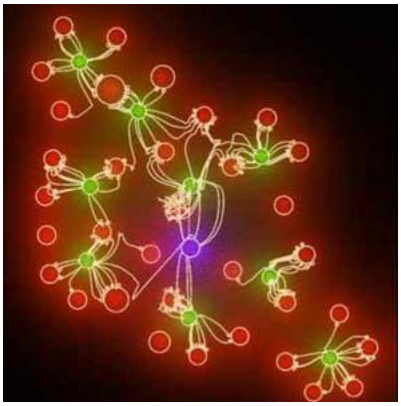
SPSS

SPSS PLS Regression & Neural Networks



Model of a hierarchical neural network in action
Image: Kurt W. Fleischer/Caltech Graphics

허 명 희, 고려대학교 통계학과 교수
E-mail: stat420@korea.ac.kr

2008.6.4
SPSS Open House

1

SPSS

Today's Lecture

PLS Regression

SPSS 16: Analyze > Regression > Partial Least Squares…

Neural Networks

SPSS 16: Analyze > Neural Networks

- Multilayer Perceptron …(MLP)

- Radial Basis Function …(RBF)

SPSS

PLS Regression의 기본특성

PLS Regression

PLS: Partial Least Squares (PLS) 편최소제곱
- Projection to Latent Structures 잠재구조사영

주요특성:

1) 다중공선성(multicollinearity)에 취약하지 않음, 이 점에서 선형회귀와 대조됨.
2) 설명변수 수가 관측 케이스 수보다 많은 경우에도 적용 가능.
3) 계량경제(econometrics), 계량화학(chemometrics) 등 여러 분야에서 활용됨.

SPSS

PLS Regression 사례

Boston Housing Data

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley, 1980.

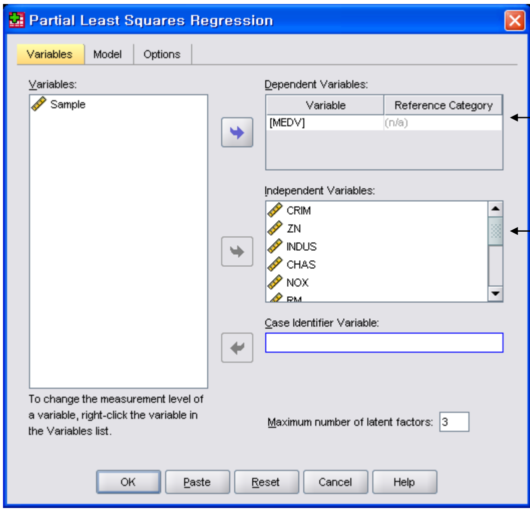
Variables:

CRIM	**per capita crime rate by town**
ZN	**proportion of residential land zoned for lots over 25,000 sq.ft.**
INDUS	**proportion of non-retail business acres per town**
CHAS	**Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)**
NOX	**nitric oxides concentration (parts per 10 million)**
RM	**average number of rooms per dwelling**
AGE	**proportion of owner-occupied units built prior to 1940**
DIS	**weighted distances to five Boston employment centres**
RAD	**index of accessibility to radial highways**
TAX	**full-value property-tax rate per $10,000**
PTRATIO	**pupil-teacher ratio by town**
B	**1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town**
LSTAT	**% lower status of the population**
MEDV	**Median value of owner-occupied homes in $1000's**

