PSEUDO LABLING

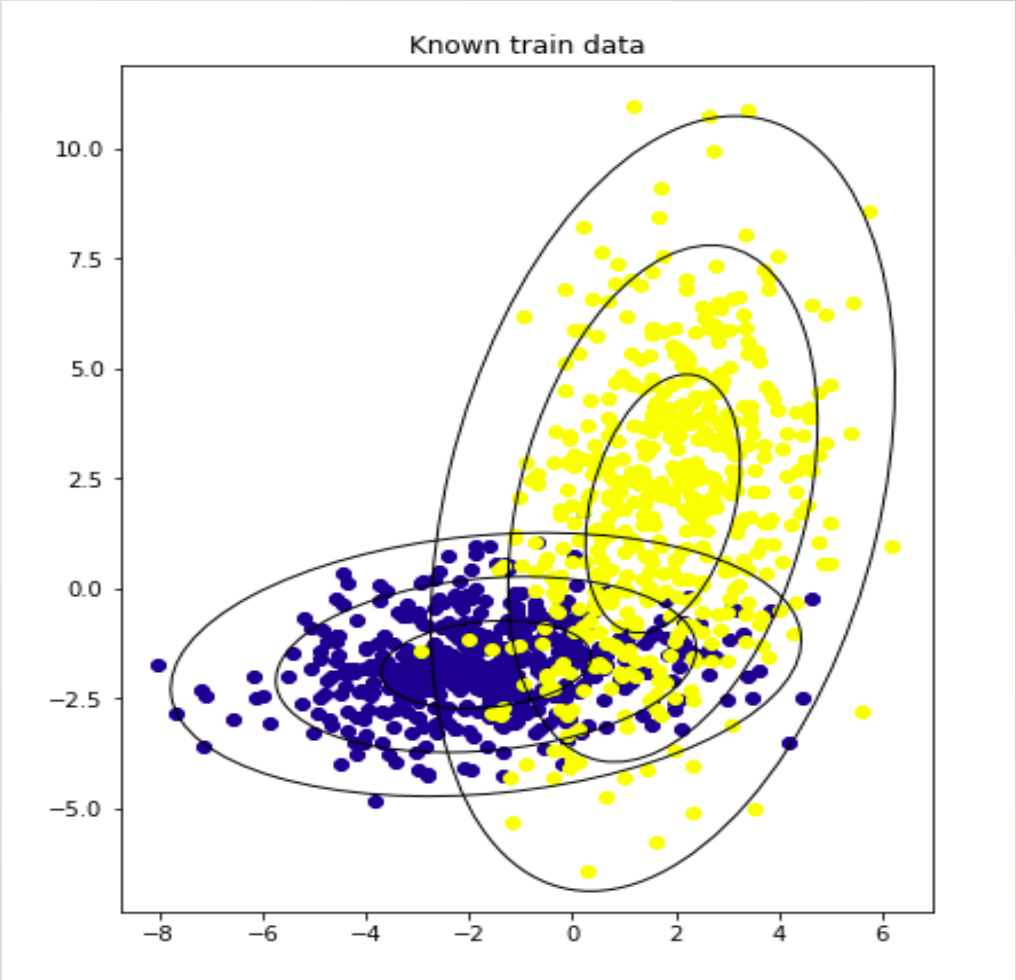
INSTANT 대회 특징

- 매우 장난스러운 대회

