|  |
| --- |
| REPORT |
| **과제 6**  Multiclass ANN/SVM |

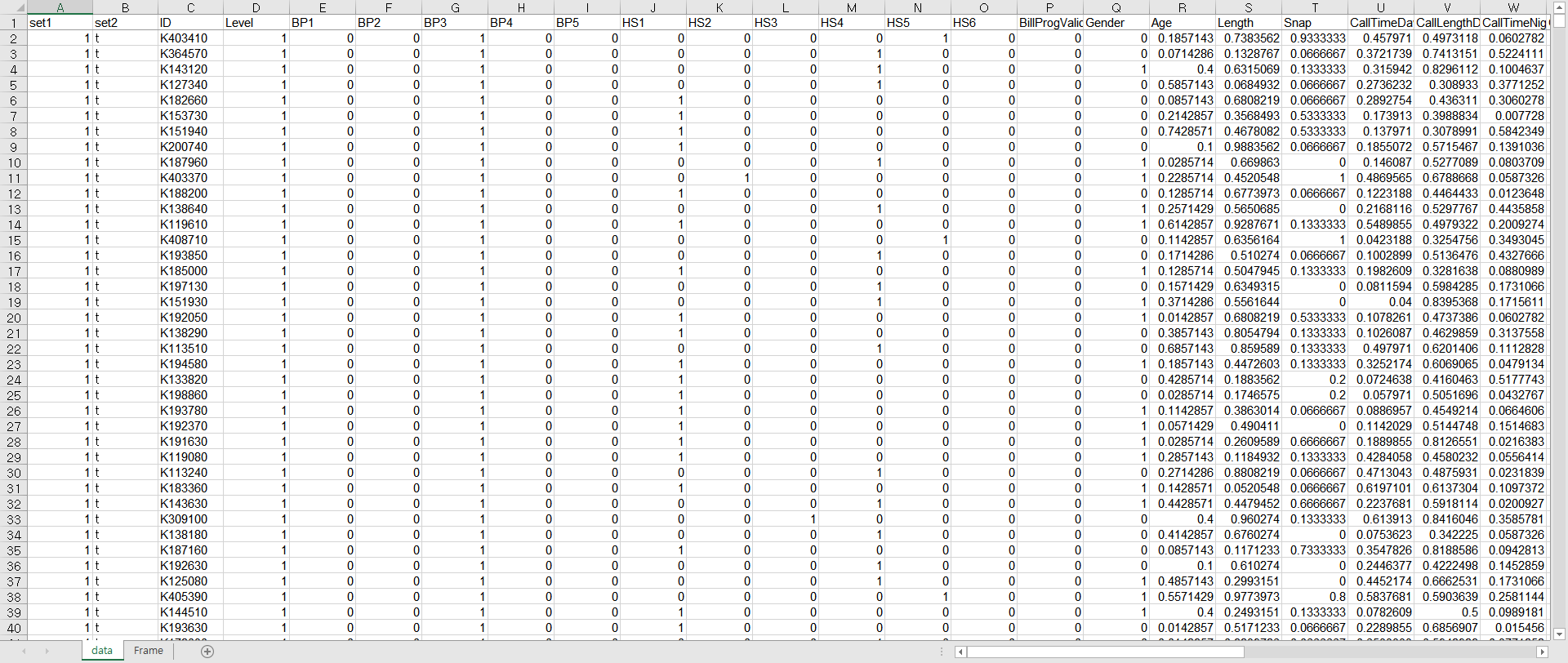




|  |  |  |
| --- | --- | --- |
| 제출일 | 2020. 05. 09 |  |
| 과목명 | 비즈니스애널리틱스개론 |  |
| 담당교수 | 안현철 |  |
| 전공 | 경영정보전공 |  |
| 학번 | 20175288 |  |
| 이름 | 최광삼 |  |

**목차**

1. **유의변수 추출**
2. 일원배치 분산검정
3. 카이제곱검정
4. 다중판별분석
5. 최종 유의변수
6. **ANN**
7. 모델 설정
8. 예측 정확도
9. **SVM**
10. LIBSVM 사용을 위한 전처리
11. 최적의 조합 및 결과
12. **유의변수 추출**



CustomerLevel.xls 파일은 통신회사의 고객의 사용량 데이터를 고객등급별로 나눈 것이다. 위 이미지는 해당 파일을 캡쳐한 것이다. 이 데이터를 토대로 고객등급을 예측하기 위해서는 먼저 예측의 정확도와 신뢰성을 확보하기 위하여 일원배치분산검정 및 카이제곱검정을 통해 예측분석을 하였을 때 유의미한 변수를 추출해야 할 필요가 있다.

검정을 실시할 때는 고객등급을 기준으로 하여 진행하였다. 연속형 변수의 경우에는 4개의 고객등급 집단에 대해 비교하여야 하기 때문에 일원배치 분산검정을 적용하였으며, 이분형 변수에는 카이제곱검정을 적용하였다.