SPSS

PLS Regression 사례

Boston Housing Data



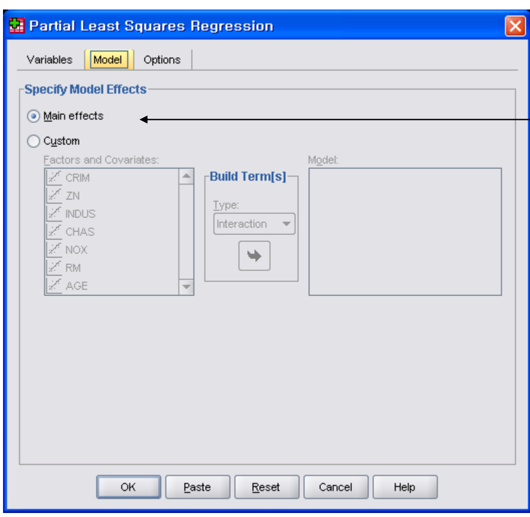
종속변수 1개

독립변수 13개

SPSS

PLS Regression 사례

Boston Housing Data

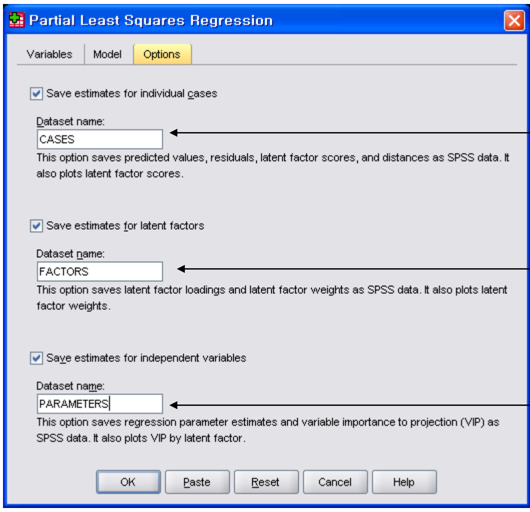


모형 (주효과)

SPSS

PLS Regression 사례

Boston Housing Data



예측값, 잔차 등

요인부하, 가중치 등

회귀 계수 등

SPSS

PLS Regression 사례

Boston Housing Data

Output:

Proportion of Variance Explained

Latent Factors	Statistics				
	X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
1	.459	.459	.499	.499	.497
2	.109	.568	.207	.706	.705
3	.071	.639	.017	.723	.721

3개의 잠재요인으로 X 변수들의 63.9%를 설명

3개의 잠재요인으로 Y 변수의 72.3%를 설명

SPSS

PLS Regression 사례

Boston Housing Data

Output: 회귀계수 (parameters)

Parameters

Independent Variables	Depende··· MEDV
(Constant)	14.322
CRIM	-.061
ZN	.019
INDUS	-.066
CHAS	2.319
NOX	-6.045
RM	5.080
AGE	-.008
DIS	-.816
RAD	.087
TAX	-.001
PTRATIO	-.757
B	.010
LSTAT	-.485

SPSS Korea Open House 9

SPSS

PLS Regression 사례

Boston Housing Data

Output: 변수 중요도 (variable importance)

Variable Importance in the Projection

Variables	Latent Factors 1	2	3
CRIM	.853	.718	.710
ZN	.792	.678	.670
INDUS	1.063	.914	.903
CHAS	.385	.627	.673
NOX	.939	.850	.866
RM	1.527	1.814	1.793
AGE	.828	.779	.775
DIS	.549	.965	.979
RAD	.838	.823	.847
TAX	1.029	.891	.883
PTRATIO	1.115	1.100	1.093
B	.732	.619	.620
LSTAT	1.620	1.606	1.606

Cumulative Variable Importance

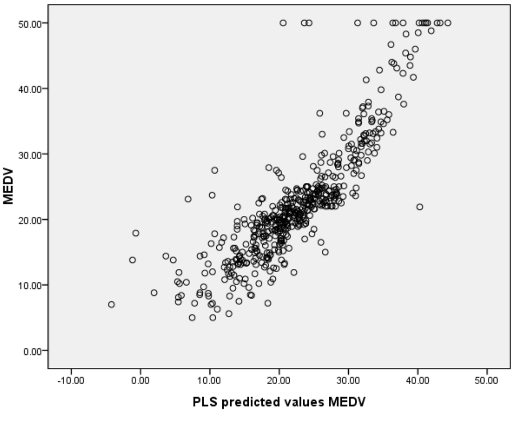
SPSS Korea Open House 10

SPSS

PLS Regression 사례

Boston Housing Data

Graph: (종속변수) 예측값과 실제값



* corr=0.85

SPSS

PLS Regression 방법론

표기 y : 종속변수, n개 케이스

X : p개 독립변수, n개 케이스

모형 $y = Xb + f$

여기서 b: regression coefficient

f: error

SPSS

PLS Regression 방법론

보통최소제곱 회귀 (Ordinary Least Squares Regression)

