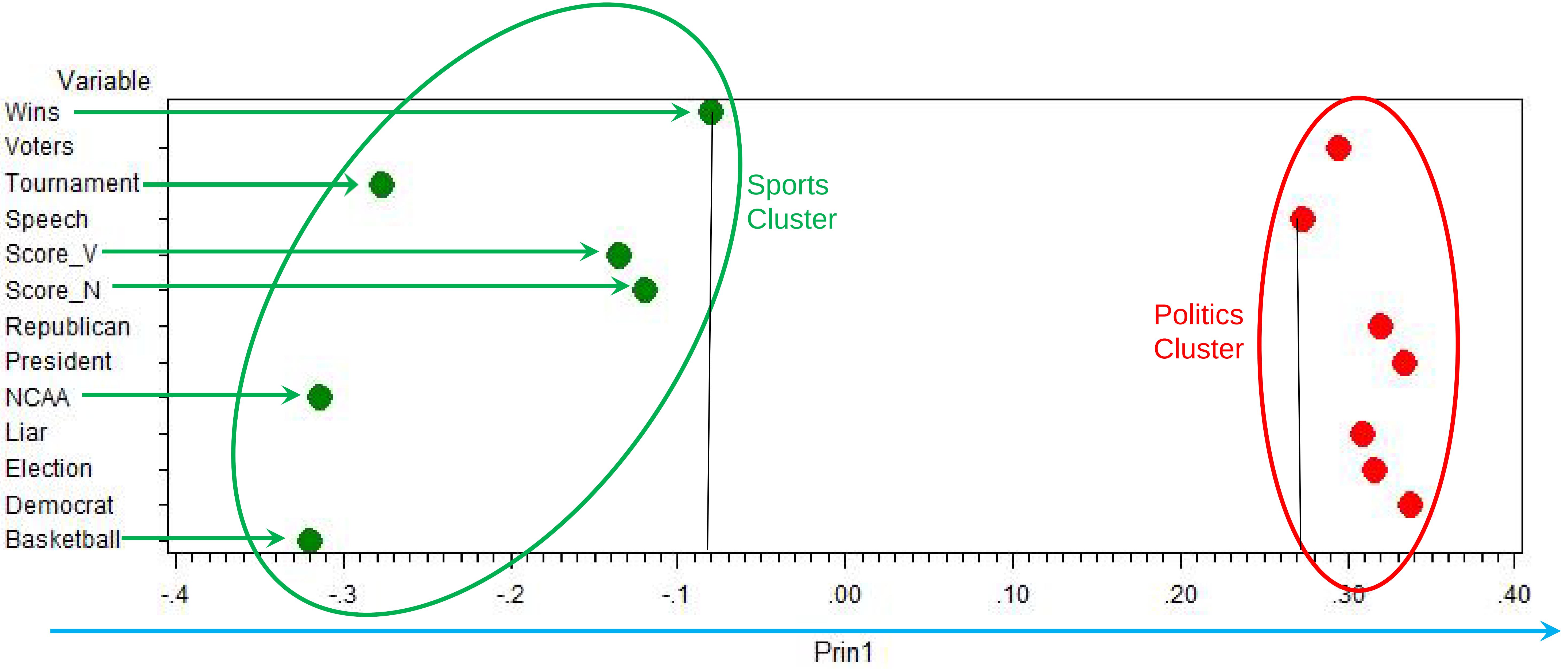
Eigenvalues of the Correlation Matrix

	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	0.000875		0.0001	1.0000