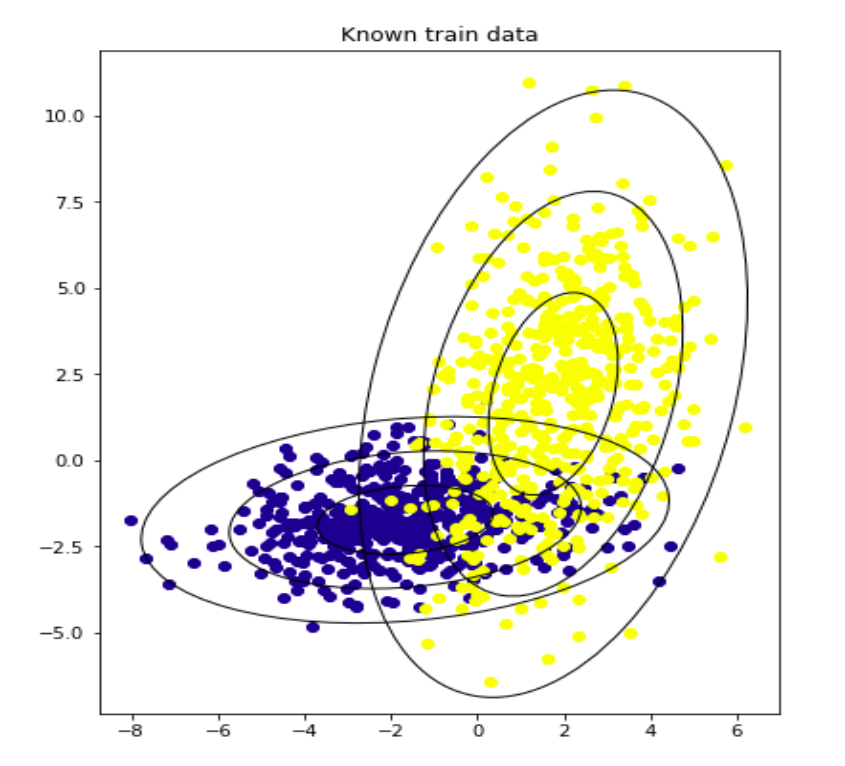
- Sklearn의 make_classification을 이용해 데이터 생성