$$\min_b \sum_{i=1}^{n} (y_i - x_i^t b)^2$$

여기서 x_i^t는 X의 i번째 행 ($i = 1, \cdots, n$).

편제곱 회귀 (PLS Regression)

$$\max_b Cov(y, Xb)$$ 단, $b^t b = 1$.

SPSS

PLS Regression 방법론

잠재인자 t_k와 가중치 w_k(Latent Factor and Weights)

$$t_k \leftarrow X_{[k-1]} w_k$$

여기서 t_k는 $n \times 1$, w_k는 $p \times 1$, $k = 1, \cdots, d$.

잠재인자 부하(Latent Factor Loadings)

$$x_j = t_1 p_{j1} + \cdots + t_d p_{jd}$$

여기서 $p_j^t = (p_{j1}, \cdots, p_{jd})$는 j번째 변수에 대한 부하.

$$j = 1, \cdots, p$$

SPSS

PLS Regression 기타사항

측정수준: 어느 변수에 대하여도 측정수준으로 scale, ordinal, nominal이 모두 허용된다.

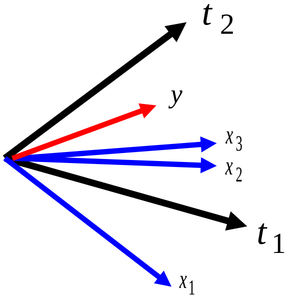
범주형 변수의 가변수 코딩: 범주형(ordinal, nominal) 변수는 모두 0-1 변수로 코딩 된다.
범주 수가 c개이면 c개의 0-1(dummy) 변수 $D_1,\cdots,D_c$가 생성.

표준화: 모든 변수는 PLS 적용에 앞서 평균 0, 표준편차 1이 되도록 …
범주형 변수의 표현을 위해 생성된 가변수에도 적용된다.

SPSS

PLS Regression 기하

기하: n 차원 공간



참고문헌

SPSS 16.0 Algorithms [PLS Algorithms, pp.582-588]
SPSS Neural Networks and PLS Regression 한글매뉴얼

SPSS

PLS Regression 응용

응용 사례: 근적외선 분광기 자료

Near-Infrared (NIR) Spectroscopy

- 파일 명: yarn_2.sav
- y = 폴리에틸렌 수지 밀도
- X = 268개 스펙트럼에서의 빛 흡수도 (p=228+ 1)
- 28개 관측(=n)

- 모형 $y = b_0 + b_1 x_1 + \cdots + b_{268} x_{268} + f$

SPSS Analyze > Regression > Partial Least Squares

SPSS

PLS Regression 응용

Proportion of Variance Explained

	Statistics				
Latent Factors	X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
1	.498	.498	.922	.922	.916
2	.202	.700	.072	.994	.993

2개의 잠재요인으로 설명된 비율

Parameters

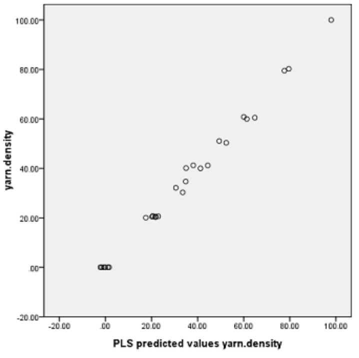
Independent Variables	Depende··· yarn.density
(Constant)	47.604
X1	-.094
X2	.355
X3	.335
X4	.235
X5	.113
X266	-3.797
X267	-3.815
X268	-3.771