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



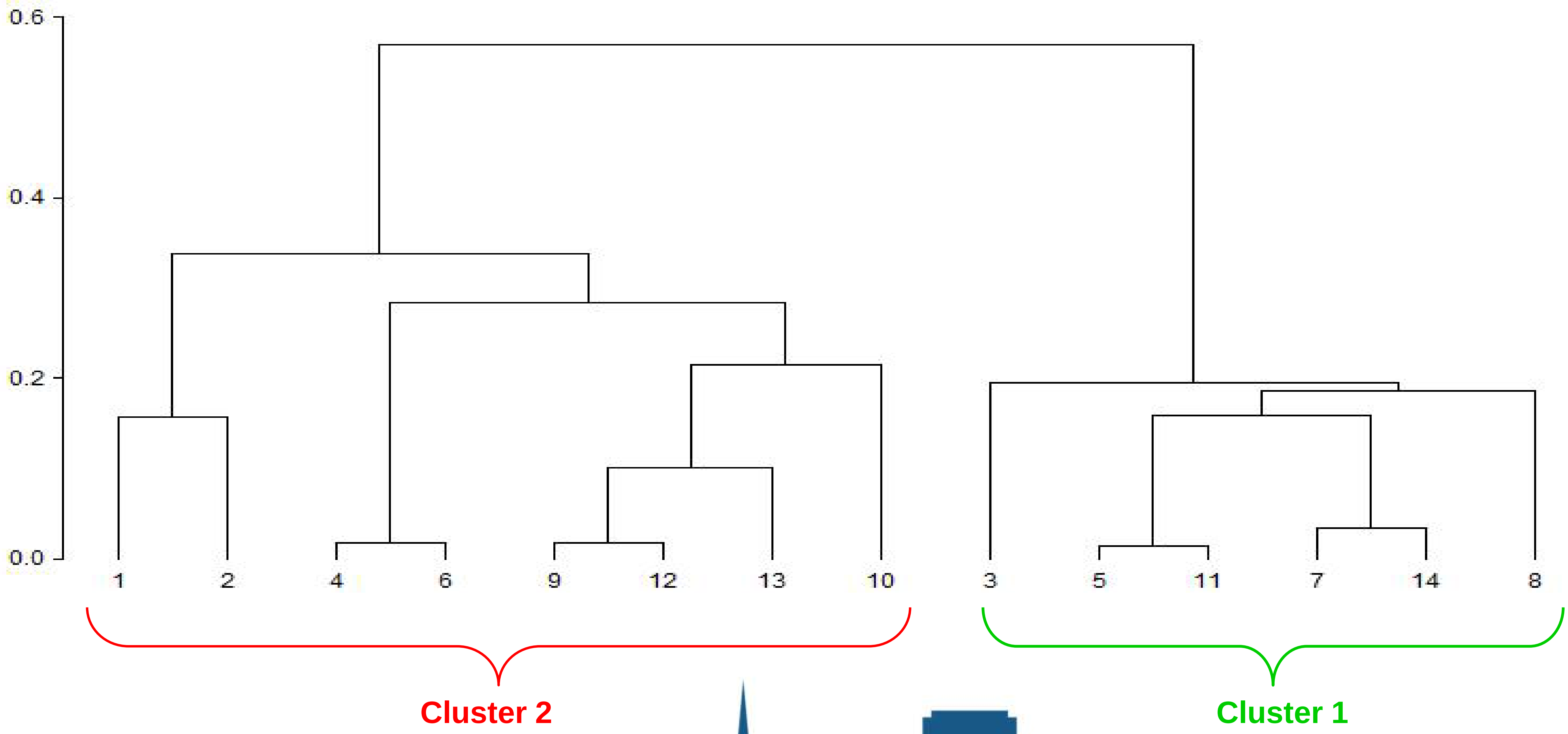
55% of the variation in these 13-dimensional vectors occurs in one dimension.



