|  |
| --- |
| REPORT |
| **과제 5**  SVM(Support Vector Machine) |





|  |  |  |
| --- | --- | --- |
| 제출일 | 2020. 05. 01 |  |
| 과목명 | 비즈니스애널리틱스개론 |  |
| 담당교수 | 안현철 |  |
| 전공 | 경영정보전공 |  |
| 학번 | 20175288 |  |
| 이름 | 최광삼 |  |

**목차**

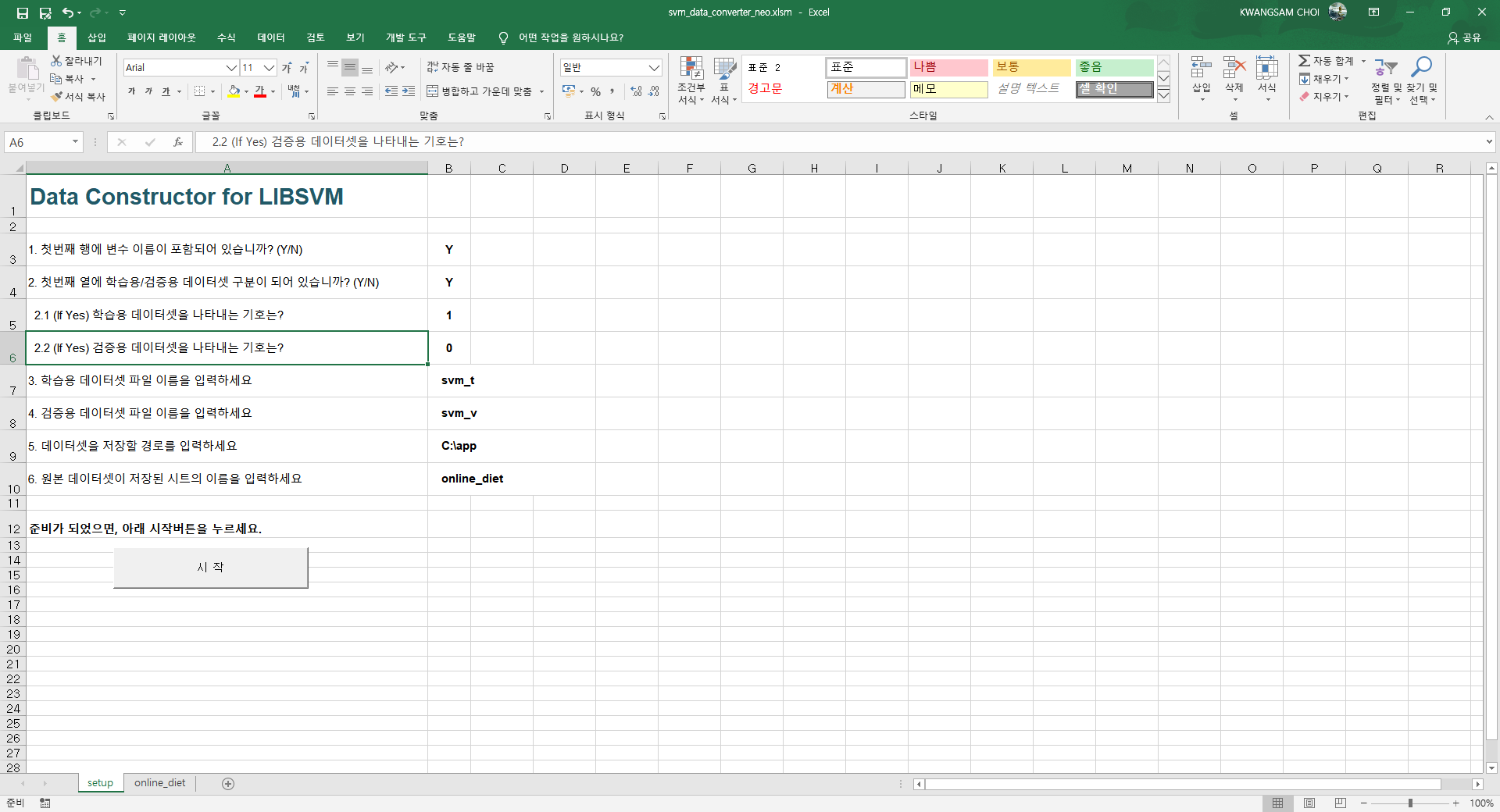
1. **SVM**
2. LIBSVM 사용을 위한 전처리
3. 최적의 조합 및 결과
4. **SVM**

SVM(Support Vector Machine)은 초평면이라는 개념을 도입하여 적은 데이터셋으로도 비선형 패턴을 반영할 수 있고 예측 정확도가 뛰어난 이분류 예측 기법이다. SVM 기법은 가볍고 빠른 성능을 보여주고 구조적 위험이 적은데다 전술한대로 적은 데이터로도 높은 예측 정확도를 보여 현존하는 분류 모형 중에서도 가장 진보한 모형으로 평가된다.