적합모형의 회귀계수 (파라미터)

SPSS

PLS Regression 응용

y: 실제값과 예측값



SPSS

PLS Regression 설치

SPSS 16.0에서 PLS 회귀를 하려면 …

- Python2.5
- SPSSPythonIntegrationPlugIn
- SPSSPLSExtensionModule

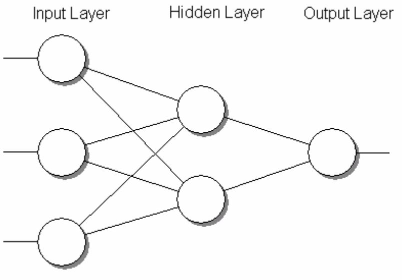
http://www.spss.com/devcentral 참조

SPSS

Neural Networks의 기본구조

신경망(神經網, neural networks):

- 몇 개의 뉴런(neuron, 신경세포)과 이것들이 배열된 층(layer)으로 구성됨
- 특정의 작업을 수행하는 뉴런들을 연결하여 자극과 반응간의 관계를 학습
- 입력 층(input layer) → 은닉 층(hidden layer) → 출력 층(output layer)



SPSS

Neural Networks : MLP

다층 퍼셉트론 (Multi-Layer Perceptron, MLP) 구조

- 입력층 뉴런(입력변수)로부터 전달되는 신호들을 모아 선형결합한다.
- $X_1, \cdots, X_p$를 설명변수(입력노드)라고 할 때 다음 뉴런에

$$L = w_1 X_1 + \cdots + w_p X_p$$

이 전달된다. 여기서 $w_1, \cdots, w_p$ 는 신경선(synapse)에 붙는 가중값(weight)

- 뉴런의 활성화
  . 로지스틱(logistic): $S = e^L/(1+e^L), \quad 0 \le S \le 1$
  . 쌍곡 탄젠트(hyperbolic tangent): $S = (e^L - e^{-L})/(e^L + e^{-L}), \quad -1 \le S \le 1$
- 출력노드
  . 연속형 : $O = L$
  . 범주형: 소프트맥스(softmax): $O_k = \dfrac{\exp(L_k)}{\sum_{j=1}^{K} \exp(L_j)}$ 여기서 k 는 범주.

SPSS

Neural Networks : RBF

방사형 기저함수(Radial Basis Fuctionstions, RBF) 구조

- 입력층 뉴런(입력변수)로부터 전달되는 신호들을 모아 중심신호와의 거리에 역비례하는 강도로 변환

$$R = \exp[-\frac{1}{2\sigma^2}\{(X_1-\mu_1)^2+\cdots+(X_p-\mu_p)^2\}]$$

여기서 $\mu_1,\cdots,\mu_p$ 는 각 신호의 중심

- 출력노드: $L = w_1 R_1 + \cdots + w_J R_J$
  - . 연속형 : $O = L$
  - . 범주형: 소프트맥스(softmax): $O_k = \frac{\exp(L_k)}{\sum_{j=1}^{K}\exp(L_j)}$

SPSS

Neural Networks의 적합과 평가

역전파 알고리즘(back-propagation algorithm):