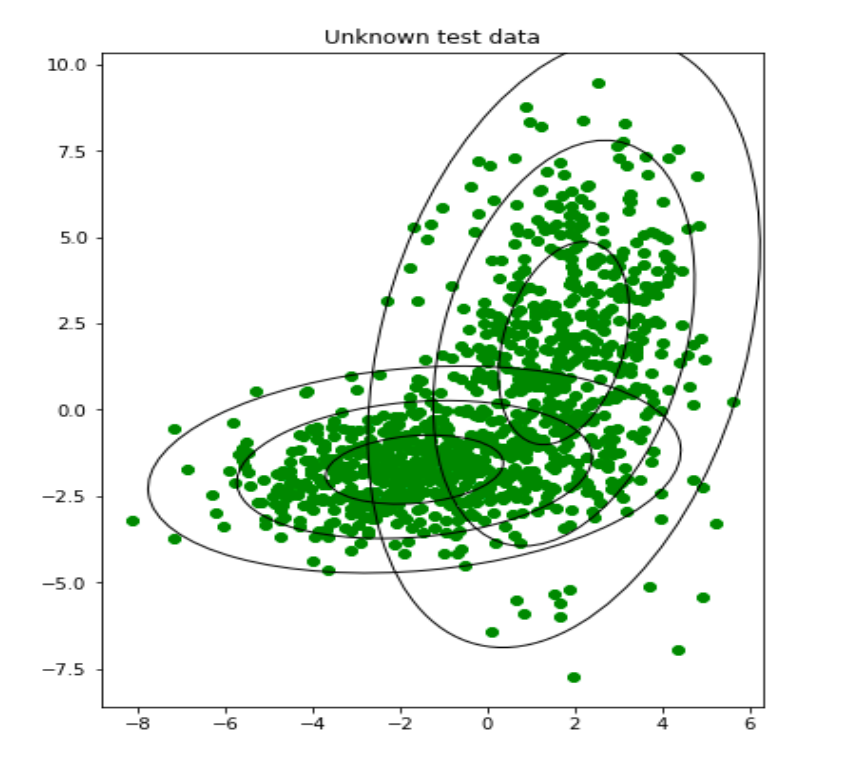
MAKE_CLASSIFICATIOIN

- 두개의 다변수 가우시안 분포 생성 후, 클래스 지정



QUADRATIC DISCRIMINANT ANALYSIS





PSEUDO LABELING

- Psedo Labeling은 확정적인 테스트 데이터를 training set에 추가하는 방법

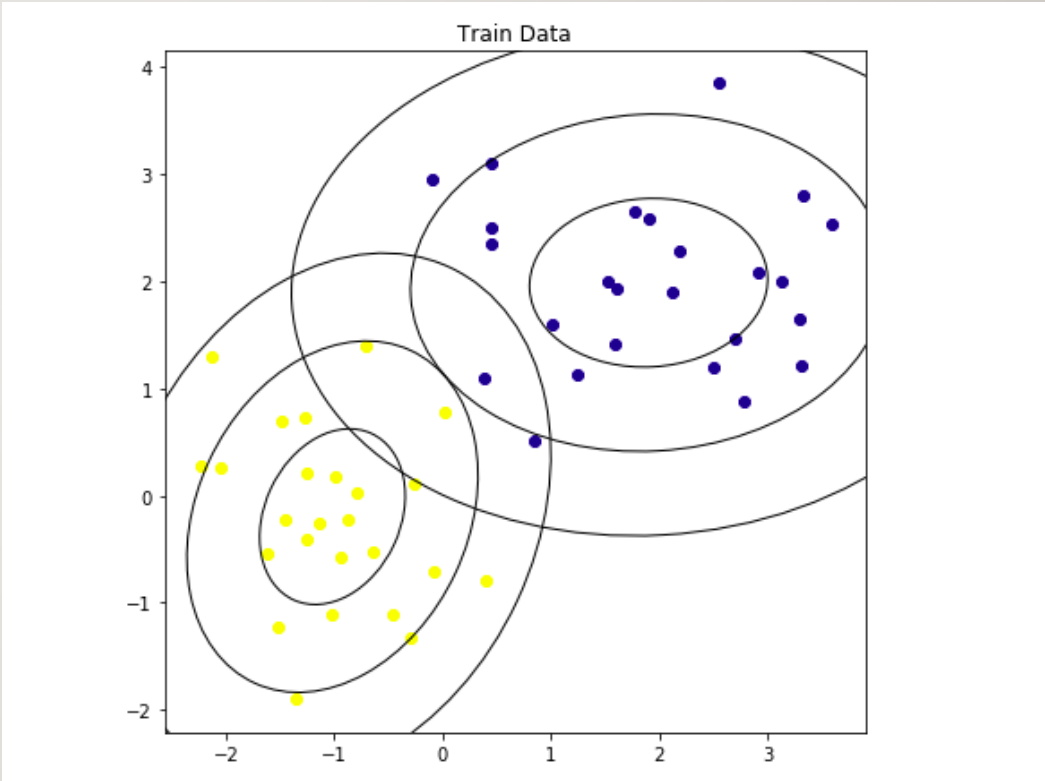
PSEUDO LABELING 의 5단계

- 1. training data를 이용해 모델 형성
- 2. test data의 label 예측
- 3. 확정적인 test dataset을 training dataset에 추가
- 4. 3번에서 만들어진 data set으로 모델 형성
- 5. 새로운 모델로 test data를 예측하고 캐글에 제출