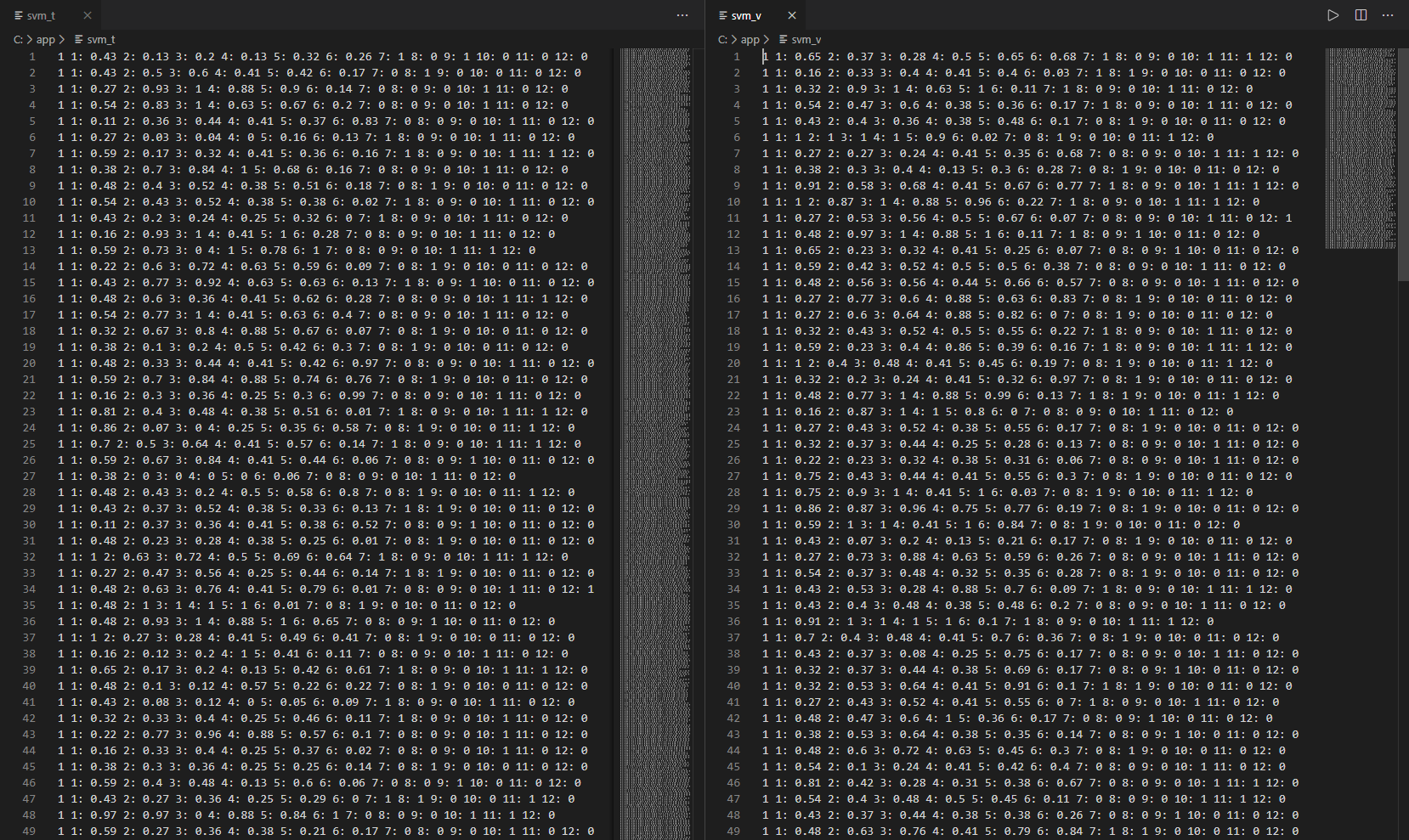
그러나 SVM 기법을 사용하기 위해서는 매우 복잡한 과정을 거쳐야 한다는 단점이 있는데, 이러한 단점을 보완하기 위하여 SVM 기법을 사용하기 쉽게한 툴이 바로 LIBSVM이다. LIBSVM은 대만의 Chih-Jen Lin 교수가 개발한 오픈소스 소프트웨어로 데이터셋만 제대로 갖추었다면 손쉽게 SVM 기법의 적용이 가능하다.

