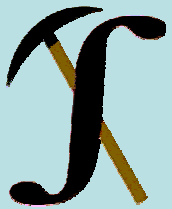
Other Methods

CAS 2004 Predictive Modeling Seminar
Chicago, Ill.

Louise Francis, FCAS, MAAA
E-mail: louise_francis@msn.com
www.data-mines.com



Francis Analytics and Actuarial
Data Mining, Inc.

Objectives

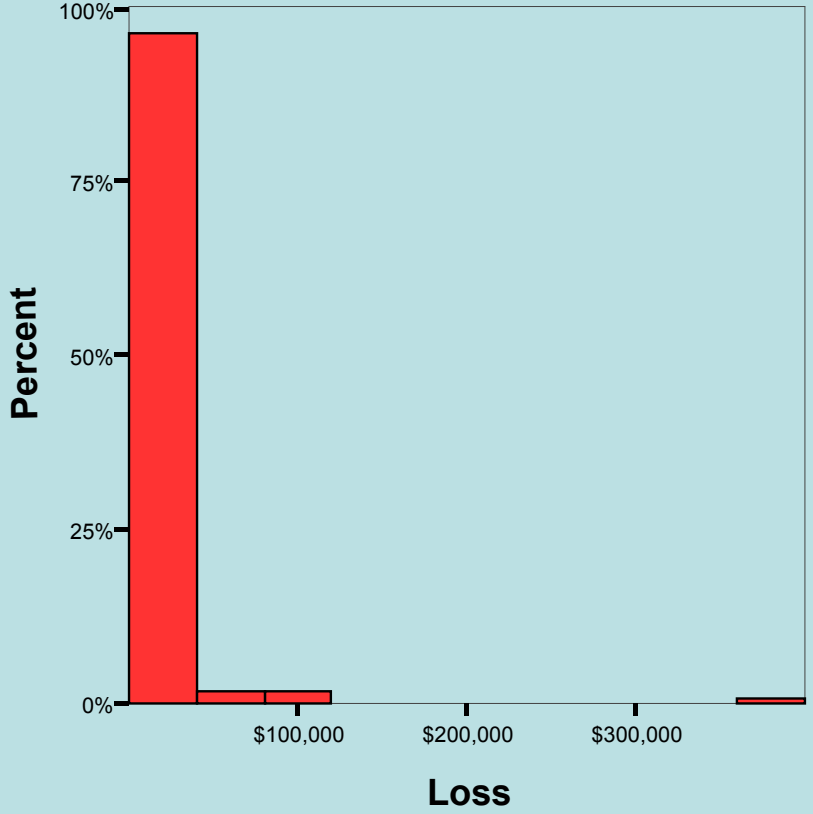
- Introduce some key analytical and data mining methods
  - Non-Parametric Regression
  - Neural Networks
  - MARS
  - Expectation Maximization

Data Challenges

- Nonlinearities
  - Relation between dependent variable and independent variables is not linear or cannot be transformed to linear
- Interactions
  - Relation between independent and dependent variable varies by one or more other variables
- Correlations
  - Predictor variables are correlated with each other

Non-Parametric Regression

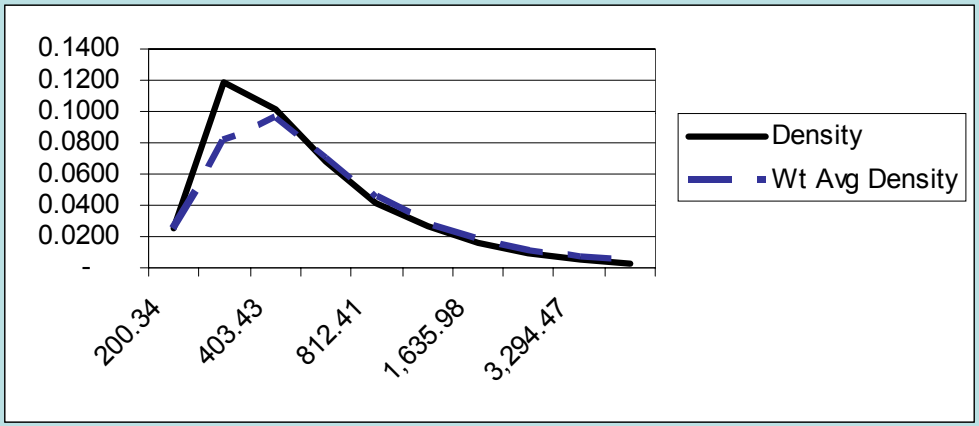
- A procedure for modeling nonlinear functions
- One of earliest applications was fitting distributions



Based on Weighted Average of Nearby Values

- Example is 3 point centered moving average
- Model: Density=f(Claim Size)
  - $f(x_t)=(d(x_{t+1})+d(x_t)+d(x_{t-1}))/3$
  - d=density=P(X=x)/h
  - h is width of histogram bin

ClaimSize	Density	Wt Avg Density
200.34	0.0250	0.0250
284.29	0.1191	0.0816
403.43	0.1007	0.0960
572.49	0.0680	0.0701
812.41	0.0417	0.0454
1,152.86	0.0264	0.0279
1,635.98	0.0155	0.0171
2,321.57	0.0095	0.0102
3,294.47	0.0057	0.0061
4,675.07	0.0033	0.0037



Kernel Estimator

- A method of weighting values of Y based on proximity to current X value
- K is weighting function
- h is window width (bandwidth)

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Examples of Kernel Estimators