SAS GLOBAL FORUM

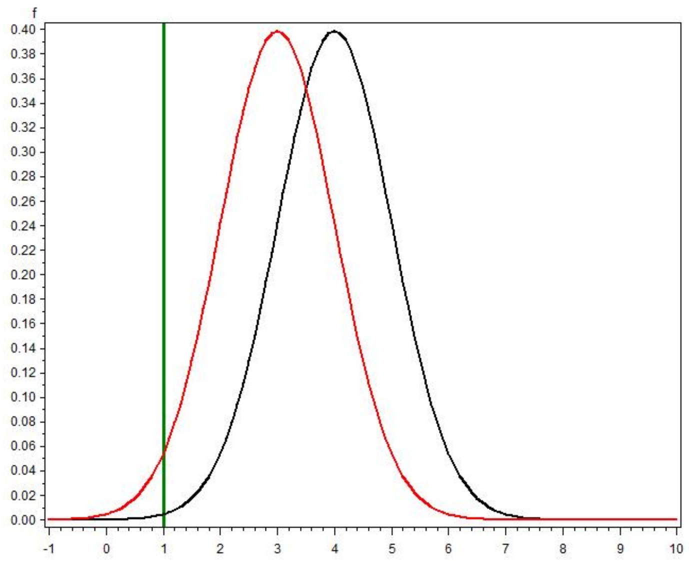
	document	CLUSTER	Prin1	Election	President	Republican	Basketball	Democrat	Voters	NCAA	Liar	Tournament	Speech	Wins	Score_V	Score_N
Sports Documents	3	1	-3.63815	0	2	0	14	0	2	12	0	16	4	24	19	30
	11	1	-3.02803	2	0	0	14	1	3	12	0	16	12	17	23	8
	5	1	-2.98347	0	0	4	16	0	0	15	2	17	3	9	0	1
	14	1	-2.48381	1	0	4	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 !