- 학습률 (eta) η : 가장 큰 경사에 주는 강도
- 모멘트(moment, alpha) α : 이제까지 이동하던 방향에 주는 강도

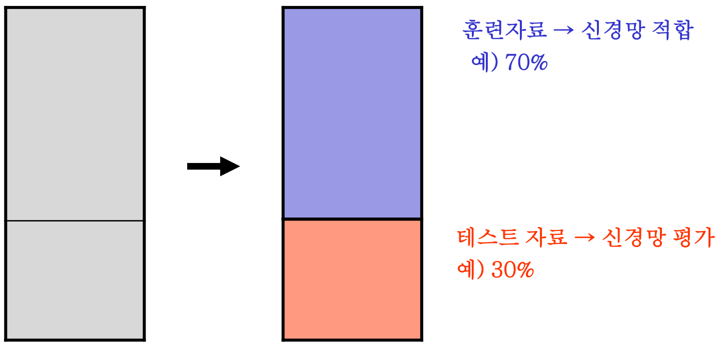
임의의 위치에서 시작하되
처음에는 큰 학습률을 적용하여 적극적으로 최고점을 탐색하지만
점차 학습률을 줄여 최고점이라고 생각되는 곳에 도달한다.
다시 다른 임의의 위치에서 시작하되 동일한 과정을 반복한다.
이런 과정을 수십 번 반복하여 가장 높은 최고점을 취한다.

최대화 과정은 눈가림 상태로 "등산"해야 하는 상황과 유사하다.
주어진 위치에서 최대 경사가 되는 방향과 이제까지 이동하던 방향을 절충하되,
도달 위치가 높아질수록 전자의 비중을 줄여 안정성을 도모한다.

SPSS

Neural Networks를 위한 자료 분할

훈련자료(training data), 테스트 자료(test data)



※ 훈련자료 50%, 테스트 자료 25%, 유보자료(hold-out) 25%

SPSS

MLP 사례

German Credit Data

- 은행에서 대출을 받은 1000명에 대한 신용관련 변수와 기록
- 종속변수 class=1 (good), 2 (bad)
- 설명변수: 총 20개 (연속형 7개, 범주형 13개)

	Name	Type	Width	Decimals
1	checking	String	4	0
2	duration	Numeric	4	1
3	history	String	4	0
4	purpose	String	4	0
5	amount	Numeric	4	2
6	savings	String	4	0
7	employ	String	4	0
8	install	Numeric	4	0
9	marital	String	4	0
10	debt	String	4	0