* 1. 일원배치 분산검정

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **분산의 동질성 검정** | | | | | |
|  | | Levene 통계량 | 자유도1 | 자유도2 | 유의확률 |
| Age | 평균을 기준으로 합니다. | .010 | 3 | 3996 | .999 |
| Length | 평균을 기준으로 합니다. | 2.066 | 3 | 3996 | .103 |
| Snap | 평균을 기준으로 합니다. | 2.605 | 3 | 3996 | .050 |
| CallTimeDay | 평균을 기준으로 합니다. | 538.628 | 3 | 3996 | .000 |
| CallLengthDay | 평균을 기준으로 합니다. | 301.610 | 3 | 3996 | .000 |
| CallTimeNight | 평균을 기준으로 합니다. | 24.586 | 3 | 3996 | .000 |
| CallLengthNight | 평균을 기준으로 합니다. | 10.161 | 3 | 3996 | .000 |
| CallTimeWeekend | 평균을 기준으로 합니다. | .709 | 3 | 3996 | .546 |
| CallLengthWeekend | 평균을 기준으로 합니다. | .575 | 3 | 3996 | .631 |
| CallLengthInternational | 평균을 기준으로 합니다. | 253.716 | 3 | 3996 | .000 |
| AvgCallLengthDay | 평균을 기준으로 합니다. | 5.589 | 3 | 3996 | .001 |
| AvgCallLengthNight | 평균을 기준으로 합니다. | .707 | 3 | 3996 | .548 |
| AvgCallLengthWeekend | 평균을 기준으로 합니다. | 1.616 | 3 | 3996 | .184 |
| CallTimeDomestic | 평균을 기준으로 합니다. | 347.687 | 3 | 3996 | .000 |
| CallLengthDomestic | 평균을 기준으로 합니다. | 86.348 | 3 | 3996 | .000 |
| AvgCallLengthDomestic | 평균을 기준으로 합니다. | 9.628 | 3 | 3996 | .000 |
| TotCallLength | 평균을 기준으로 합니다. | 91.337 | 3 | 3996 | .000 |
| CallLengthPay | 평균을 기준으로 합니다. | 31.169 | 3 | 3996 | .000 |
| ChargeMin | 평균을 기준으로 합니다. | 392.143 | 3 | 3996 | .000 |
| ChargeDomestic | 평균을 기준으로 합니다. | 698.387 | 3 | 3996 | .000 |
| ChargeTotal | 평균을 기준으로 합니다. | 358.429 | 3 | 3996 | .000 |
| AvgChargePerUsage | 평균을 기준으로 합니다. | 16.980 | 3 | 3996 | .000 |
| PercentDay | 평균을 기준으로 합니다. | 126.724 | 3 | 3996 | .000 |
| PercentNight | 평균을 기준으로 합니다. | 167.598 | 3 | 3996 | .000 |
| PercentWeekend | 평균을 기준으로 합니다. | 493.573 | 3 | 3996 | .000 |
| PercentInternational | 평균을 기준으로 합니다. | 29.694 | 3 | 3996 | .000 |

일원배치 분산검정 실시 결과, 분산의 동질성 검정에서 Age / Length / CallTimeWeekend / CallLengthWeekend / AvgCallLengthNight / AvgCallLengthWeekend등 6개 변수에 대해 P값이 0.05보다 큰 값을 보이고 있다. 이는 95% 신뢰수준 하에서 귀무가설을 기각할 수 없다는 뜻이므로 해당 변수에 대해서는 분산이 동질하게 분포되어 있음을 알 수 있다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | | |
|  | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
| Age | 집단-간 | .058 | 3 | .019 | .554 | .645 |
| Length | 집단-간 | .731 | 3 | .244 | 2.943 | .032 |
| CallTimeWeekend | 집단-간 | .102 | 3 | .034 | 1.802 | .145 |
| CallLengthWeekend | 집단-간 | .276 | 3 | .092 | 2.892 | .034 |
| AvgCallLengthNight | 집단-간 | .001 | 3 | .000 | .252 | .860 |
| AvgCallLengthWeekend | 집단-간 | .008 | 3 | .003 | .782 | .504 |

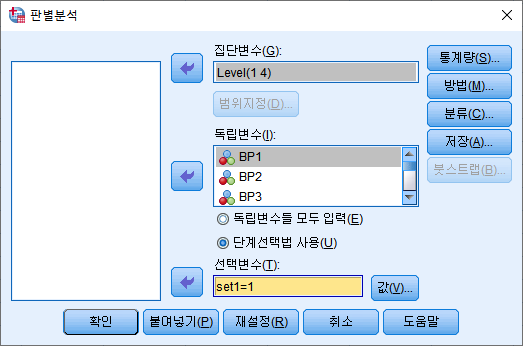
데이터의 분산이 동질하다고 판단된 6개 변수에 대한 일원분산분석 값을 살펴보면 Age / CallTimeWeekend / AvgCallLengthNight / AvgCallLengthWeekend의 4개 변수에 대해서 유의확률이 0.05보다 높게 나타나 95% 신뢰수준 하에서 귀무가설을 기각할 수 없다. 따라서 고객등급에 따라 유의미한 차이를 보이는 변수는 Length / CallLengthWeekend 2개인 것을 확인할 수 있다.

* 1. 카이제곱검정

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

13개 이분형 변수에 대한 카이제곱검정을 실시한 결과, HS5와 Gender 2개의 변수는 유의확률이 0.05를 초과하여 95% 신뢰수준 하에서 고객등급과 관련이 없는 것으로 나타났다. 따라서 고객등급에 따라 유의미한 차이를 보이는 변수는 BP1 / BP2 / BP3 / BP4 / BP5 / HS1 / HS2 / HS3 / HS4 / HS6 / BillProgValid 11개인 것을 알 수 있다.

* 1. 다중판별분석



다중판별분석에는 위 이미지와 같은 설정을 따랐다. 집단변수로는 고객등급(1~4)을 지정하였고, 독립변수에는 앞서 실시한 검정 결과에서 유의하다고 판단된 13개의 변수를 사용하였다. 또한 단계선택법을 사용하여 다중판변분석에 사용하도록 하였다.

선택변수에는 테스트용 데이터셋인 1을 선택하였으며, 분류 메뉴에서 요약표를 표시할 수 있게 하여 예측정확도까지 나타낼 수 있도록 하였다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **분류 결과a,b** | | | | | | | | |
|  |  |  | Level | 예측 소속집단 | | | | 전체 |
|  |  |  | 1 | 2 | 3 | 4 |
| 선택된 케이스 | 원래값 | 빈도 | 1 | 420 | 377 | 3 | 0 | 800 |
| 2 | 267 | 428 | 105 | 0 | 800 |
| 3 | 6 | 232 | 488 | 74 | 800 |
| 4 | 0 | 0 | 179 | 621 | 800 |
| % | 1 | 52.5 | 47.1 | .4 | .0 | 100.0 |
| 2 | 33.4 | 53.5 | 13.1 | .0 | 100.0 |
| 3 | .8 | 29.0 | 61.0 | 9.3 | 100.0 |
| 4 | .0 | .0 | 22.4 | 77.6 | 100.0 |
| 선택되지 않은 케이스 | 원래값 | 빈도 | 1 | 107 | 93 | 0 | 0 | 200 |
| 2 | 73 | 104 | 23 | 0 | 200 |
| 3 | 1 | 64 | 116 | 19 | 200 |
| 4 | 0 | 0 | 42 | 158 | 200 |
| % | 1 | 53.5 | 46.5 | .0 | .0 | 100.0 |
| 2 | 36.5 | 52.0 | 11.5 | .0 | 100.0 |
| 3 | .5 | 32.0 | 58.0 | 9.5 | 100.0 |
| 4 | .0 | .0 | 21.0 | 79.0 | 100.0 |
| a. 원래의 선택 집단 케이스 중 61.2%이(가) 올바로 분류되었습니다. | | | | | | | | |
| b. 원래의 비선택 집단 케이스 중 60.6%이(가) 올바로 분류되었습니다. | | | | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **표준화 정준 판별함수 계수** | | | |
|  | 함수 | | |
| 1 | 2 | 3 |
| BP1 | .106 | .026 | -.160 |
| BP2 | .946 | 1.126 | -.335 |
| BP3 | 1.472 | .349 | .176 |
| BP5 | .512 | .666 | .111 |
| HS1 | .192 | -.188 | -.396 |
| HS2 | -.238 | .216 | .224 |
| HS3 | -.285 | .149 | .412 |
| HS6 | -.279 | .200 | .074 |
| BillProgValid | .156 | .006 | .710 |