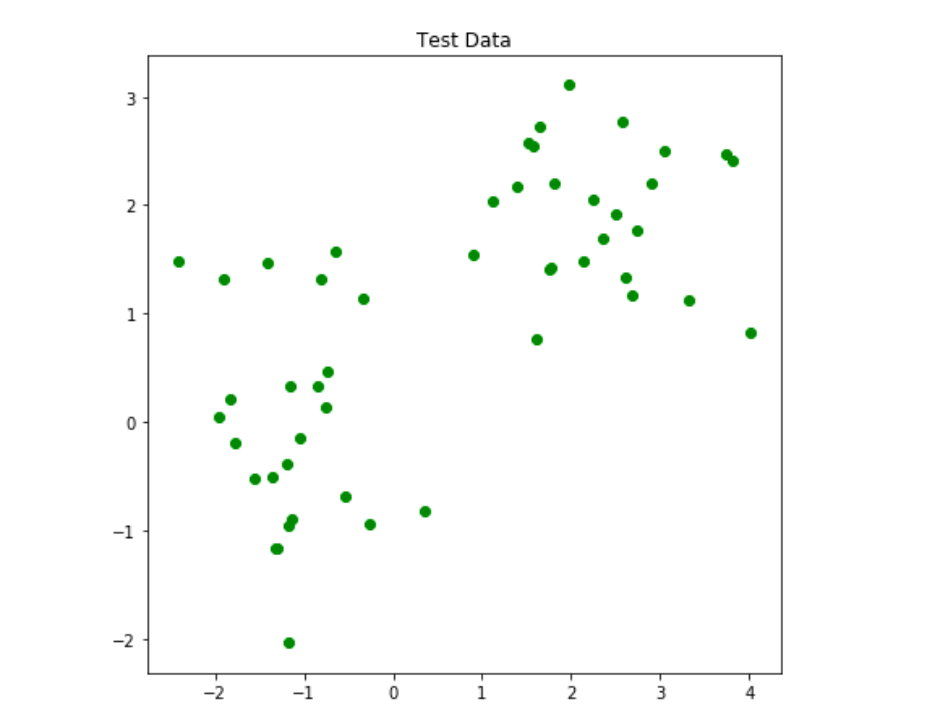
1. BUILD FIRST MODEL

- 50개의 training data

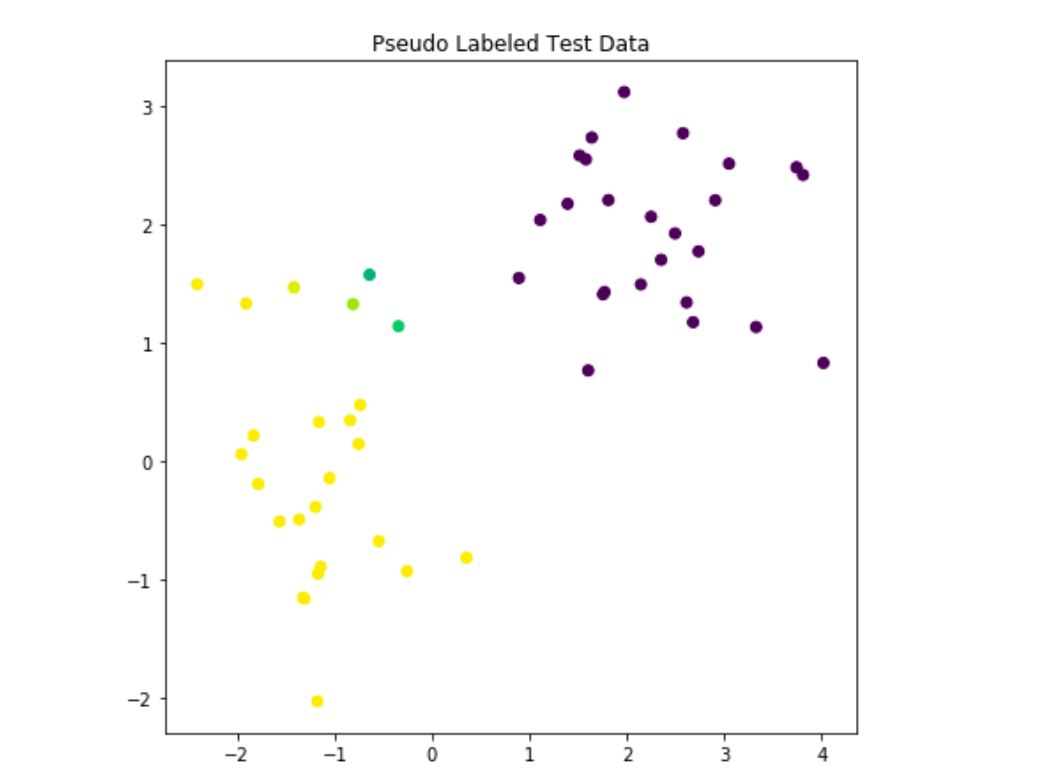
(target = 1 25개, target =0 25개)을 가지고 QDA를 이용해서 모델 형성