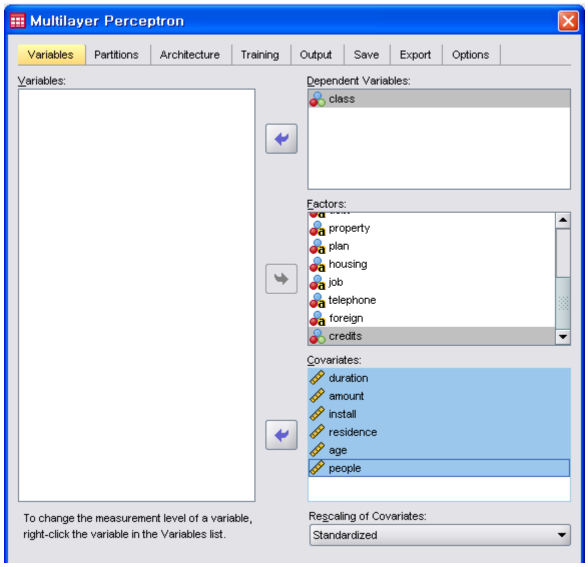
↓

21 class

SPSS

MLP 사례

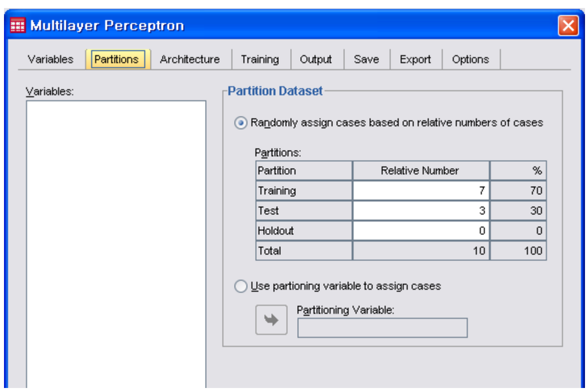
MLP Variables
German Credit



SPSS

MLP 사례

MLP Partitions
German Credit

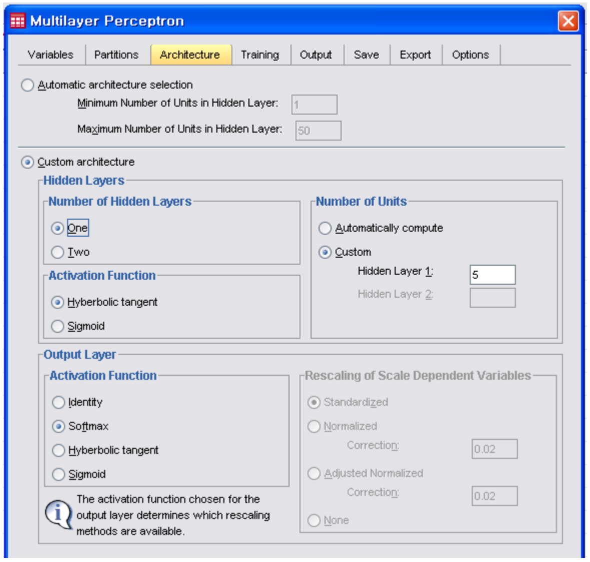


SPSS

MLP 사례

MLP Architecture

German Credit

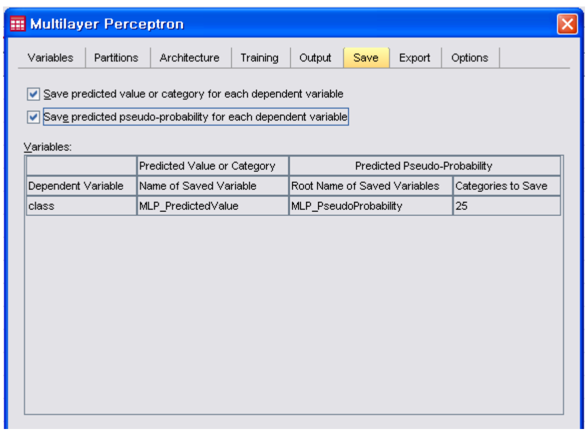


SPSS

MLP 사례

MLP Save

German Credit



SPSS

MLP 사례

OUTPUT: Neural Networks for German Credit Data

Classification

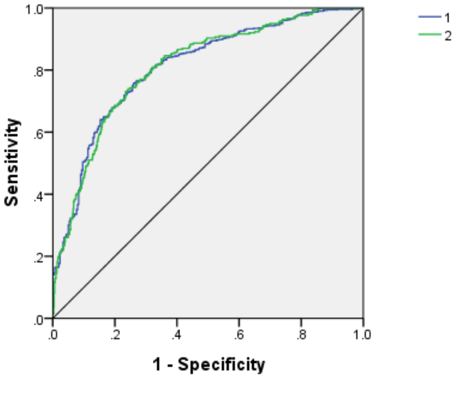
Sample	Observed	Predicted 1	Predicted 2	Predicted Percent Correct
Training	1	419	62	87.1%
	2	88	110	55.6%
	Overall Percent	74.7%	25.3%	77.9%
Testing	1	181	38	82.6%
	2	50	52	51.0%
	Overall Percent	72.0%	28.0%	72.6%

Dependent Variable: class

SPSS

MLP 사례

OUTPUT: Neural Networks for German Credit Data



Dependent Variable: class

Area Under the Curve

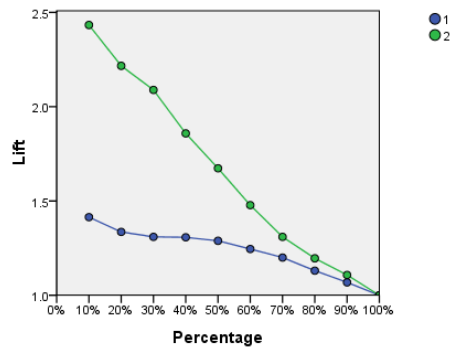
		Area
class	1	.794
	2	.794

※ Specificity (특이도) = Probability of correctly identifying a negative
Sensitivity (민감도) = Probability of correctly identifying a positive

SPSS

MLP 사례

OUTPUT: Neural Networks for German Credit Data



Dependent Variable: class

향상도(lift):

예측모형이 제공하는 점수(유사확률)를 기준으로 전체(=훈련+테스트) 자료의 상위 p%에 범주 값을 판단해 넣는 경우 그 중에서 실제 그 범주인 케이스들 가운데 옳게 분류된 케이스의 퍼센트(%)가 p의 몇 배인가?