일원배치 분산검정과 카이제곱검정을 통해 고객등급에 따른 차이가 있다고 판단된 13개 변수를 사용하여 다중판별분석을 실시한 결과, 다중판별분석의 예측정확도는 학습용 데이터에서 **61.2%**, 검증용 데이터에서 **60.6%**이다.

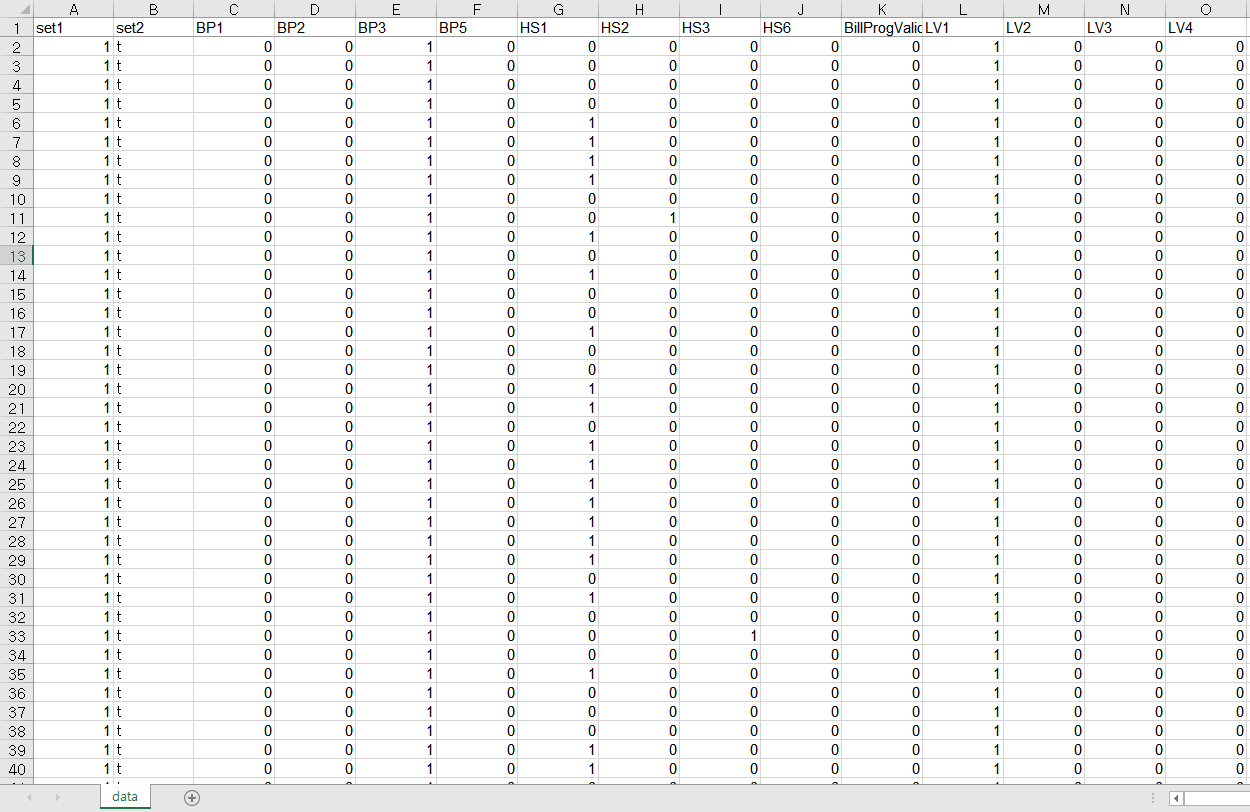
다중판별분석에서 고객등급을 예측하는데 사용된 변수는 총 9개로 나타났다.

* 1. 최종 유의변수

원본 데이터셋에 있는 39개의 변수 가운데 일원배치 분산검정 및 카이제곱검정을 통해 13개의 변수가 실제 고객등급을 예측하는데 유의하다고 판명되었다. 이 변수를 사용하여 다중판별분석을 통한 예측을 실시한 결과, 9개 변수(BP1 / BP2 / BP3 / BP5 / HS1 / HS2 / HS3 / HS6 / BillProgValid)만 사용하고도 충분히 예측이 가능한 것으로 나타났다.

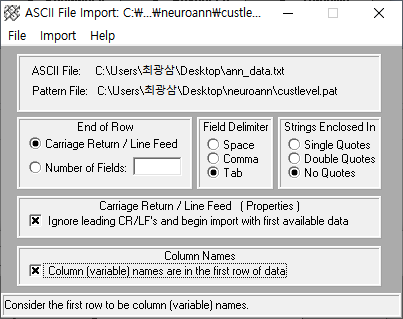
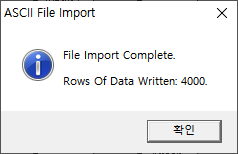
따라서 최종 유의변수는 9개(BP1 / BP2 / BP3 / BP5 / HS1 / HS2 / HS3 / HS6 / BillProgValid)이다.

