

Multiclass ANN/SVM 실습 보고서

Y2023011 윤요섭

I. 자료 및 분석 방법

본 보고서는 CustomerLevel.xlsx 데이터를 이용하여 Neuroshell2® R4.2와 LibSVM을 통해 Multiclass ANN/SVM 실습과제를 수행하였다. CustomerLevel.xlsx 데이터 변수와 이에 대한 설명은 다음과 같다.

ID	고객ID
Churn	고객 이탈여부
TotCharge	납부요금 (총통화요금+기본요금)
YearlyBasis	연간납부요금
Level	고객 등급
Gender	성별
Age	나이
Length	휴대전화 서비스 사용기간
Snap	단선허수
BillProg	요금제 유형
BP1	CAT50
BP2	CAT100
BP3	CAT200
BP4	PLAY100
BP5	PLAY300
Handset	단말기 유형
HS1	ASAD
HS2	BS
HS3	CAS
HS4	S
HS5	SOP
HS6	WC
CallTimeDay	주간통화횟수
CallLengthDay	주간통화시간
CallTimeNight	야간통화횟수
CallLengthNight	야간통화시간
CallTimeWeekend	주말통화횟수
CallLengthWeekend	주말통화시간
CallLengthInternational	국제통화시간_분
AvgCallLengthDay	평균주간통화시간
AvgCallLengthNight	평균야간통화시간
AvgCallLengthWeekend	평균주말통화시간
CallTimeDomestic	국내통화횟수
CallLengthDomestic	국내통화시간_분
AvgCallLengthDomestic	평균국내통화시간
TotCallLength	총통화시간_분
CallLengthPay	요금부과시간

ChargeMin	분당통화요금
ChargeDomestic	국내통화요금
ChargeTotal	총통화요금
BillProgValid	기본요금대비 납부요금이 큰지 여부
AvgChargePerUsage	평균납부요금(=부과요금/총통화시간)
PercentDay	주간통화비율
PercentNight	야간통화비율
PercentWeekend	주말통화비율
PercentInternational	국제통화비율(=국제통화시간/국내통화시간)

46개의 변수와 별개로, set1은 훈련용/테스트 데이터, set2는 훈련용/검증용/테스트 데이터 컬럼으로 설정하였고, 46개의 변수 중 종속변수는 Level(고객 등급)으로 설정하였다. 나머지 45개의 변수 중 최종 독립변수는 카이제곱 분석 등과 MDA의 단계별 선택 방법을 통해 선정하였다.

먼저, 13개의 이산형 독립변수와 종속변수(Level)의 카이제곱 검정 결과를 통해, 95% 신뢰 수준 하에서 유의확률이 0.05이하인 독립변수 11개(BP1, BP2, BP3, BP4, BP5, HS1, HS2, HS3, HS4, HS6, BillProgValid)를 1차적으로 독립변수로 선정하였다.

나머지, 26 개의 연속형 독립변수와 종속변수(Level)의 일원배치 분산분석을 통해, 95% 신뢰 수준에서 유의확률이 0.05 이하인 독립변수 18 개(Length, CallTimeDay, CallLengthDay, CallTimeNight, CallLengthWeekend, CallLengthInternational, CallTimeDomestic, CallLengthDomestic, AvgCallLengthDomestic, TotCallLength, CallLengthPay, ChargeMin, ChargeDomestic, ChargeTotal, AvgChargePerUsage, PercentDay, PercentNight, PercentWeekend)를 1 차적 독립변수로 선정하였다.

이렇게 총 39 개의 1 차 독립변수를 선정한 뒤에, 2 차 독립변수를 MDA 의 단계별 선택 방법을 통해 선정하였다.

MDA 의 단계적 선택 방법을 통해 변수를 선정하기 위해 판별분석을 시행한 결과, 13 개의 변수(TotCallLength, BP4, CallLengthDomestic, BP3, BP2, ChargeDomestic, BillProgValid, HS1, HS4, BP1, HS6, HS3, CalTlmeDay)가 단계선택 통계량에서 진입된 변수로 선택되었다.

분류 결과 학습 데이터에서는 83.8%가 올바른 분류가 되었으며, 테스트 데이터에서는 85.4%가 올바른 분류가 되었다.