SPSS

MLP 사례

OUTPUT: Neural Networks for German Credit Data

*SPSS DataSet

	class	MLP_PredictedValue	MLP_PseudoProbability_1	MLP_PseudoProbability_2
1	1	1	0.978	0.022
2	2	2	0.258	0.742
3	1	1	0.965	0.035
4	1	1	0.536	0.464
5	2	2	0.341	0.659
6	1	1	0.828	0.172
7	1	1	0.960	0.040
8	1	1	0.599	0.401
9	1	1	0.970	0.030
10	2	2	0.246	0.754
11	2	2	0.222	0.778
12	2	2	0.129	0.871
13	1	1	0.876	0.124
14	2	1	0.572	0.428
15	1	2	0.292	0.708

New Variables:

1) MLP_PredictedValue = 예측범주
2) MLP_PseudoProbability_1 = 범주 1 예측 확률
3) MLP_PseudoProbability_2 = 범주 2 예측 확률

SPSS

RBF 사례

Telco Data

- 전화회사, 1000명에 대한 인구사회적 변수와 통신 서비스 정보
- 종속변수 custcat=1, 2, 3, 4

  * 1 = “Basic service”, 2 = “E-service”, 3 = “Plus service”, 4 = “Total service”
- 변수 수: 총 42개

telco.sav [SPSS16\Samples\]

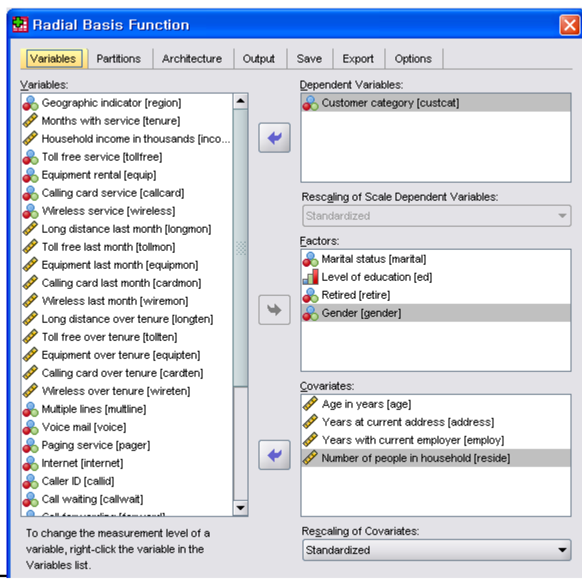
	region	tenure	age	marital	address	income	ed	employ
1	2	13	44	1	9	64.00	4	5
2	3	11	33	1	7	136.00	5	5
3	3	68	52	1	24	116.00	1	29
4	2	33	33	0	12	33.00	2	0
5	2	23	30	1	9	30.00	1	2
6	2	41	39	0	17	78.00	2	16
7	3	45	22	1	2	19.00	2	4
8	2	38	35	0	5	76.00	2	10