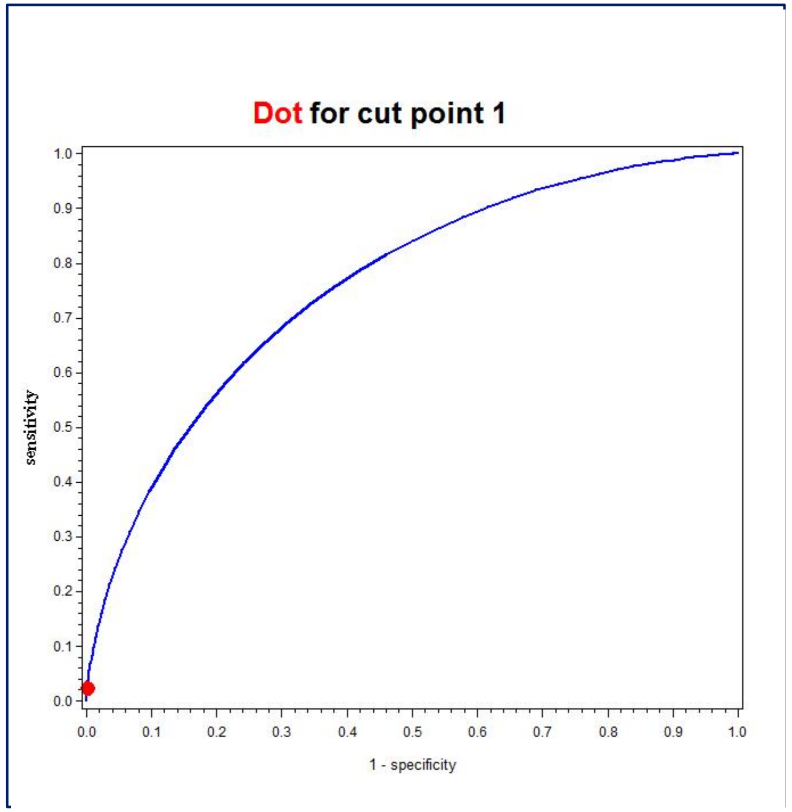
Receiver Operating Characteristic Curve

Cut point 1



Logits of 1s red

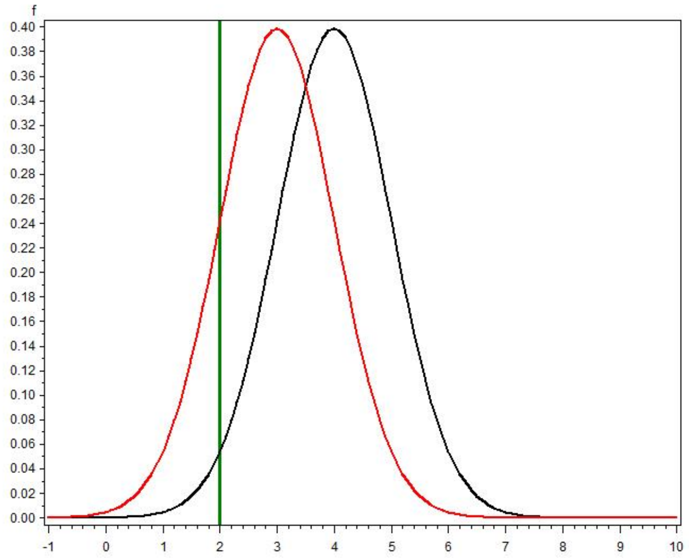
Logits of 0s black



SAS GLOBAL FORUM

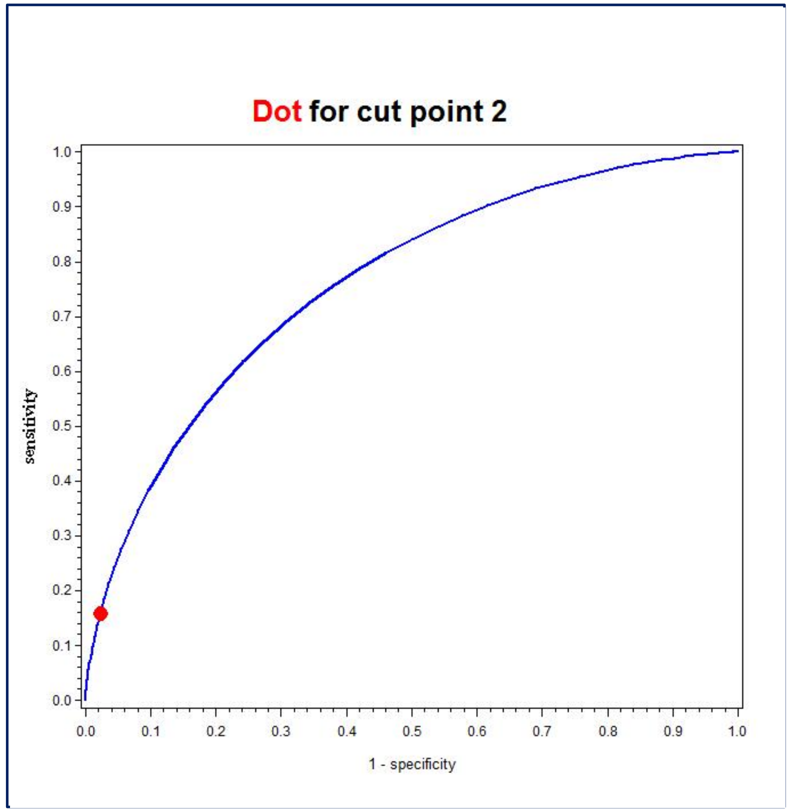
Receiver Operating Characteristic Curve