Biweight $\frac{15}{16}(1-t^2)^2$ for $|t|<1$

0 otherwise

Triangular $1-|t|$ for $|t|<0$, 0 otherwise

Gaussian $\frac{1}{\sqrt{2\pi}}e^{-1/2t^2}$

Rectangular $\frac{1}{2}$ for $|t|<1$

$t=\frac{x-x_i}{h}$

An Example with Varying Bandwidth

- P(x) = # Claims in interval/Total
- Density = P(x) / Interval width
- Interval width is based on log(Claim Size) so it gets larger as claims get larger

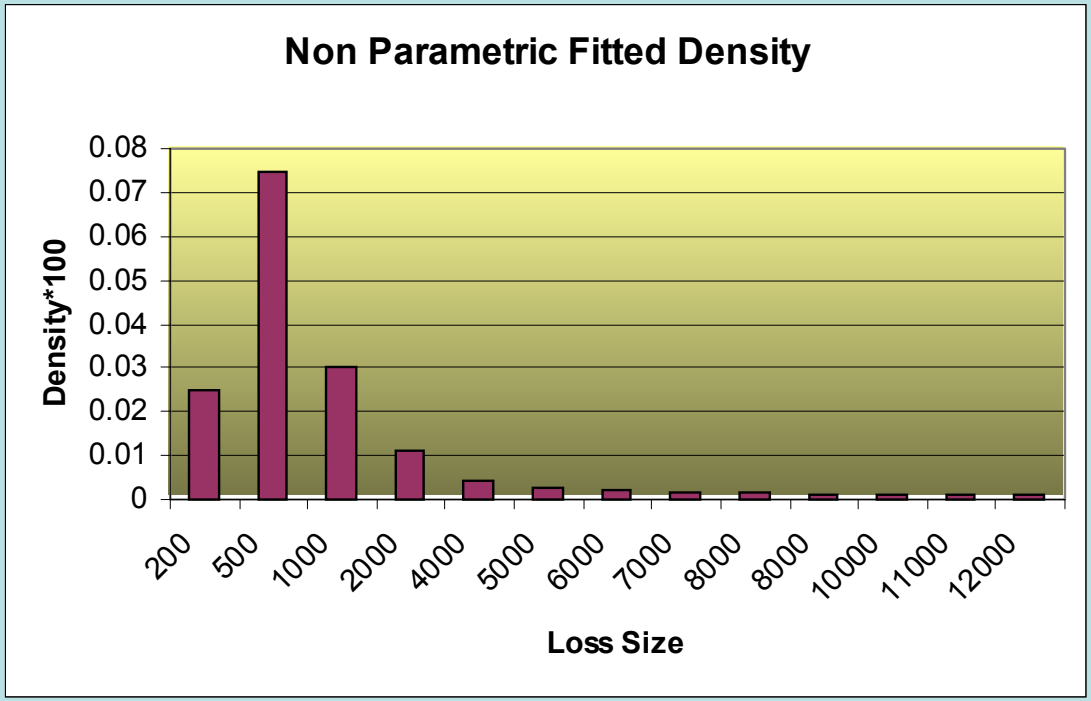
ln(Claim Size)	ClaimSize	P(X)	(P(x))/h=Density
5.3	200.34	0.05	0.0250
5.65	284.29	0.1	0.1191
6	403.43	0.12	0.1007
6.35	572.49	0.115	0.0680
6.7	812.41	0.1	0.0417
7.05	1,152.86	0.09	0.0264
7.4	1,635.98	0.075	0.0155
7.75	2,321.57	0.065	0.0095
8.1	3,294.47	0.055	0.0057
8.45	4,675.07	0.045	0.0033
8.8	6,634.24	0.04	0.0020
9.15	9,414.44	0.03	0.0011
9.5	13,359.73	0.025	0.0006