분류결과^{a,b}

			Level	예측 소속집단				전체
				1	2	3	4	
선택된 케이스	원래값	빈도	1	714	86	0	0	800
			2	0	705	95	0	800
			3	0	138	586	76	800
			4	0	0	124	676	800
		%	1	89.3	10.8	.0	.0	100.0
			2	.0	88.1	11.9	.0	100.0
			3	.0	17.3	73.3	9.5	100.0
			4	.0	.0	15.5	84.5	100.0
선택되지 않은 케이스	원래값	빈도	1	180	20	0	0	200
			2	0	179	21	0	200
			3	0	29	152	19	200
			4	0	0	28	172	200
		%	1	90.0	10.0	.0	.0	100.0
			2	.0	89.5	10.5	.0	100.0
			3	.0	14.5	76.0	9.5	100.0
			4	.0	.0	14.0	86.0	100.0

a. 원래의 선택 집단 케이스 중 83.8%이(가) 올바르게 분류되었습니다.

b. 원래의 비선택 집단 케이스 중 85.4%이(가) 올바르게 분류되었습니다.

13개의 독립변수를 최종 독립변수로 지정한 뒤, Neuroshell2® R4.2과 LibSVM을 활용하여 통신회사의 고객등급 예측을 위한 모델을 도출하였다.

II. 분석 과정과 결과 및 해석

II-가 Multiclass ANN을 통한 분석

Test Set Extract



Files Extracted

Number of rows in training set (.trn) = 2400

Number of rows in test set (.tst) = 800

Number of rows in production file (.pro) = 800

Pattern (.pat) file was not altered.

확인


ANN을 통한 분석에서는 학습용 데이터는 2400건, 검증용 데이터와 테스트용 데이터는 각각 800


건으로 설정하고 실습을 진행하였다.


Learning: C:\W...WBITW1-1학기W비즈니스애널리틱스WwkWwk08W과제06 수행용W... — □ ×


File Run Options Help

Training Graphics

Training Set Average Error  **Epochs Elapsed**

Test Set Average Error  **Intervals Elapsed**

Error Factor Ranges  **Training Set Patterns**

Error Factor Ranges  **Test Set Patterns**

There are 2400 training patterns.

☐ learning events: 258000

☒ learning epochs: 107

☒ last average error: 0.0056025

☒ minimum average error: 0.0053880

☐ epochs since min. avg. error: 7

There are 800 test patterns.

Calibration interval (events): 200

☐ last average (internal) error: 0.0058191

☐ minimum average error: 0.0049807

☐ events since min. avg. error: 20000

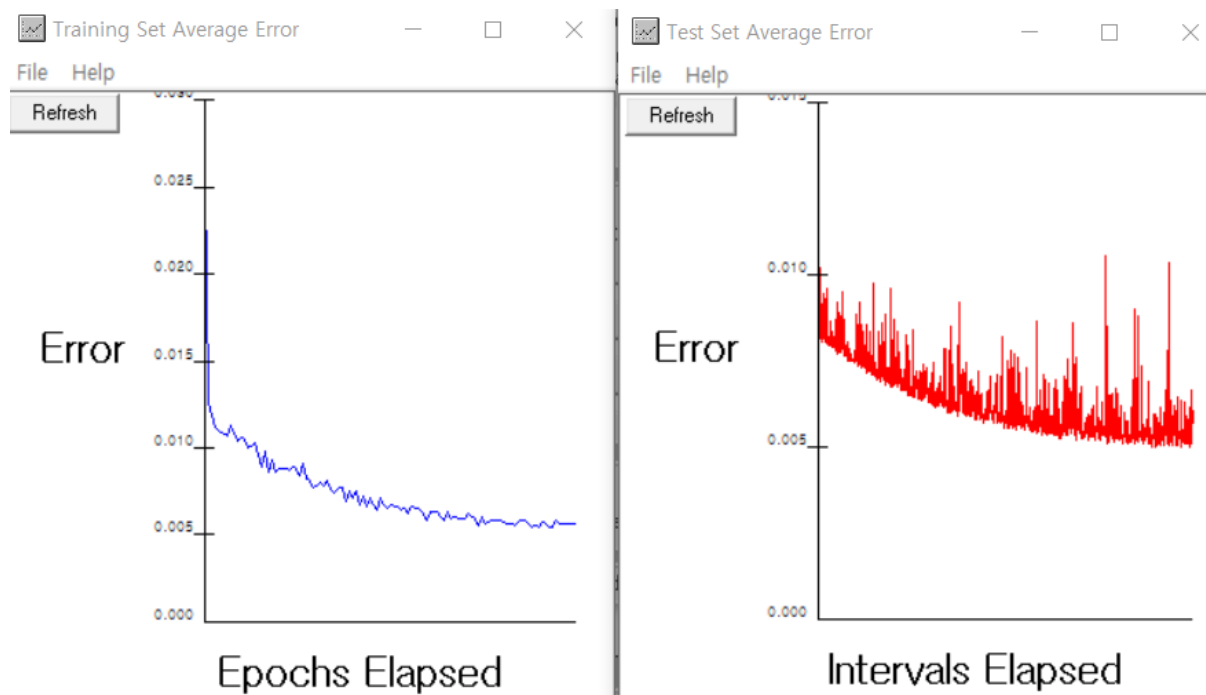
Automatically Save Training on

☐ best training set ☒ best test set ☐ no auto save

Training Time (hhh:mm:ss)

000:00:08

Check boxes above to display selected statistics. Training is slowed with more graphs/statistics. Apply net to see true error.