Cut point 2



Logits of 1s red

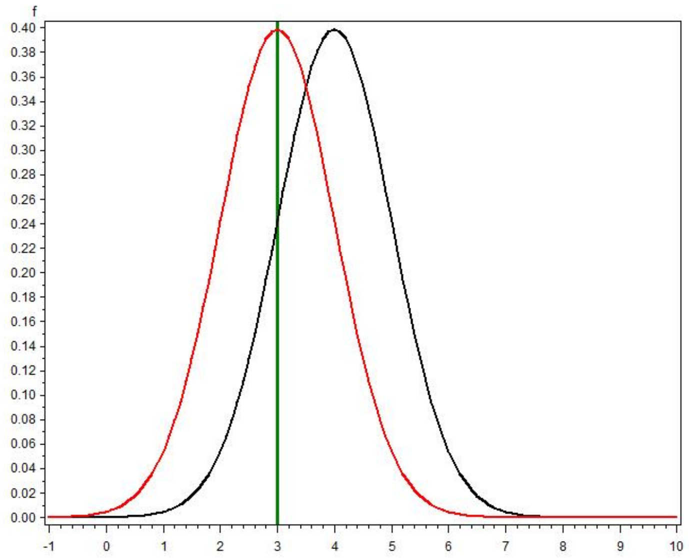
Logits of 0s black



SAS GLOBAL FORUM

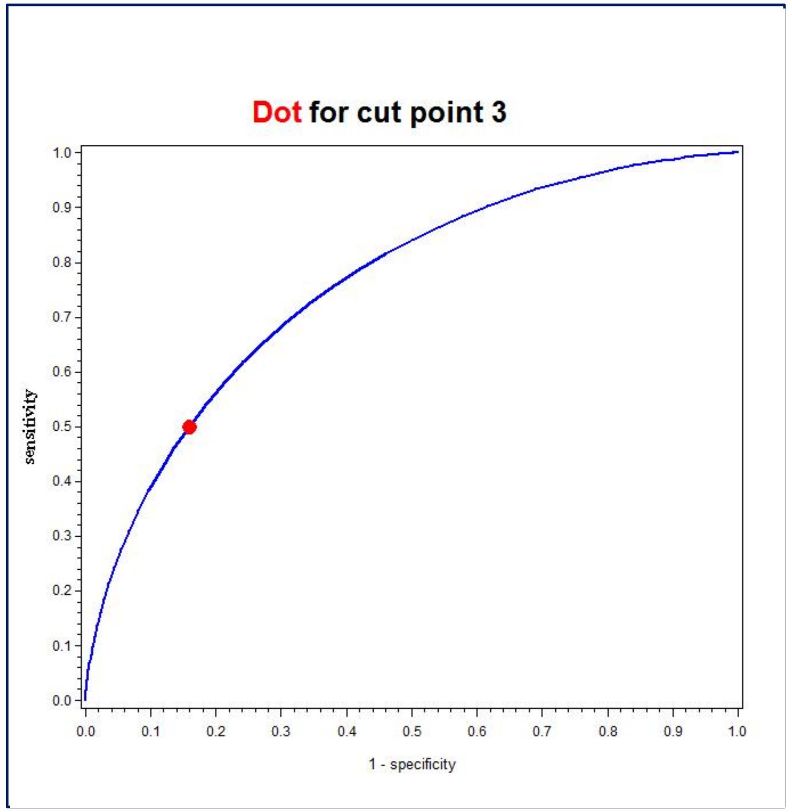
Receiver Operating Characteristic Curve

Cut point 3



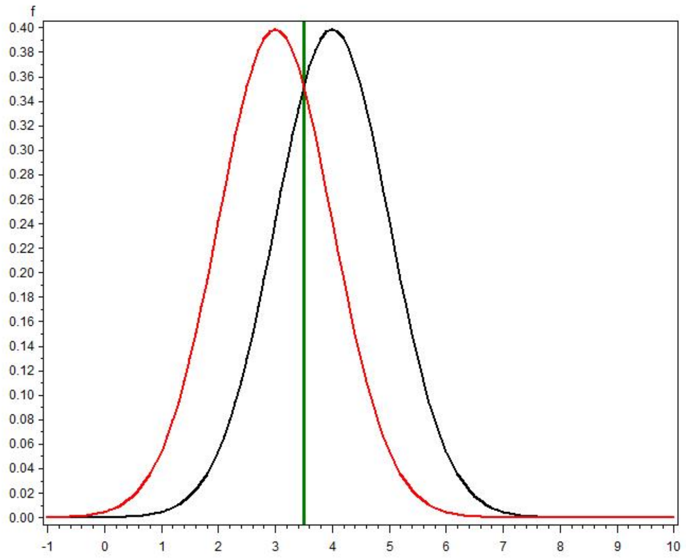