Example of Computing Kernel Estimate

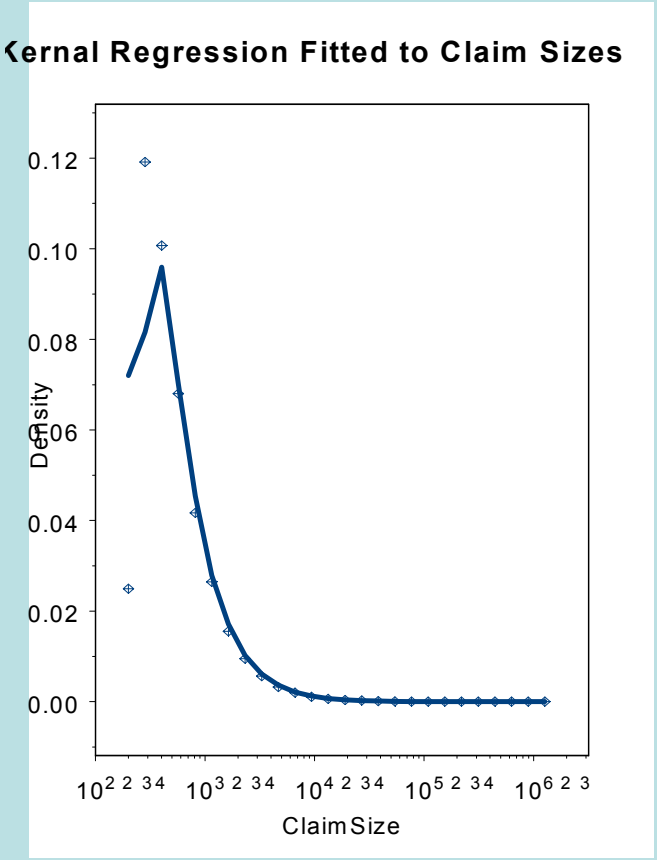
	X	500		1000	
	f(x)	0.0749		0.0302	
Claim Size	Density	exp(-t^2))	Weight	exp(-t^2))	Weight
200.34	0.0250	0.0000	0.000	0.0000	0.000
284.29	0.1191	0.0014	0.001	0.0000	0.000
403.43	0.1007	0.5184	0.331	0.0000	0.000
572.49	0.0680	0.8321	0.531	0.0017	0.001
812.41	0.0417	0.1835	0.117	0.5426	0.346
1,152.86	0.0264	0.0253	0.016	0.8174	0.521
1,635.98	0.0155	0.0040	0.003	0.1768	0.113
2,321.57	0.0095	0.0009	0.001	0.0243	0.016
3,294.47	0.0057	0.0003	0.000	0.0038	0.002
4,675.07	0.0033	0.0001	0.000	0.0008	0.001
6,634.24	0.0020	0.0001	0.000	0.0003	0.000
9,414.44	0.0011	0.0000	0.000	0.0001	0.000
13,359.73	0.0006	0.0000	0.000	0.0001	0.000
18,958.35	0.0004	0.0000	0.000	0.0000	0.000
26,903.19	0.0003	0.0000	0.000	0.0000	0.000
38,177.44	0.0001	0.0000	0.000	0.0000	0.000
54,176.36	0.0001	0.0000	0.000	0.0000	0.000
76,879.92	0.0000	0.0000	0.000	0.0000	0.000
109,097.80	0.0000	0.0000	0.000	0.0000	0.000
154,817.15	0.0000	0.0000	0.000	0.0000	0.000
219,695.99	-	0.0000	0.000	0.0000	0.000
311,763.45	-	0.0000	0.000	0.0000	0.000
442,413.39	0.0000	0.0000	0.000	0.0000	0.000
627,814.49	-	0.0000	0.000	0.0000	0.000
890,911.17	-	0.0000	0.000	0.0000	0.000
1,264,263.12	-	0.0000	0.000	0.0000	0.000
Total			1.000		1.000

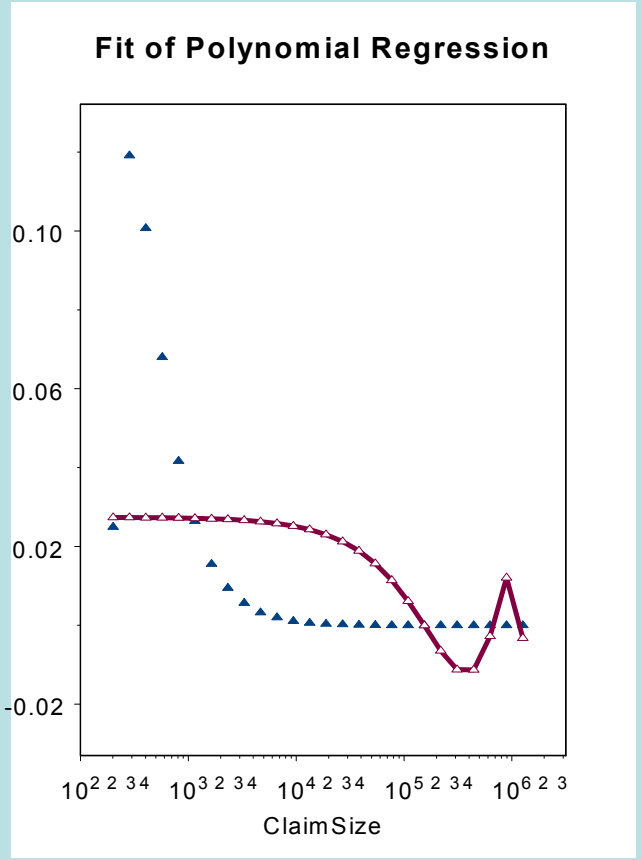
Fitted Density Function

- Endpoints a problem
- Ignored kernel function and used closest Y value



Kernel Model vs Polynomial Regression





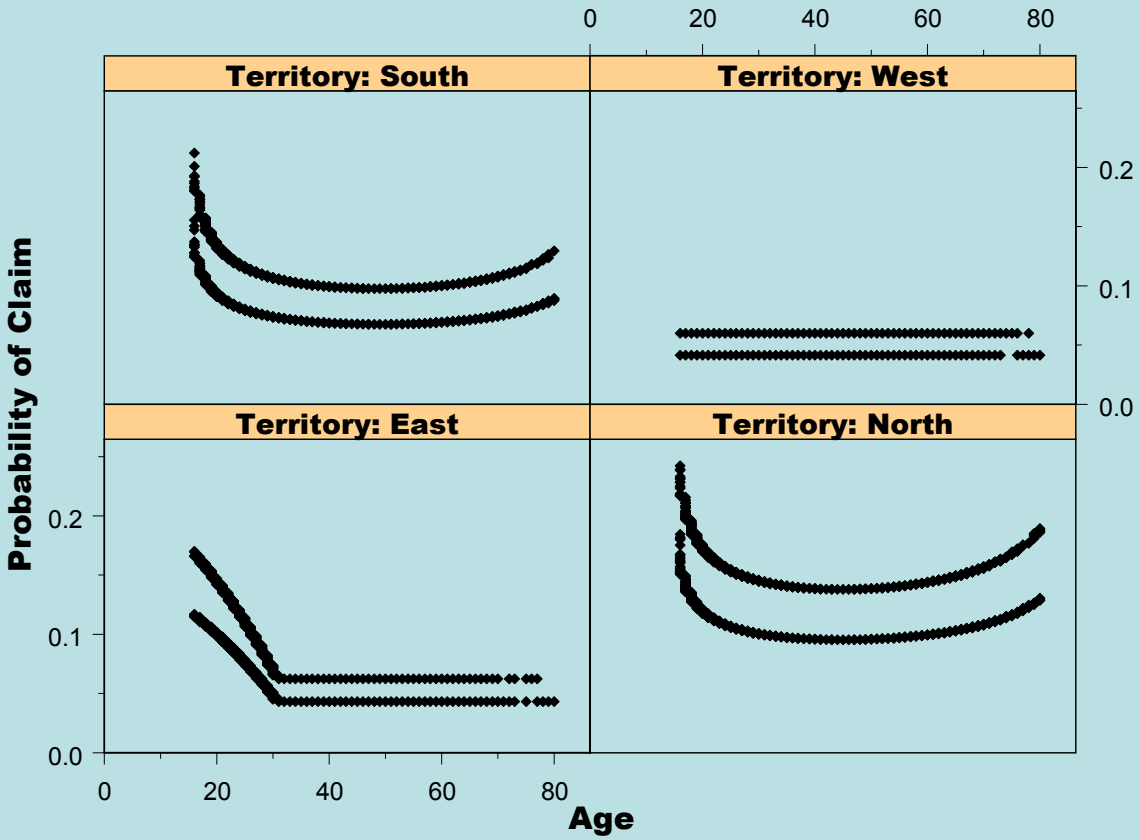
Francis Analytics and Actuarial
Data Mining, Inc.