* 1. 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.



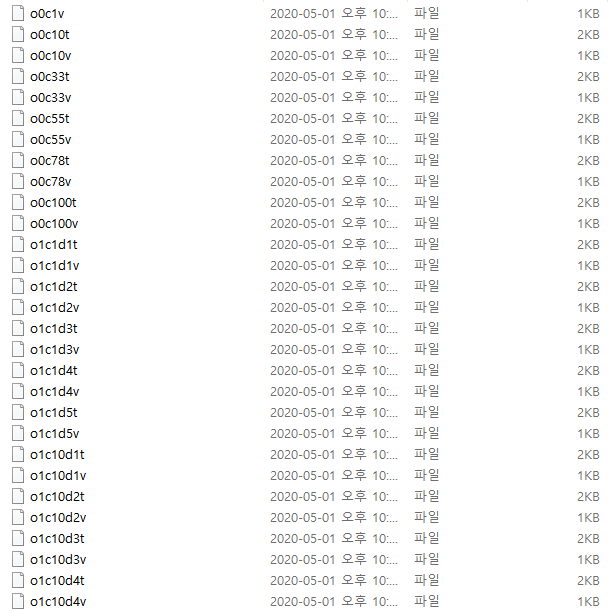
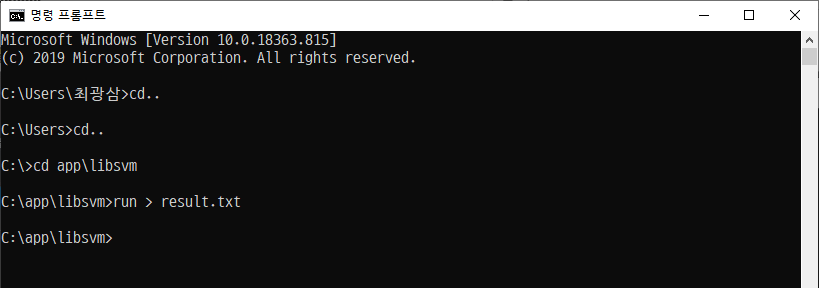
위 이미지는 svm\_data\_converter를 실제로 실행한 모습이며 사용된 엑셀 데이터셋은 미리 전처리 과정을 거친 online\_diet 데이터를 사용하였다. 학습용 데이터 파일은 svm\_t, 검증용 데이터 파일은 svm\_v로 생성하였다.

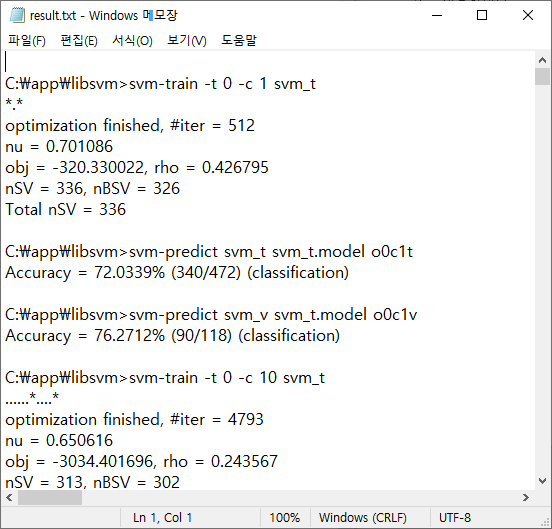


실제 생성된 파일을 열어보면 위와 같이 LIBSVM에 적용할 수 있는 형태로 변환된 것을 확인할 수 있다.