Logits of 1s red

Logits of 0s black



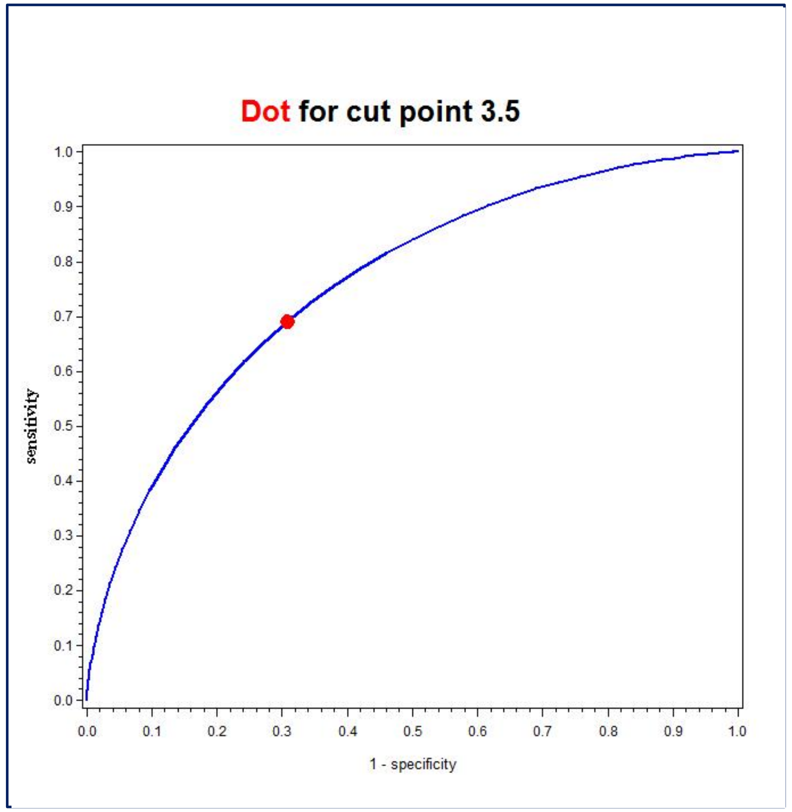
Receiver Operating Characteristic Curve

Cut point 3.5



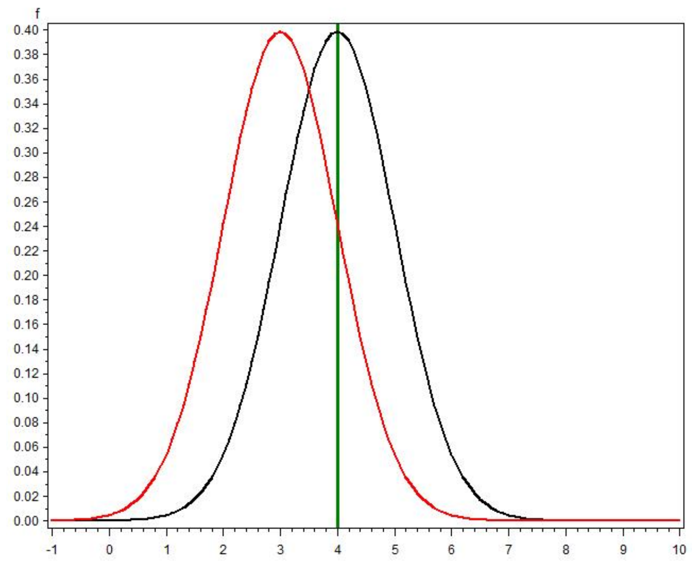
Logits of 1s red

Logits of 0s black



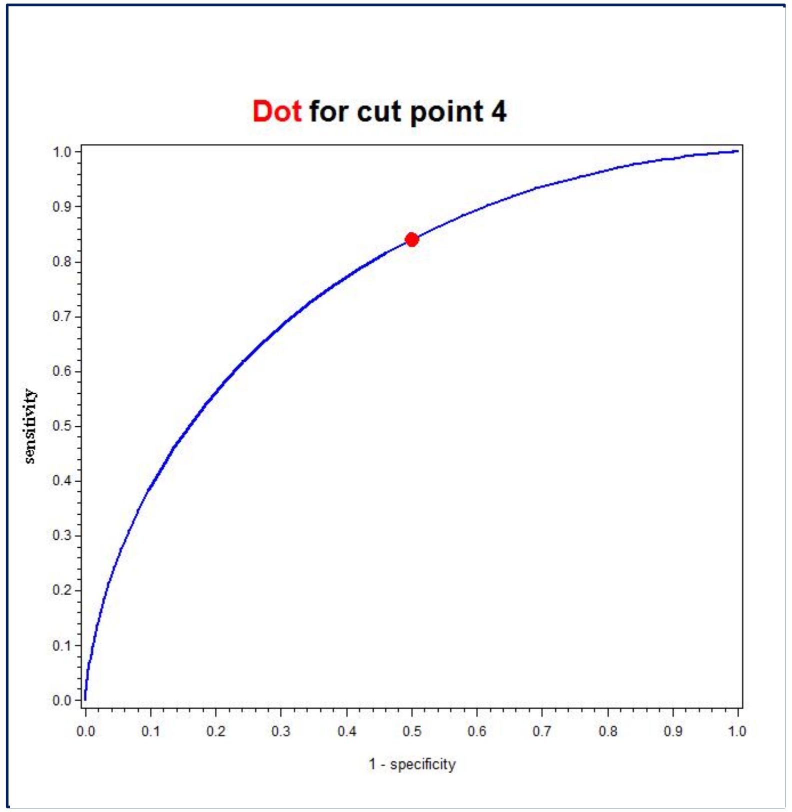