Example Data

- Simulated Data for Automobile Claim Frequency
- Three Factors
  - Territory
    - Four Territories
  - Age
    - Continuous Function
  - Mileage
    - High, Low

Simulated Example: Probability of Claim vs. Age

by Territory and Mileage Group



Francis Analytics and Actuarial
Data Mining, Inc.

Independent Probabilities for Each Variable

Claim Count * Territory

Mean

		Claim Count
Territory	East	.06
	North	.13
	South	.10
	West	.05
	Total	.10

Claim Count * Mileage Group

Mean

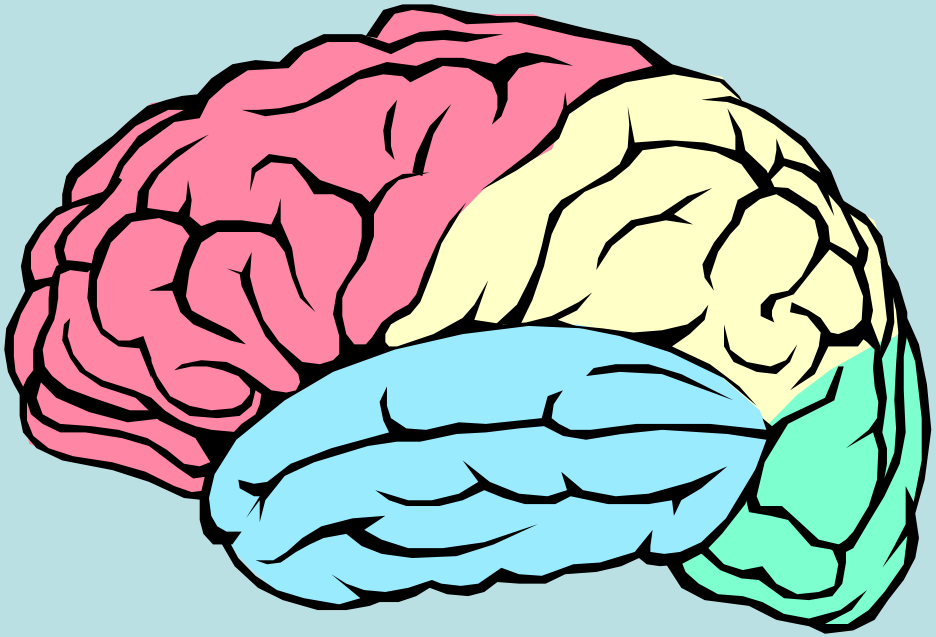
		Claim Count
Mileage Group	High	.12
	Low	.08
	Total	.10

Mean

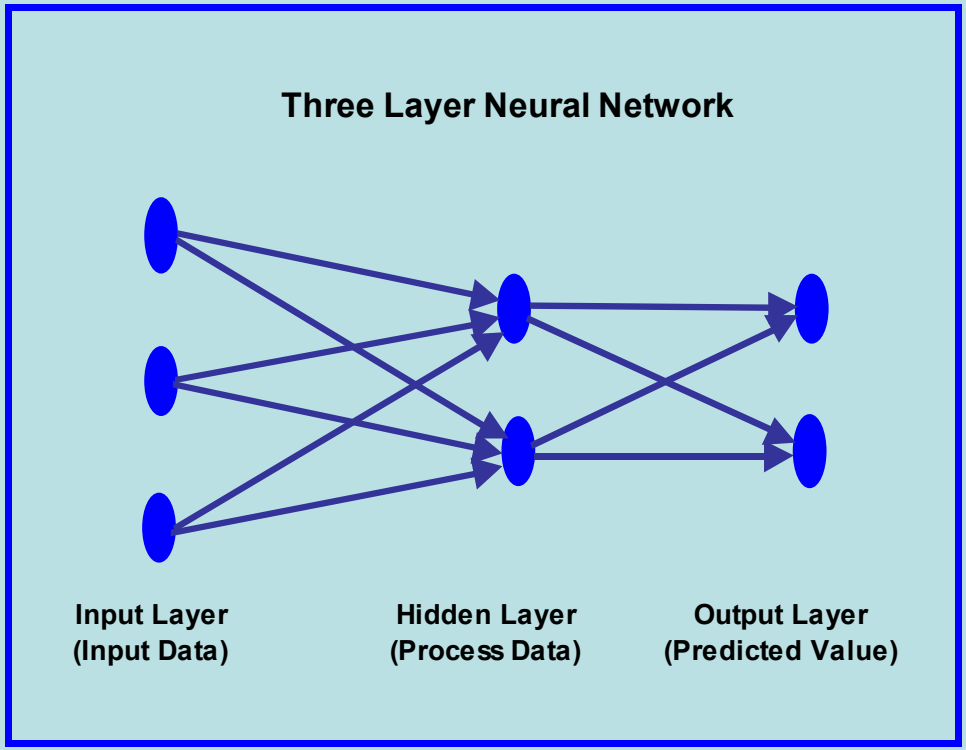
		Claim Count
Age Group	18.5	.13
	25.0	.11
	35.0	.09
	45.0	.09
	55.0	.09
	65.0	.10
	75.0	.13
	85.0	.18
	Total	.10