2.TEST DATA의 LABEL 예측



2. TEST DATA의 LABEL 예측

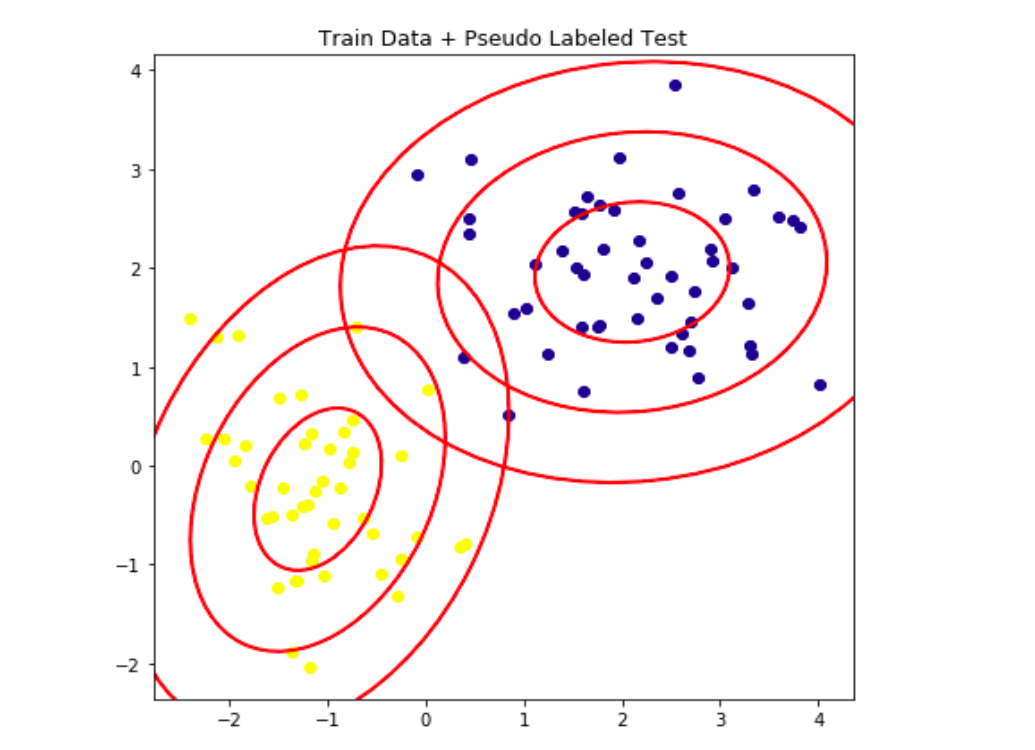


3. PSEUDO LABEL DATA를 추가해 준다.

- $Pr(y=1|x)>0.99$, $Pr(y=0|x)>0.99$ 인 점들을 training dataset에 추가해준다.

4. PSEUDO LABEL DATA를 추가해 모델을 만든다.

- 이렇게 형성된 90개의 train data들로 QDA 모델을 형성한다.



5 새로운 모델로 TEST DATA를 예측하고 제출

왜 PSEUDO LABELING이 효과가 있을까?

QDA에서 효과적인 이유

- QDA는 P-차원의 점들을 이용해 초 타원체을 찾는다.
- 더 많은 점이 있다면 QDA는 중심과 초 타원체를 더 잘 찾을 수 있게 된다.
- 그 결과 예측을 더 잘할 수 있게 된다.

다른 모델에서는?

- 모든 모델들은 p-차원에서 target =1 과 target=0 일 때의 모양을 찾는 형태로 시각화 할 수 있기 때문에 Pseudo Labeling이 모든 모델에 도움을 줄 수 있다.

단점은 없을까?

- Over-fitting이 문제일 수도 있을 것 같다.

- 현실 문제에선 쓸 수 없는 방법이 아닐까 싶다.

Q&A