실습을 진행한 이후, Neuroshell의 결과를 View Output File을 통해 Network(1) 열을 복사한 뒤, 소수점 첫 째자리에서 반올림 하여 실제 값과 비교한 결과는 아래 사진과 같다.

set2	Accuracy
t	0.935833
p	0.94375
v	0.93875

학습, 검증, 테스트 데이터에서 모두 93% 이상의 정답을 맞춘 것을 통해, 학습 및 검증이 잘 되었다고 판단할 수 있다.

II-나 Multiclass SVM을 통한 분석

LibSVM는 가장 무난한 성능을 보이는 OAO(One-Against-One) 방법을 이용하여 다분류 SVM을 구현한다. OAO 모델의 경우, K개의 클래스로 분류해야 하는 문제인 경우 $\binom{K}{2}$ 개의 이분류기를 구축하여 다수결의 원칙에 따라 가장 많이 선택된 Class를 최종 선택하는 기법이다.

종속변수 Level이 총 1부터 4까지 4개(1:VIP, 2:Gold, 3:Silver, 4:Normal)로 이루어져있으므로, 이럴 경우, $\binom{4}{2}$ 가 되어, 모델을 총 6개 학습하여야 한다.

실습에서 사용할 파라미터를 정하기 위해 각 커널과 파라미터에 변화에 대한 학습데이터와 테스트(검증)데이터의 결과값을 추출 한 뒤, 상위 10개를 뽑았을 때, 결과는 아래와 같다.

구분1(커널)	구분2[C]	구분3[D]	Training	Validation
다항식커널	C=78	4	99.09%	97.63%
다항식커널	C=55	4	98.75%	97.50%
다항식커널	C=100	5	99.44%	97.38%
다항식커널	C=78	5	99.41%	97.25%
다항식커널	C=55	3	98.09%	97.13%
다항식커널	C=100	3	98.38%	97.13%
다항식커널	C=100	4	99.16%	97.13%
다항식커널	C=33	4	98.47%	97.00%
다항식커널	C=55	5	99.31%	97.00%
다항식커널	C=78	3	98.38%	96.88%

그 중에서 가장 결과가 좋았던 Validation의 모델 파라미터 설정 값은 다항식 커널에 파라미터 C는 78, gamma는 1(default), coef()는 1, degree는 4였다.

아래 사진을 통해 모델 설정값 명령어, 모델 실행 시, 모델을 총 6개 학습하는 과정, 예측 및 정확도 측정 과정을 볼 수 있다.

```
C:\Temp\libsvm>svm-train -t 1 -c 78 -g 1 -r 1 -d 4 svm_t
...*...*
optimization finished, #iter = 8173
nu = 0.075061
obj = -7091.076827, rho = -1.102403
nSV = 153, nBSV = 95
*
optimization finished, #iter = 709
nu = 0.001875
obj = -117.022914, rho = -2.133217
nSV = 24, nBSV = 1
*
optimization finished, #iter = 86
nu = 0.000040
obj = -2.495609, rho = -1.872618
nSV = 10, nBSV = 0
...*...*
optimization finished, #iter = 21368
nu = 0.045863
obj = -3996.294945, rho = -5.355064
nSV = 106, nBSV = 51
*
optimization finished, #iter = 788
nu = 0.000131
obj = -8.146489, rho = -4.863175
nSV = 21, nBSV = 0
...*...*
optimization finished, #iter = 47223
nu = 0.024839
obj = -2195.768889, rho = -13.440972
nSV = 57, nBSV = 32
Total nSV = 337

C:\Temp\libsvm>svm-predict svm_t svm_t.model o1c78d4t
Accuracy = 99.0938% (3171/3200) (classification)

C:\Temp\libsvm>svm-predict svm_v svm_t.model o1c78d4v
Accuracy = 97.625% (781/800) (classification)
```

이를 통해, MDA부터 MANN/MSVM까지 실습을 진행하였고, MSVM의 경우가 가장 높은 정확도가 나타난 것을 확인할 수 있다.

-끝-