Neural Networks

- Developed by artificial intelligence experts – but now used by statisticians also
- Based on how neurons function in brain



Neural Network Structure



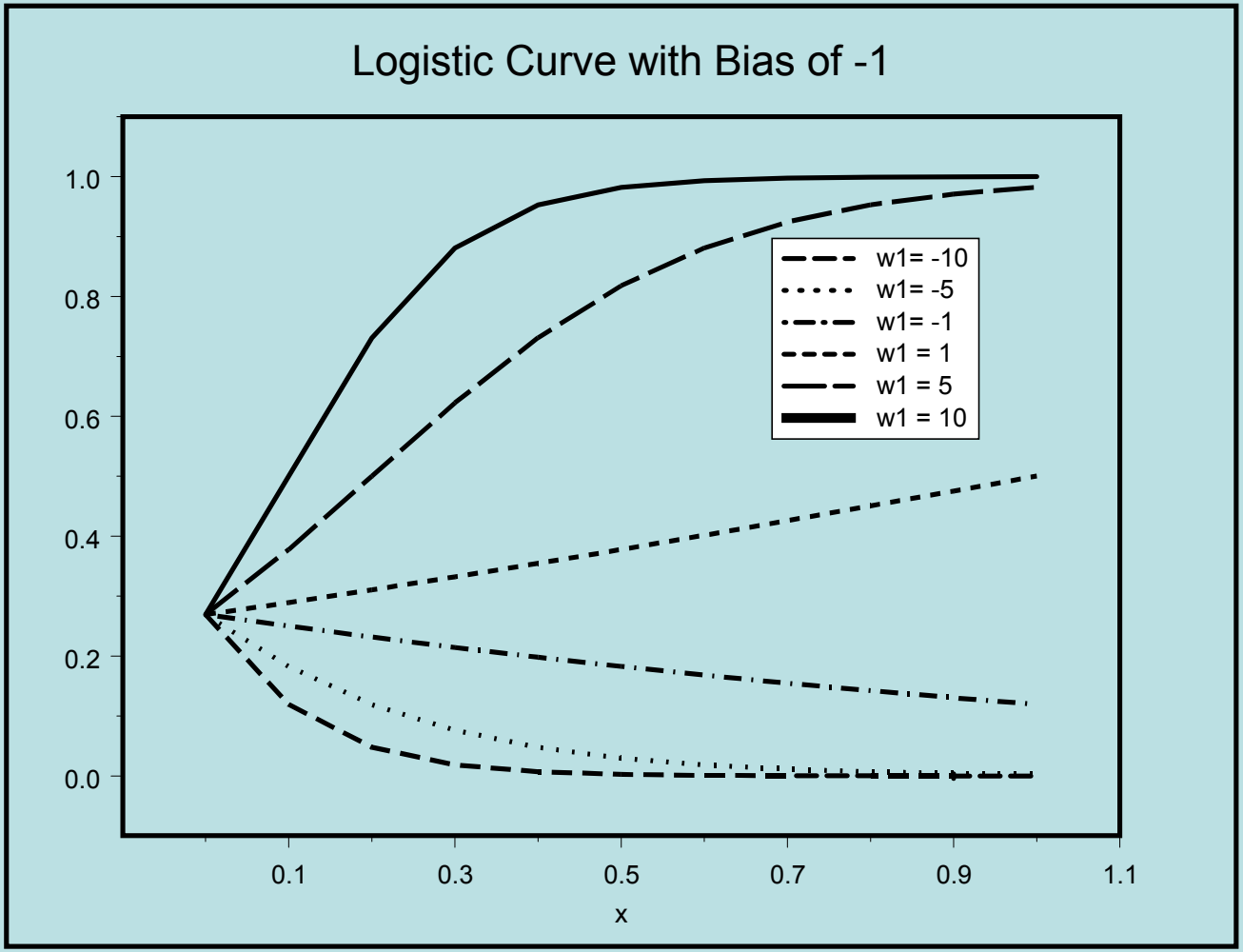
The Activation Function

- The sigmoid logistic function

$$f(Y) = \frac{1}{1 + e^{-Y}}$$

$$Y = w_0 + w_1 * X_1 + w_2 X_2 \ldots + w_n X_n$$

Variety of Shapes with Logistic Curve



Universal Function Approximator

- The feedforward neural network with one hidden layer is a universal function approximator

- Theoretically, with a sufficient number of nodes in the hidden layer, any continuous nonlinear function can be approximated

Function if Network has One Hidden Node

$$h = f(X; w_{0,} w_1) = f(w_0 + w_1 X) = \frac{1}{1 + e^{-(w_0 + w_1 X)}}$$

$$f(f(X; w_0, w_1); w_2, w_3) = \frac{1}{1 + e^{-(w_2 + w_3 \frac{1}{1 + e^{-w_0 + w_1 X}})}}$$

Neural Networks

- Fit by minimizing squared deviation between fitted and actual values
- Can be viewed as a non-parametric, non-linear regression
- Often thought of as a “black box”
  - Due to complexity of fitted model it is difficult to understand relationship between dependent and predictor variables

How Many Hidden Nodes for Neural Network?