1. **ANN**

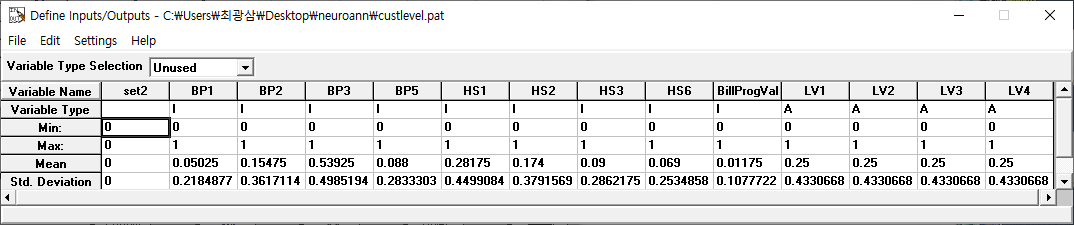


인공신경망 예측에 앞서 모형에서 Multiclass를 받아들일 수 있도록 고객등급이 입력되어 있는 Level 변수를 4개의 이분류 변수로 분류하였다. 고객등급이 1인 데이터는 LV1에서 1로, 아닌 데이터는 0으로 표시하였다. 나머지 등급에 대해서도 해당 등급에 해당하면 1의 값을, 해당하지 않으면 0의 값을 가지는 것으로 하여 LV2, LV3, LV4로 나누었다.

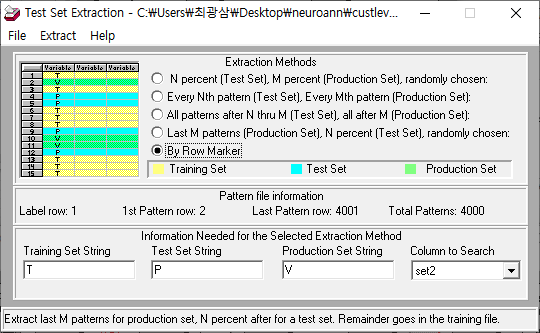
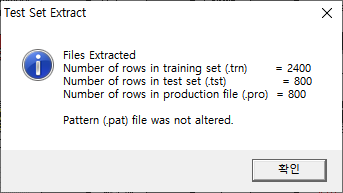
1. 모델 설정

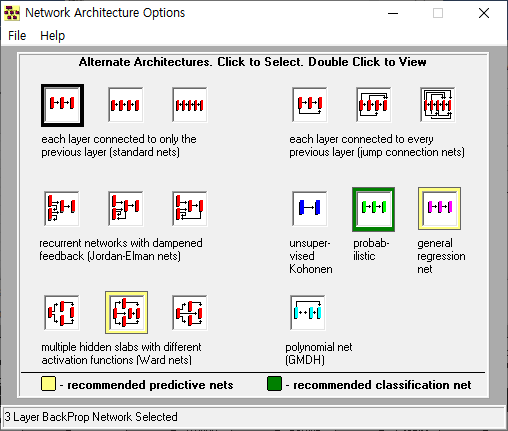
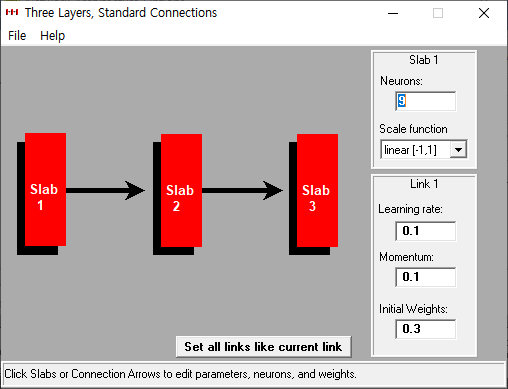
위 이미지와 같은 설정으로 4000개의 데이터를 Neuroshell2를 통해 불러왔다.

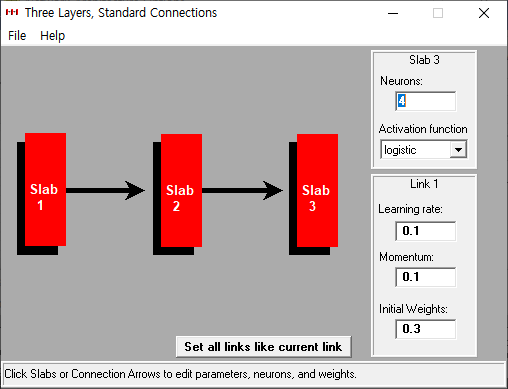
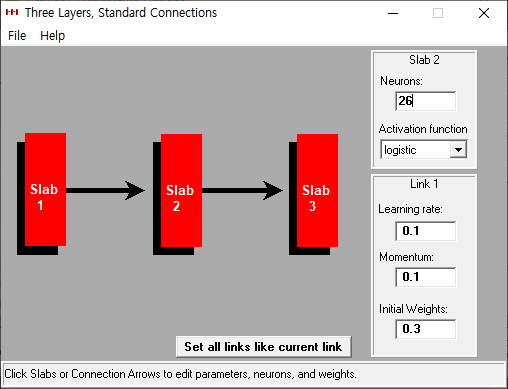


독립변수에 대해서는 Input으로 입력받게 하였고, 고객등급을 나타내는 4개의 변수(LV1, LV2, LV3, LV4)는 Actual Output으로 지정하였다. 데이터 구분값에 대한 정보(S: 학습용, P: 테스트용, V: 검증용)가 있는 set2 변수는 공란으로 두었다.

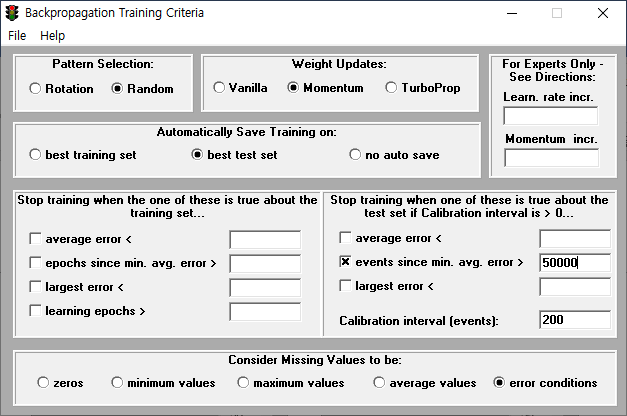
 

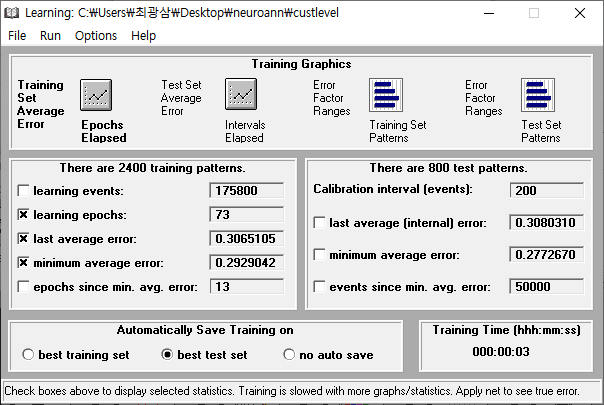
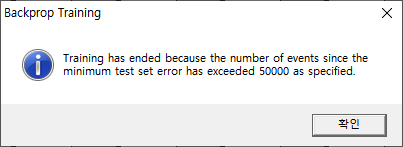
행 구분을 사용하여 set2 변수(T, P, V)를 기준으로 인공신경망 예측 모델에서 데이터를 구분하도록 설정하였다.