SPSS

RBF 사례

RBF Variables

Telco 데이터

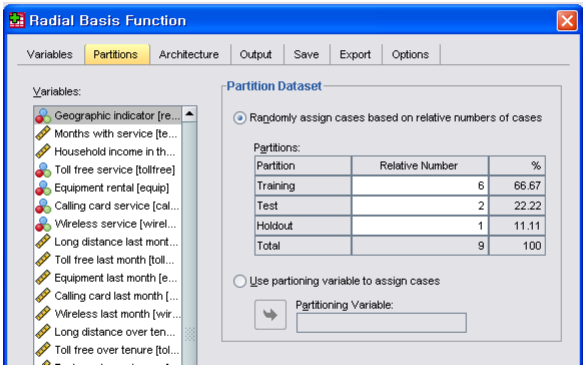


SPSS

RBF 사례

RBF Partitions

Telco 데이터

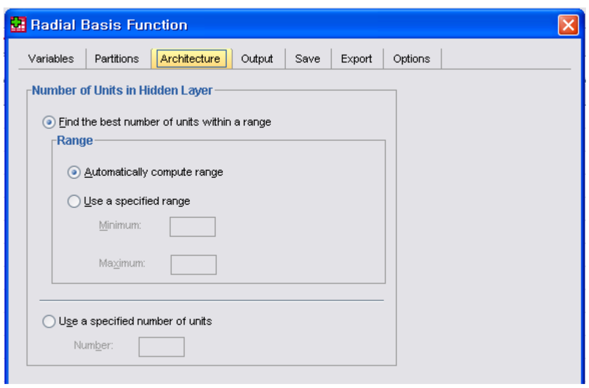


SPSS

RBF 사례

RBF Architecture

Telco 데이터



SPSS

RBF 사례

OUTPUT: Neural Networks for Telco Data

Classification

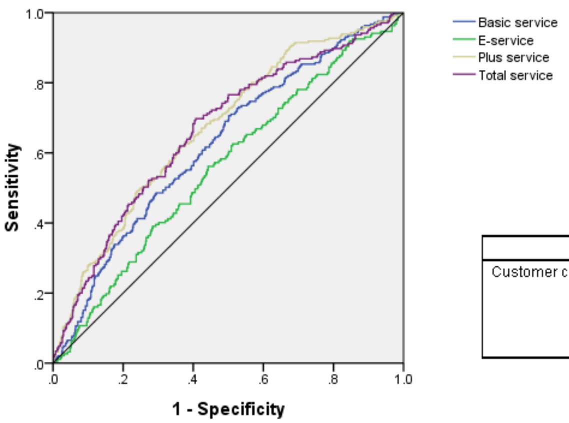
Sample	Observed	Predicted: Basic service	Predicted: E-service	Predicted: Plus service	Predicted: Total service	Predicted: Percent Correct
Training	Basic service	83	0	72	30	44.9%
	E-service	39	0	65	34	.0%
	Plus service	41	0	122	26	64.6%
	Total service	44	0	53	62	39.0%
	Overall Percent	30.8%	.0%	46.5%	22.7%	39.8%
Testing	Basic service	19	0	34	7	31.7%
	E-service	14	0	23	12	.0%
	Plus service	9	0	34	14	59.6%
	Total service	11	0	17	18	39.1%
	Overall Percent	25.0%	.0%	50.9%	24.1%	33.5%
Holdout	Basic service	6	0	8	7	28.6%
	E-service	7	0	12	11	.0%
	Plus service	11	0	17	7	48.6%
	Total service	12	0	9	10	32.3%
	Overall Percent	30.8%	.0%	39.3%	29.9%	28.2%

Dependent Variable: Customer category

SPSS

RBF 사례

OUTPUT: Neural Networks for Telco Data



Dependent Variable: Customer category

Area Under the Curve

		Area
Customer category	Basic service	.625
	E-service	.560
	Plus service	.670
	Total service	.665

SPSS

References for Neural Networks

참고문헌

SPSS 16.0 Algorithms [MLP Algorithms, pp.470-479, RBF Algorithms, pp.647-652]
SPSS Neural Networks 16.0
SPSS Neural Networks and PLS Regression 한글매뉴얼

데이터 마이닝 모델링과 사례 **[제2판]** 허명회/이용구, 한나래 **2008.**

SPSS

Questions and Answers