- Too few nodes: Don't fit the curve very well
- Too many nodes: Over parameterization
  - May fit noise as well as pattern

How Do We Determine the Number of Hidden Nodes?

- Use methods that assess goodness of fit
- Hold out part of the sample
- Resampling
  - Bootstrapping
  - Jacknifing
- Algebraic formula
  - Uses gradient and Hessian matrices

Understanding the Model:
Variable Importance

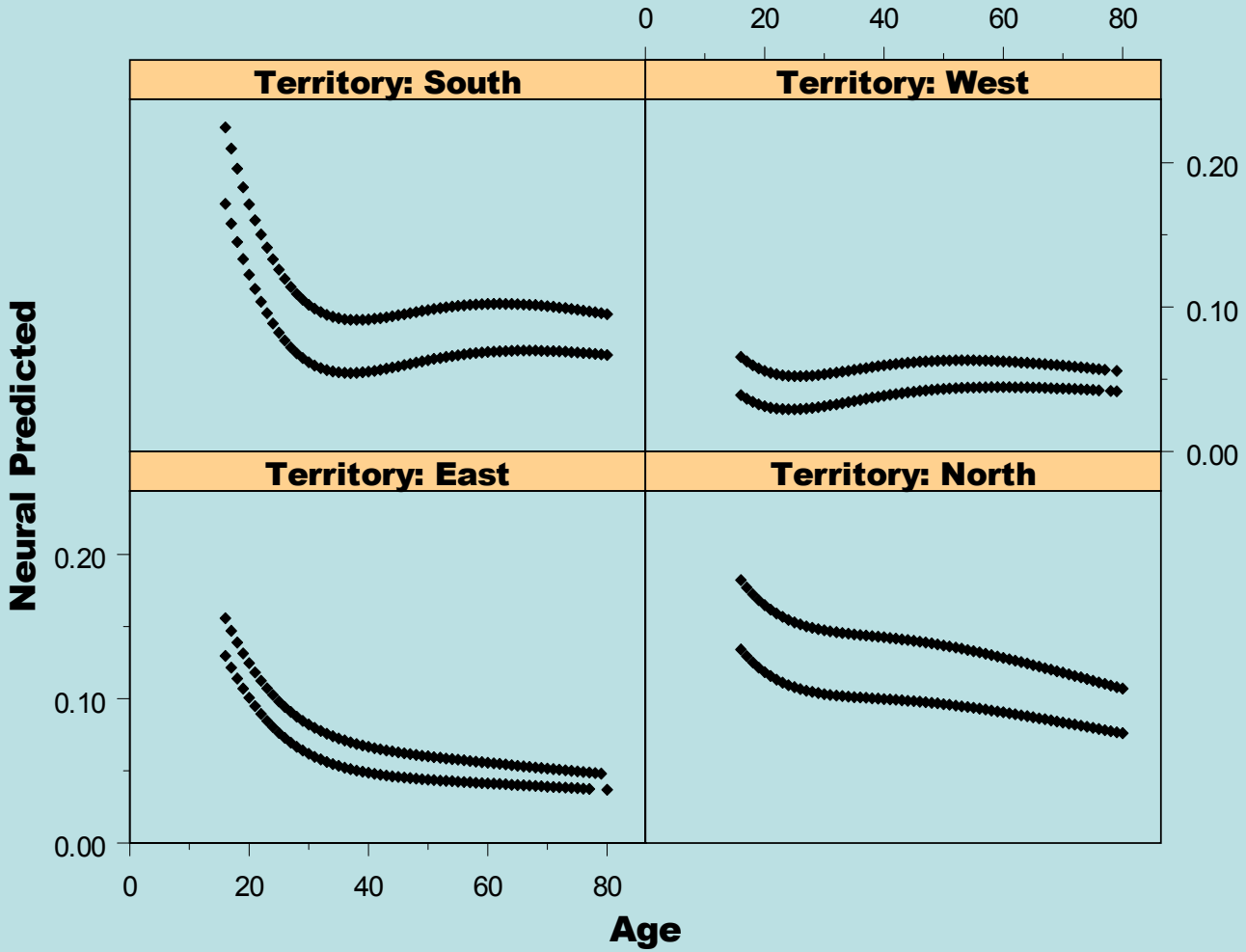
- Look at weights to hidden layer
- Compute sensitivities:
  - a measure of how much the predicted value's error increases when the variables are excluded from the model one at a time