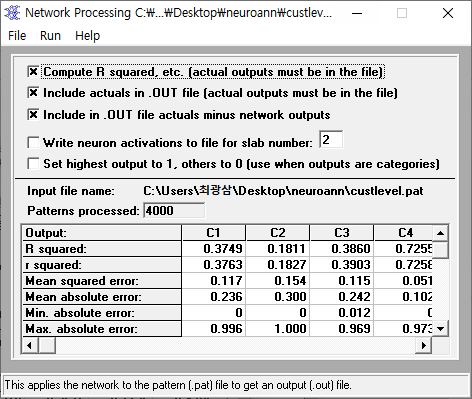
 

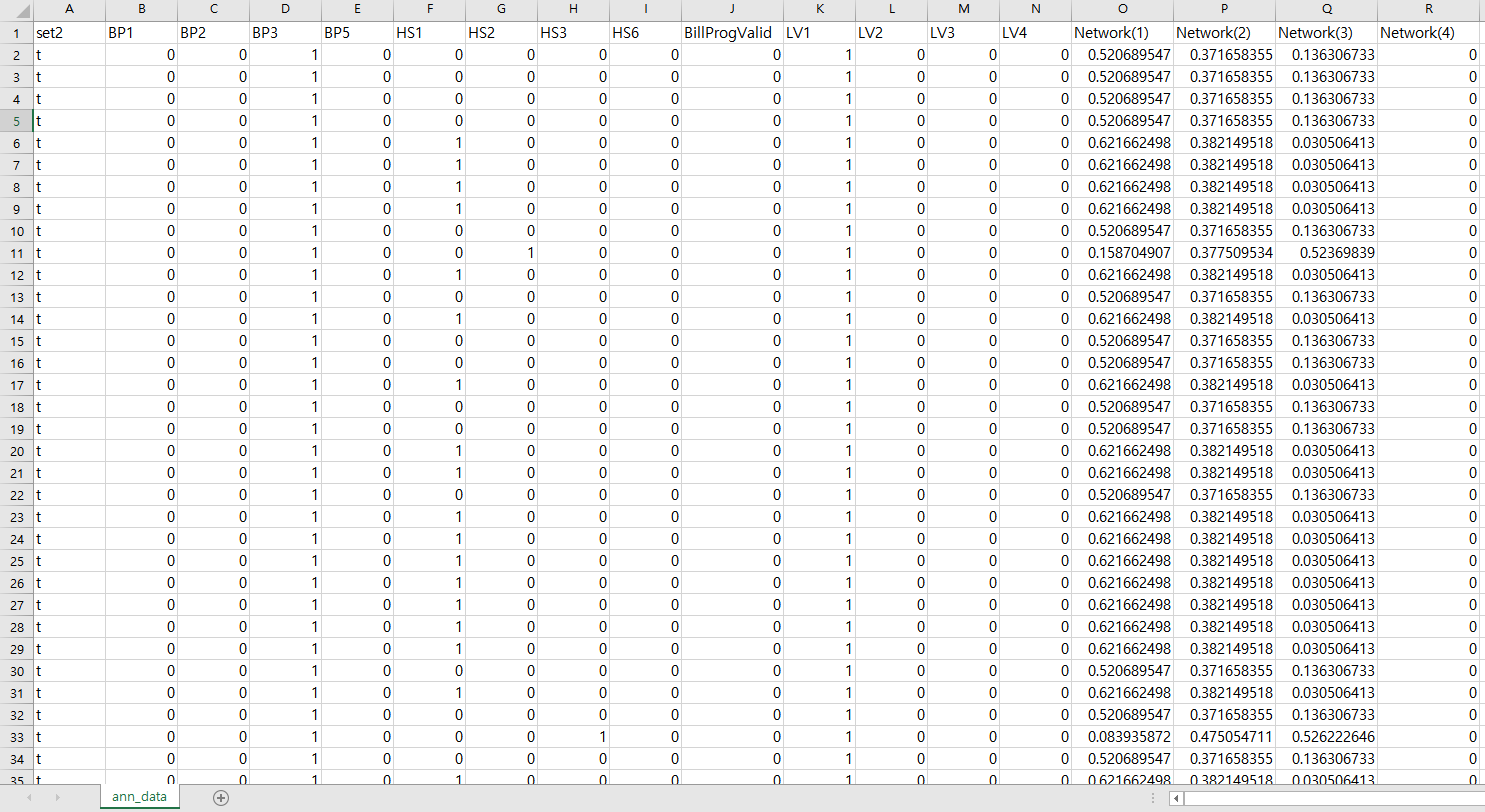
인공신경망 설계에서는 3개 층 네트워크를 선택하였으며, 입력층 값(Slab1, 9)과 출력층 값(Slab3, 4)을 비교하여 둘의 합계의 2배가 되는 값((9 + 4) \* 2 = 26)을 노드 값(Slab2)으로 설정하였다.



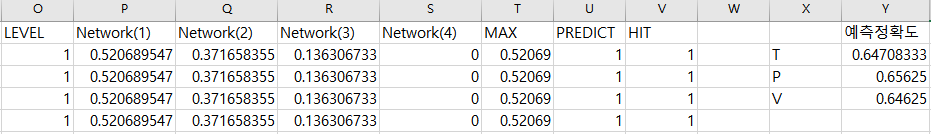
학습 조건에는 평균 에러값의 최소값에 도달한 이후에도 50000번 반복해서 학습하도록 지정하고 200번 반복할 때마다 에러가 얼마나 발생하는지 체크하도록 한 뒤, 학습을 진행하였다.





Neuroshell2내에서 연산을 실시하여 그에 대한 결과 파일을 생성하도록 하였다. 그런 다음에는 엑셀로 결과에 대한 데이터 값을 복사하여 예측 정확도를 체크하기 용이하게 하였다.