Receiver Operating Characteristic Curve

Cut point 4



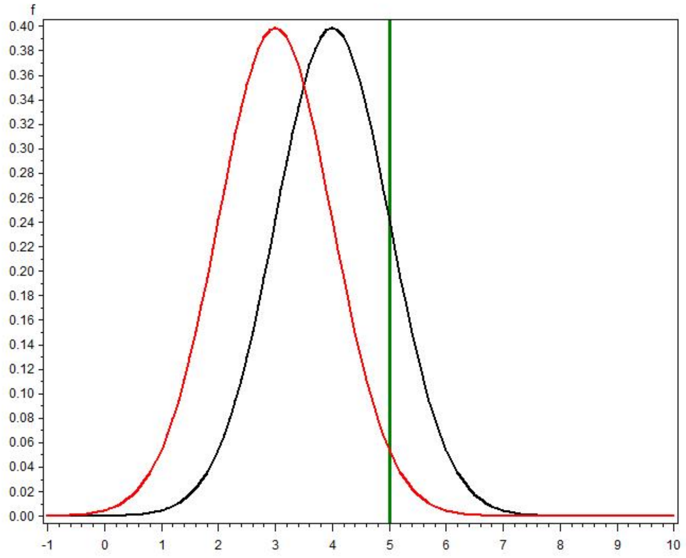
Logits of 1s red

Logits of 0s black



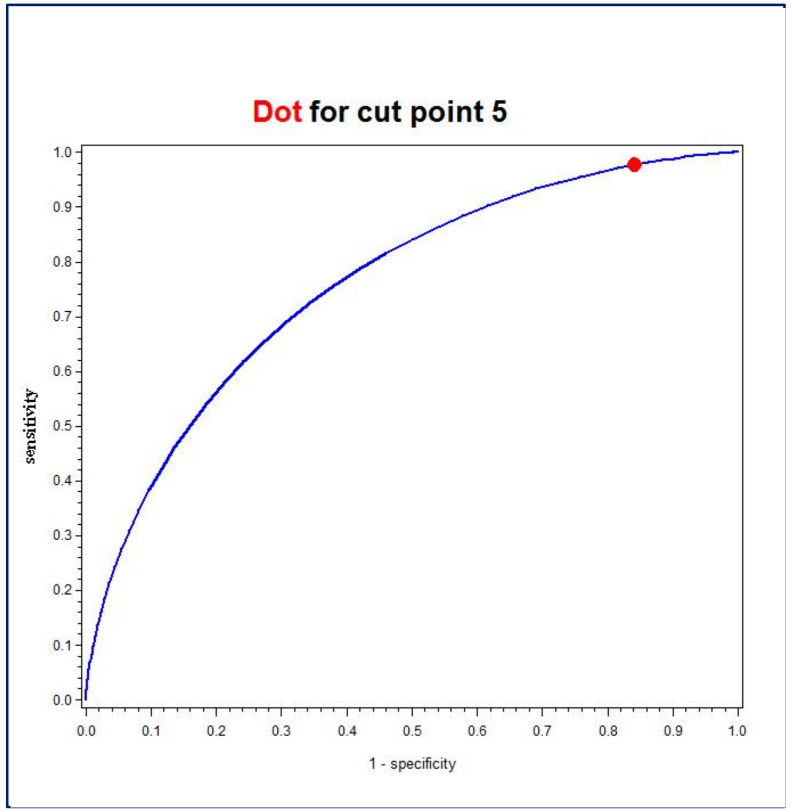
Receiver Operating Characteristic Curve

Cut point 5