Importance Ranking

- Neural Network and Mars ranked variables in same order

Variable	Neural Net Rank	MARS Rank
Territory	1	1
Age	2	2
Mileage	3	3

Visualizing Fitted Neural Network



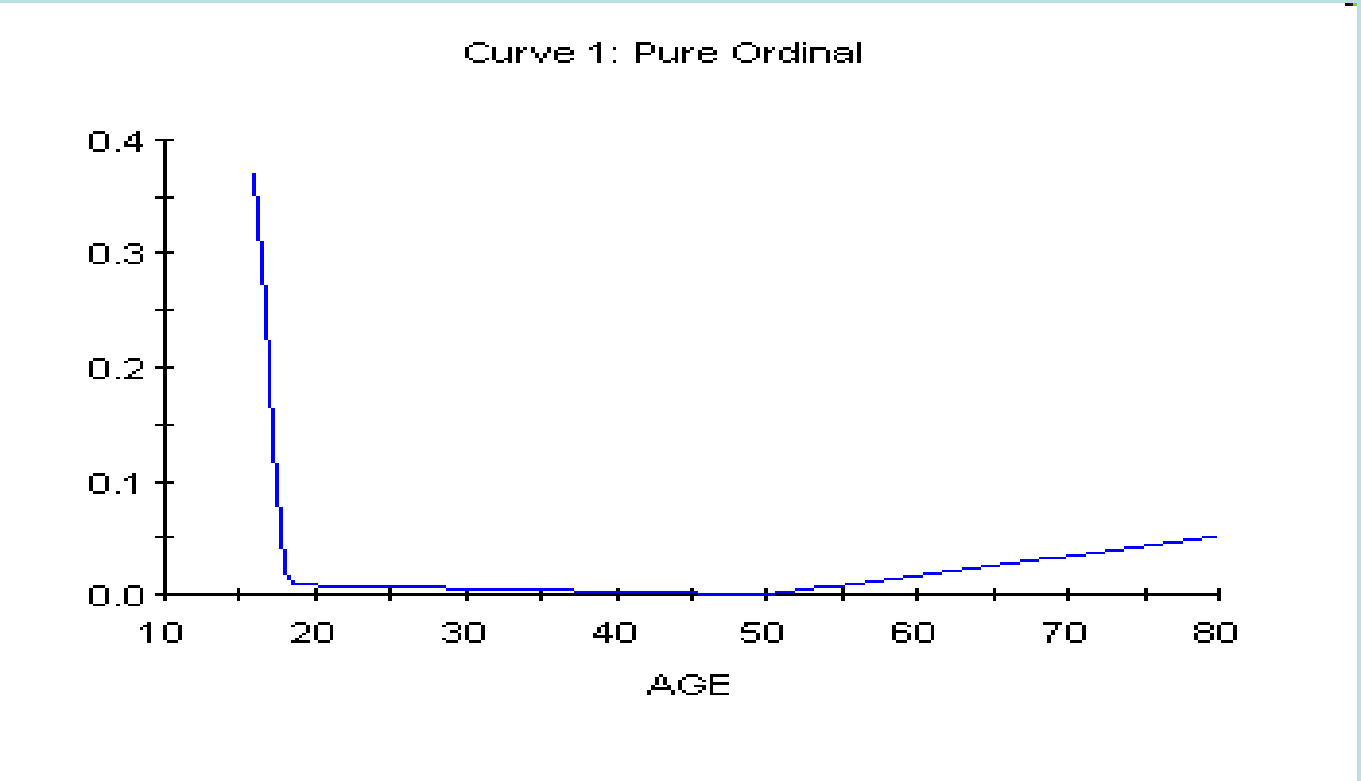
Francis Analytics and Actuarial Data Mining, Inc.

MARS

- Multivariate Adaptive Regression Splines
- An extension of regression which
  - Uses automated search procedures
  - Models nonlinearities
  - Models interactions
  - Produces a regression-like formula

Nonlinear Relationships

- Fits piecewise regression to continuous variables



Interactions

- Fits basis functions (which are like dummy variables) to model interactions
  - An interaction between Territory=East and Mileage can be modeled by a dummy variable which is 1 if the Territory=East and mileage =High and 0 otherwise.

Goodness of Fit Statistics

- Generalized Cross-Validation

$$GCV = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - k/N}\right]^2$$

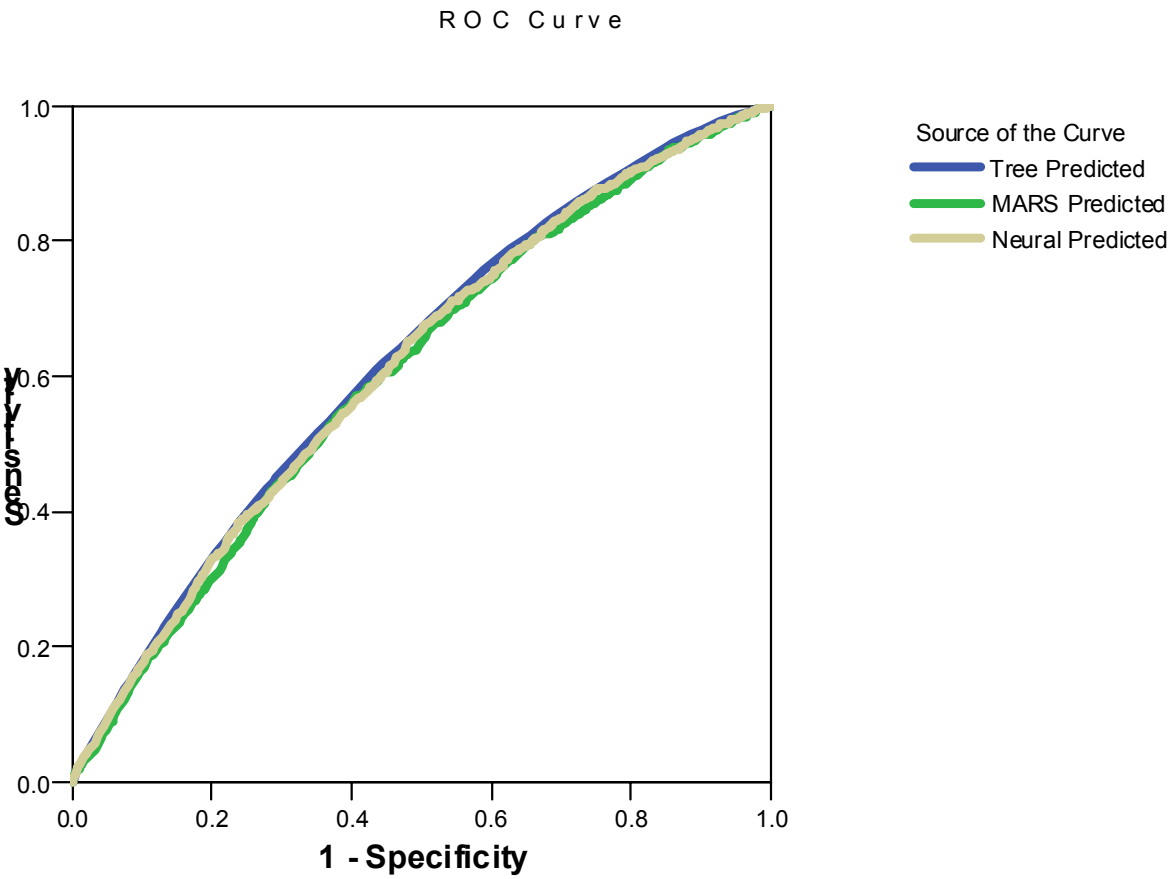
where N is the number of observations
y is the dependent variable
x is the independent variable(s)
k is the effective number of parameters or degrees of freedom in the model.