1. 예측 정확도

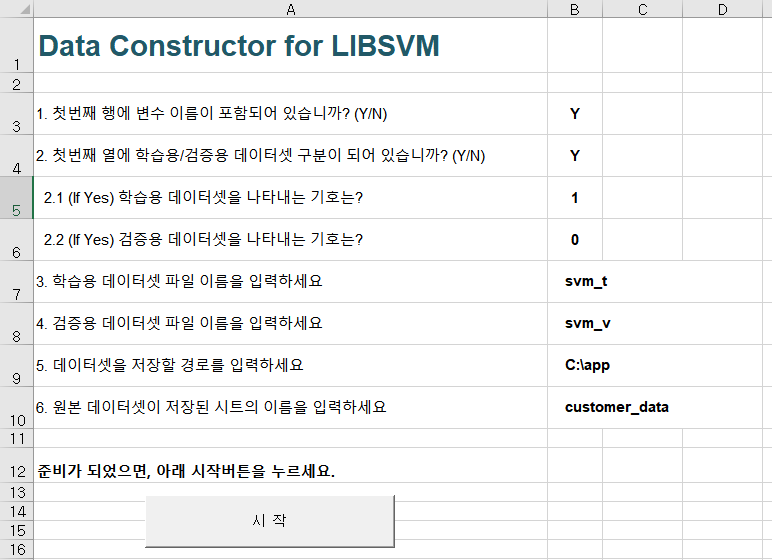


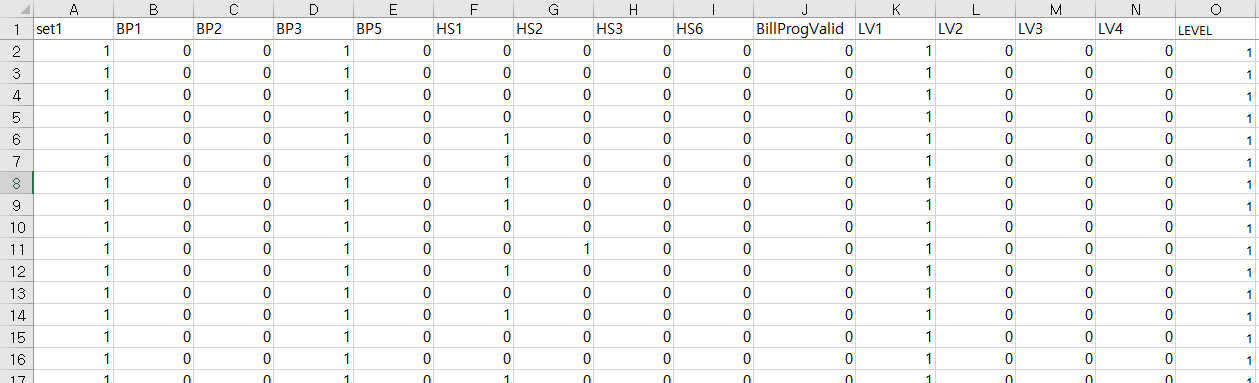
엑셀의 MAX 함수를 이용하여 네 개의 출력값 중 가장 높은 값을 찾아내어, PREDICT 변수에 MAX 변수의 값이 Network(1)와 같은 경우에는 1, Network(2)와 같은 경우에는 2, Network(3)과 같은 경우에는 3, 나머지 경우에 대해서는 4의 값을 가지도록 하였다.

또한 IF 함수를 이용하여 원래 이분류한 고객등급을 원래의 값으로 바꿔놓아, LEVEL 변수와 PREDICT 변수의 값이 같으면 HIT 변수에 1로 표시되도록 하였다.

각 데이터셋에 대한 예측 정확도를 알아보기 위하여 AVERAGEIF 함수를 이용하여 데이터셋에 대한 HIT 변수 값의 평균을 도출하도록 하였다. 그 결과, 학습용 데이터셋에 대해서는 약 **65%**의 예측 정확도를 보였고 테스트용 데이터셋에서는 약 **66%**의 예측 정확도를 보였다. 검증용 데이터셋에서도 약 **65%**로 높은 예측 정확도를 보였다.