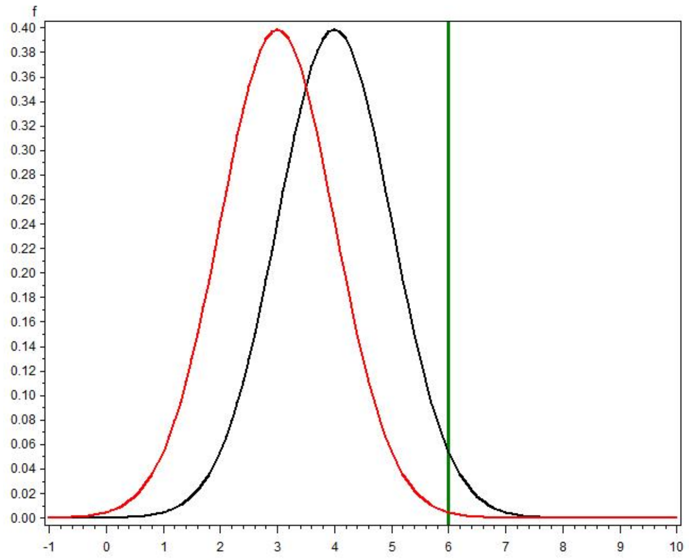
Logits of 1s red

Logits of 0s black



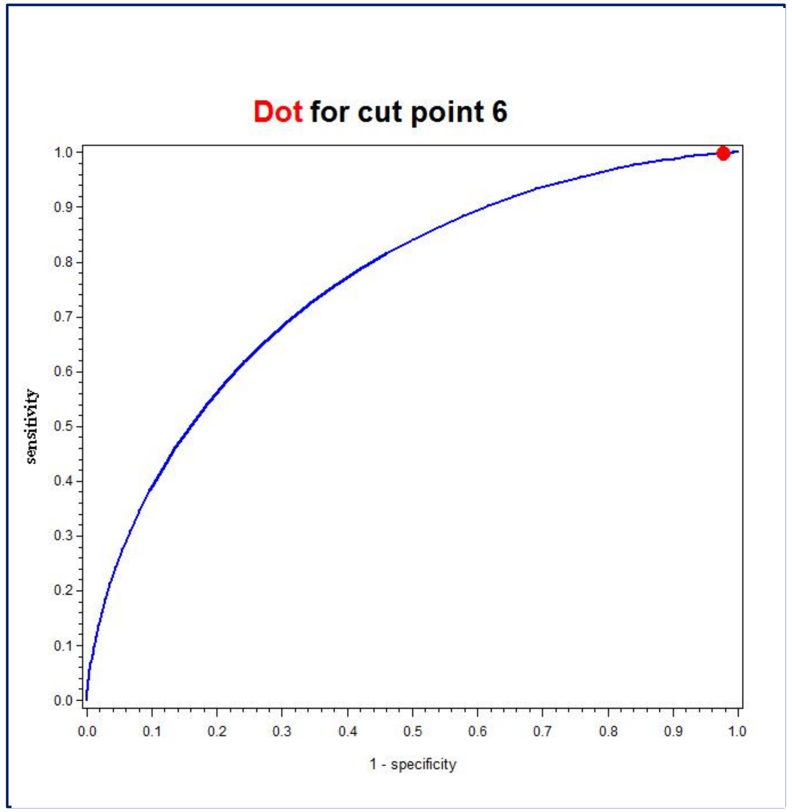
Receiver Operating Characteristic Curve

Cut point 6



Logits of 1s red

Logits of 0s black



SAS GLOBAL FORUM

"LOOKS LIKE MY TIME IS ALMOST UP"