Fitted MARS Model

```
Basis Functions:
 BF1 = ( TERRITORY = 2 OR TERRITORY = 3);
 BF3 = (MILEAGE = HIGH);
 BF5 = ( TERRITORY = 1 OR TERRITORY = 2);
 BF7 = max(0, AGE - 50.000);
 BF8 = max(0, 50.000 - AGE );
 BF9 = max(0, AGE - 18.000);
 BF10 = max(0, 18.000 - AGE );
 BF11 = ( TERRITORY = 2 OR TERRITORY = 4) * BF10;
 BF13 = ( TERRITORY = 1) * BF9;
 BF17 = max(0, AGE - 19.000) * BF3;
 BF18 = max(0, 19.000 - AGE ) * BF3;
 BF19 = max(0, AGE - 22.000) * BF3;

Model
 Y = -3.887 + 0.044 * BF1 + 0.032 * BF5 - 0.121 * BF7 +
0.124 * BF8+ 0.123 * BF9 - 0.071 * BF11 - .979823E-03 *
BF13 - 0.011 * BF17 - 0.049 * BF18 + 0.011 * BF19;
```

ROC Curves for the Data Mining Methods



Francis Analytics and Actuarial
Data Mining, Inc.

Correlation

- Variable gender added
- Its only impact on probability of a claim: correlation with mileage variable – males had higher mileage
  - MARS did not use the variable in model
  - CART used it in two places to split tree
  - Neural Network ranked gender as least important variable

Expectation Maximization (EM)

- One of main applications is filling in missing values
- Also used in for applications where there is no closed form solution and deriving an estimate is particularly challenging otherwise

Expectation Maximization (EM)

- Two Steps
  - Expectation

$$Q(\Theta, \Theta_{i-1}) = E[\sum \ln(p(x_i, y_i \mid \Theta \mid x_i, \Theta_{i-1})]$$

  - Maximization

$$\Theta_i = \max(Q(\Theta, \Theta_{i-1}))$$

The Fraud Study Data

- 1993 Automobile Insurers Bureau closed Personal Injury Protection claims
- Dependent Variables
  - Suspicion Score
  - Expert assessment of Fraud
- Predictor Variables
  - Red flag indicators
  - Claim file variables – like legal representation, payment amount, etc.

Example: Suspicion Score vs Legal Representation

Report

Suspicion Level

		Mean	N	Std. Deviation
Legal representation	1	.6275	639	1.67096
	2	3.2500	616	2.78249
	Total	1.9147	1255	2.63395

Example of EM

- Suppose 10% of values for legal representation are missing
- We can use relation between suspicion score and legal representation to estimate missing legal
- We then estimate Y from estimated values of X
- For some distributions iteration may be necessary

Expectation Maximization

- Expectation

$$E(x/y) = \mu_x + \frac{\sigma_{xy}}{\sigma_x}(y - \mu_y)$$

- Maximization
  - Using estimated value of legal representation maximize joint liklihood of x and y

Expectation Maximization Assuming Multivariate Normal

- Covariance
- X1 X2
- X1 0.25147 0.67024
- X2 0.67024 7.09093
- ==========

- final log-likelihood = -1649.7

- difference in the log-likelihood (or log posterior density) = 0.0025512

- maximum absolute relative change in parameter estimate on last iteration = 0.00034302

- call:
- emGauss.default(object = Miss)

Beginners Library

- Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997
- Silverman, B.W. Density Estimation for Statistics and Data Analysis, Chapman and Hall
- Smith, Murry, *Neural Networks for Statistical Modeling*, International Thompson Computer Press, 1996
- Fox, John, *An R and S-PLUS Companion to Applied Regression*, Sage Publications, 2002
- Francis, Louise, ”Martian Chronicles: Is MARS Better than Neural Networks”, 2003 CAS Spring Forum

Francis Analytics and Actuarial
Data Mining, Inc.