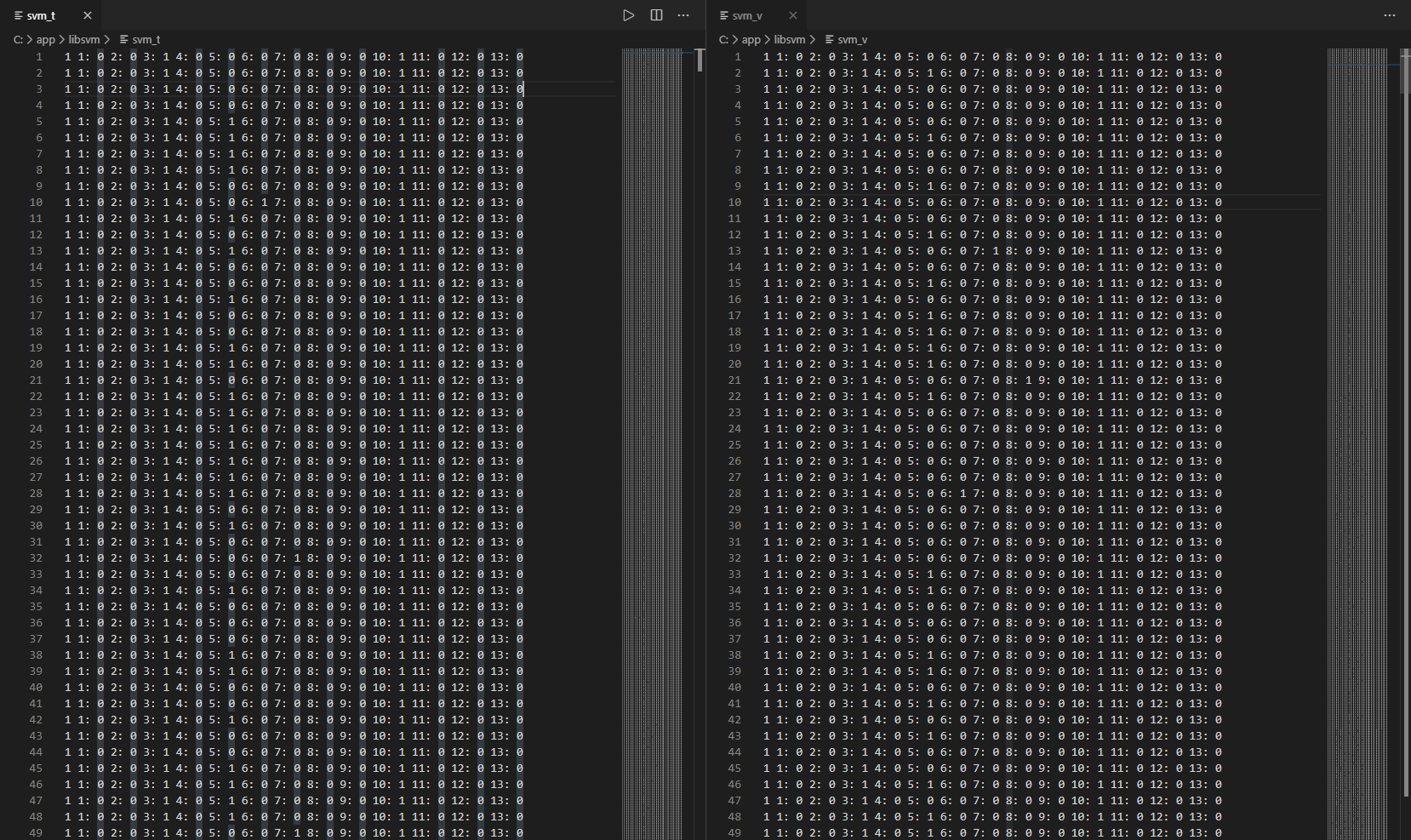
1. **SVM**
2. LIBSVM 사용을 위한 전처리





LIBSVM을 사용하려면 먼저 학습용 데이터와 검증용 데이터를 나누어 각각의 파일로 만들어야 한다. 이를 위해서 국민대학교의 안현철 교수가 제작한 ‘svm\_data\_converter’를 사용하였다. svm\_data\_converter은 엑셀의 매크로 기능을 이용하여 엑셀 데이터 파일을 가지고 몇몇 설정만 거치면 간편하게 LIBSVM 사용을 위한 파일을 생성해준다.

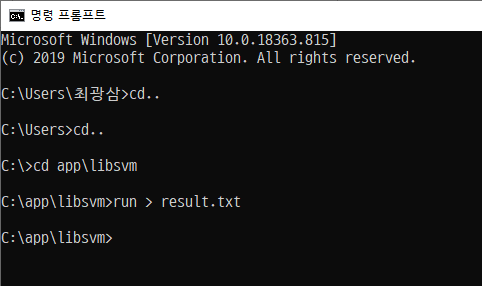
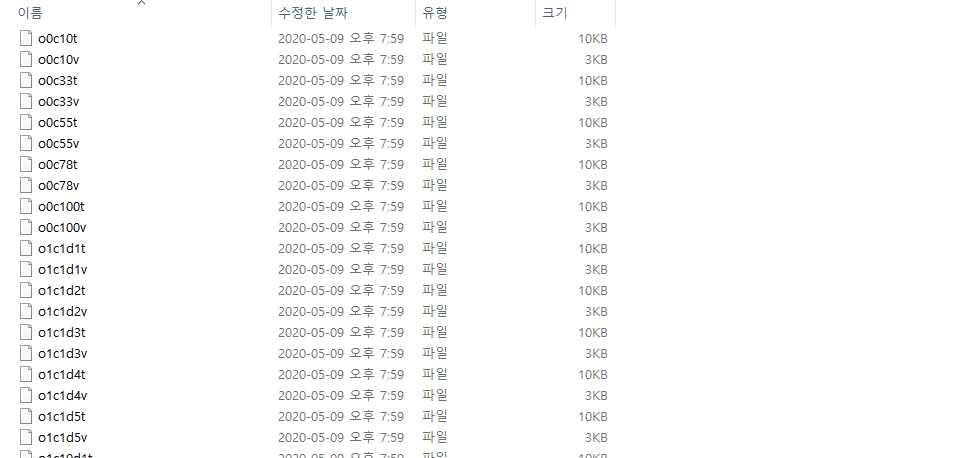
위 이미지는 svm\_data\_converter를 실제로 실행한 모습이며 사용된 엑셀 데이터셋은 미리 전처리 과정을 거친 online\_diet 데이터를 사용하였다. 학습용 데이터 파일은 svm\_t, 검증용 데이터 파일은 svm\_v로 생성하였다.

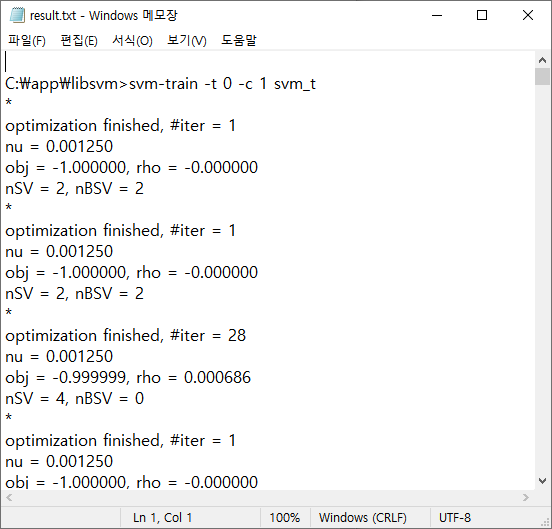


실제 생성된 파일을 열어보면 위와 같이 LIBSVM에 적용할 수 있는 형태로 변환된 것을 확인할 수 있다.