* 1. 최적의 조합 및 결과

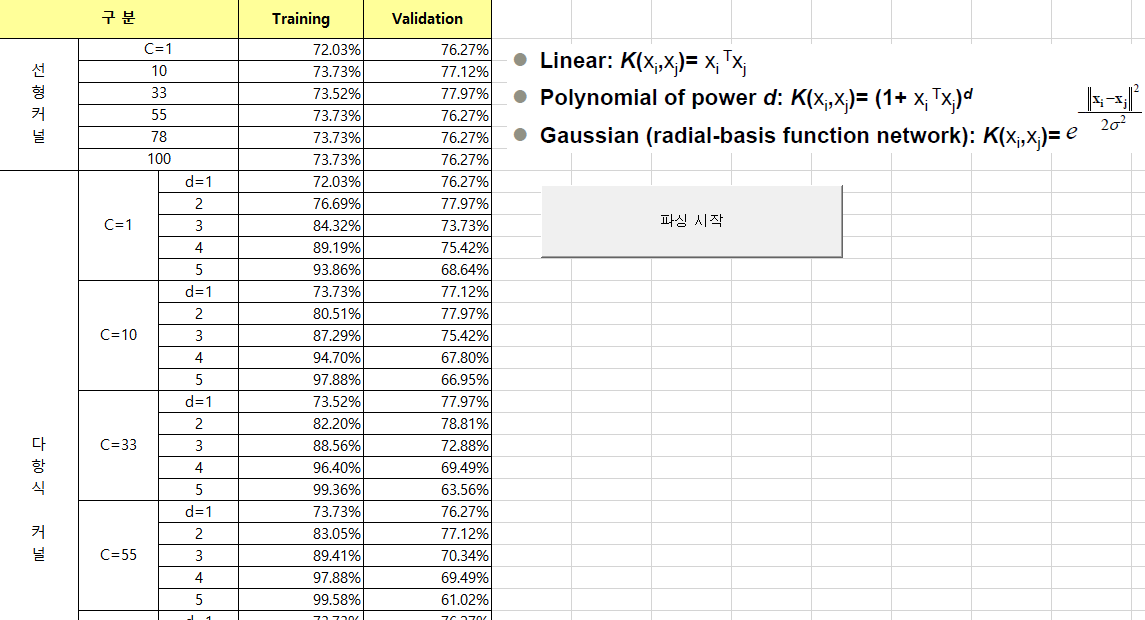
LIBSVM은 커널 타입, 감마값 등 무수히 많은 파라미터의 조합으로 데이터를 활용하여 예측을 실행할 수 있다. 가장 최적의 조합을 찾는 것이 높은 예측 정확도를 갖는데 중요한 요소로 작용한다. 최적의 값을 찾기 위해서는 안현철 교수가 제작한 ‘run.dat’ 프로그램과 ‘svm\_parser’를 이용하였다. run.dat은 LIBSVM 내에서 사용할 수 있는 여러 개의 파라미터 구성(-t, -c, -g, -r, -d 등)을 한번에 실행하여 최적 값을 찾기 위해 예측값을 여러 번 구해야 하는 사용자의 불편을 줄여준 프로그램이다. svm\_parser는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

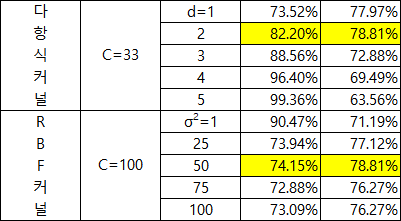




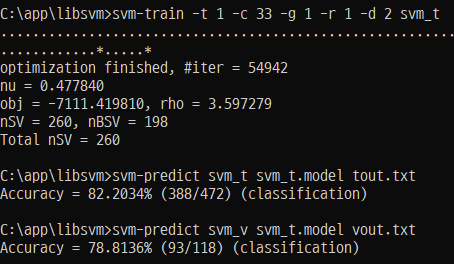
실제 CMD를 통하여 run.bat을 실행시키면 이미지와 같이 다양한 구성이 실행된 것을 알 수 있다. 실행 결과는 result.txt에 저장하도록 하였다.

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

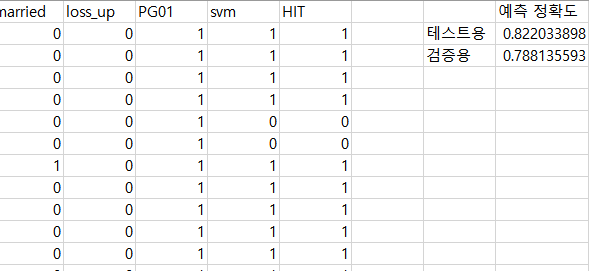




이 중에서 검증용 데이터의 예측 정확도가 가장 높은 경우는 78.81%로 두 건이 발견된다. 이 두 건 가운데 학습용 데이터의 예측 정확도가 높은 것은 82.20%이다. 따라서 LIBSVM에서 online\_diet 데이터 셋을 사용하여 여러 조합을 실행하였을 때 다항식 커널에서 -c 값을 33으로 주고, -d의 값을 2로 주었을 때 가장 높은 예측 정확도를 보인다는 사실을 확인할 수 있다.



이를 실제로 LIBSVM을 실행하여 해당 값을 적용해보면 테스트용 데이터의 예측 정확도는 82.2034%, 검증용 데이터의 예측 정확도는 78.8136%로 전술한 바와 동일한 값이 나오는 것을 알 수 있다.



LIBSVM을 통해 생성된 tout.txt(테스트용), vout.txt(검증용)의 값을 원본 데이터에 대입하고 엑셀의 IF, AVERAGEIF 함수를 이용하여 재현해보았을 때도 동일한 예측 정확도를 얻을 수 있었다.