1. 최적의 조합 및 결과

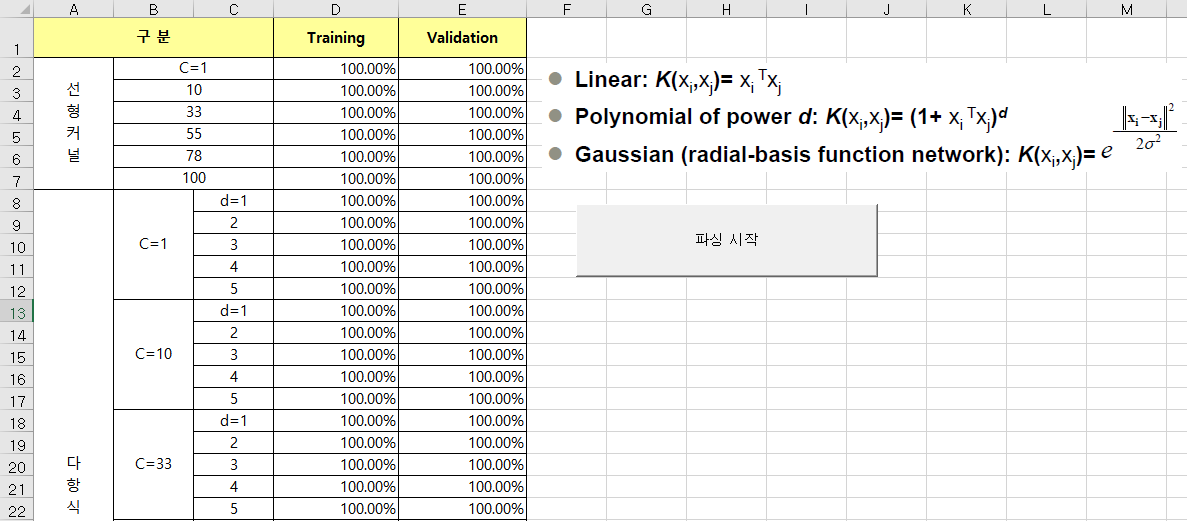
LIBSVM은 커널 타입, 감마값 등 무수히 많은 파라미터의 조합으로 데이터를 활용하여 예측을 실행할 수 있다. 가장 최적의 조합을 찾는 것이 높은 예측 정확도를 갖는데 중요한 요소로 작용한다. 최적의 값을 찾기 위해서는 안현철 교수가 제작한 ‘run.dat’ 프로그램과 ‘svm\_parser’를 이용하였다. run.dat은 LIBSVM 내에서 사용할 수 있는 여러 개의 파라미터 구성(-t, -c, -g, -r, -d 등)을 한번에 실행하여 최적 값을 찾기 위해 예측값을 여러 번 구해야 하는 사용자의 불편을 줄여준 프로그램이다. svm\_parser는 앞서 run.dat을 사용하여 얻은 결과물을 엑셀로 보기 쉽게 나열하여 가장 높은 예측 정확도를 보이는 조합과 그에 따른 LIBSVM 결과를 찾는데 도움을 준다.

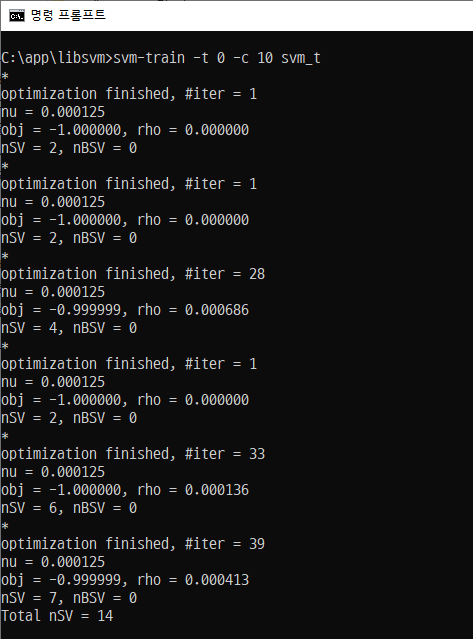
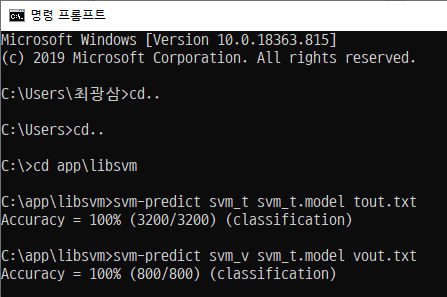


실제 CMD를 통하여 run.bat을 실행시키면 이미지와 같이 다양한 구성이 실행된 것을 알 수 있다. 실행 결과는 result.txt에 저장하도록 하였다.

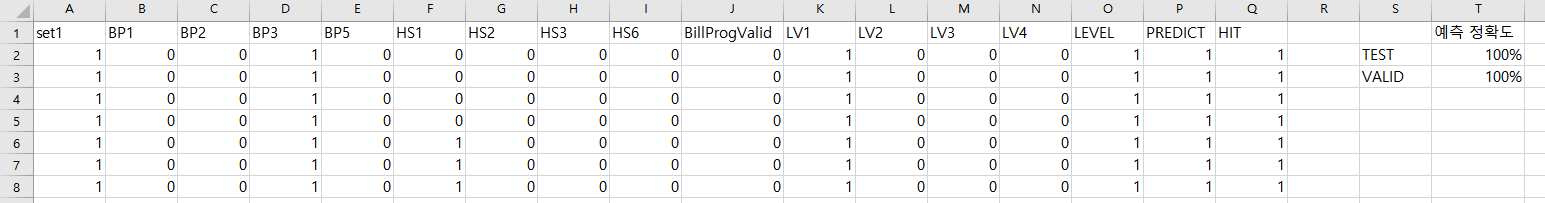
결과 파일에 있는 값을 svm\_parser에 넣어 실행시키면 다음과 같은 결과를 얻을 수 있다.



선형 커널, 다항식 커널, RBF 커널 등 최적의 값을 따로 찾아서 사용하지 않아도 어떤 조합을 사용한 경우에도 예측 정확도가 100%에 육박하는 것으로 나타났다.

이 같은 결과를 검증하기 위해 svm\_parser에서 예측 정확도가 100%로 추정된 한 가지 사례를 svm-train 값에 적용하여 예측모델(svm\_t.model)을 생성하였다. 생성된 예측모델을 svm-predict를 이용하여 테스트용 데이터셋과 검증용 데이터셋에 적용하였을 때 두 데이터셋 모두에서 **100%**의 예측 정확도를 보이는 것을 알 수 있다.



svm-predict를 실시하여 생성된 예측값인 tout.txt와 vout.txt의 값을 원본 엑셀에 복사하여 예측 정확도를 확인하였다. 검증에는 엑셀의 IF, AVERAGEIF 함수를 사용하였다. 그 결과 테스트용 데이터셋에 대한 예측 정확도는 **100%**, 검증용 데이터셋에 대한 예측 정확도 역시 **100%**로 나타나, SVM 기법을 사용하였을 경우에 다중판별분석을 사용하였을 때(예측 정확도: 약 61%) 또는 인공신경망 기법을 사용하였을 때(예측 정확도: 약 65%)보다 매우 높은 예측 정확도를 보이는 것을 확인할 수 있다.