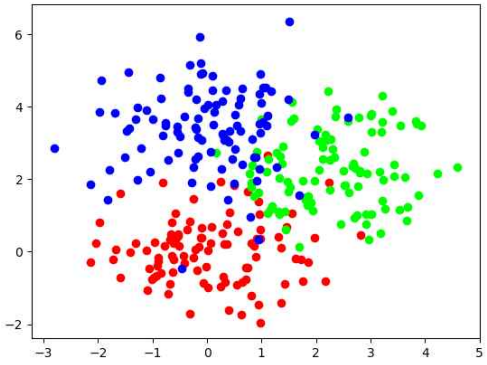
2019년도 1학기 패턴인식 Homework #2

- **개요**
  - 데이터를 생성하고, Bayesian classifier와 Linear model에 대한 실험을 수행
- **요구사항**
  - classifier의 성능을 측정하기 위한 데이터와 regressor의 성능을 측정하기 위한 데이터를 생성한다.
  - classification 데이터
    - ✓ 3개의 군집(클래스)을 이루는 2차원 데이터를 생성한다.
    - ✓ 각 클래스의 데이터는 가우시안 분포를 따르도록 100개씩 생성하며, 평균과 분산은 자유롭게 설정한다.
    - ✓ 이 중 무작위로 70% 데이터를 학습 데이터로 사용하고, 나머지를 테스트 데이터로 사용한다.

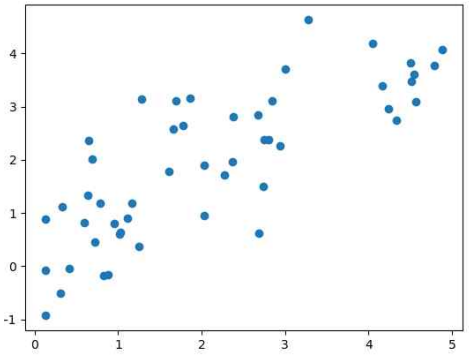


[그림 1] classification 데이터 예시

  - regression 데이터
    - ✓ 적절한 기울기와 절편을 가진 직선을 설정하고, 정규분포를 따르는 노이즈를 더하여 50개의 데이터를 생성한다.

$$y = w_1 x + w_0 + \epsilon$$
$$x, y, w_1, w_0, \epsilon \in R$$
$$\epsilon \sim N(0, 1^2)$$

    - ✓ 이 중 전반부 70%는 학습 데이터로 사용하고, 나머지는 테스트 데이터로 사용한다.



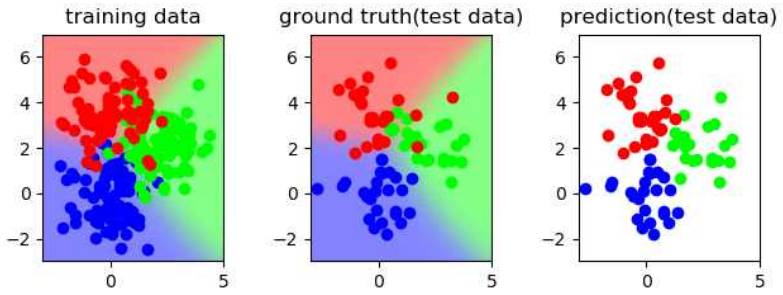
[그림 2] regression 데이터 예시

  - classification 데이터에 대해 Naive Bayesian classifier, KNN classifier, EM

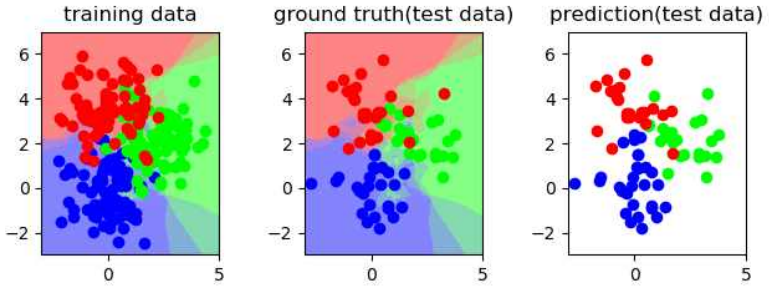
clustering, Logistic regression 그리고 SVM을 이용하여 분류 성능을 측정하고, regression 데이터에 대해 Linear regression의 성능을 측정한다.

- **예시**

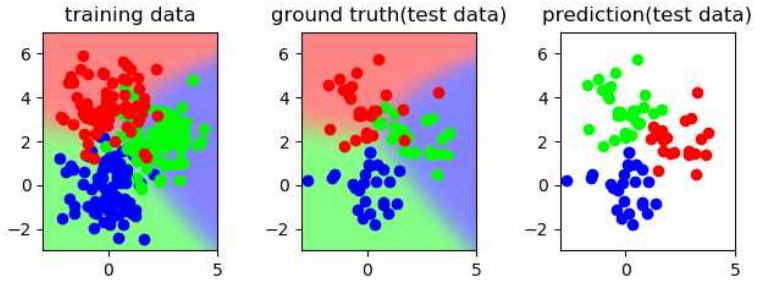
  - Naive Bayesian classifier



  - KNN classifier

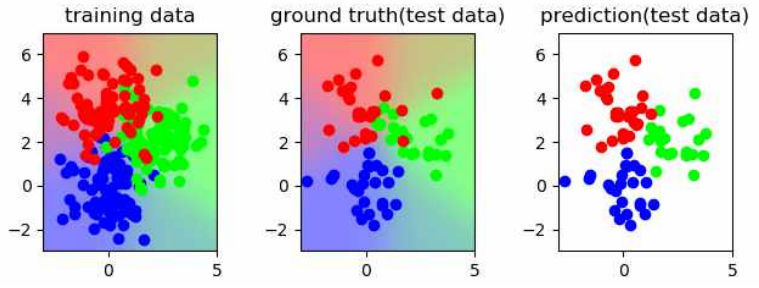


  - EM clustering

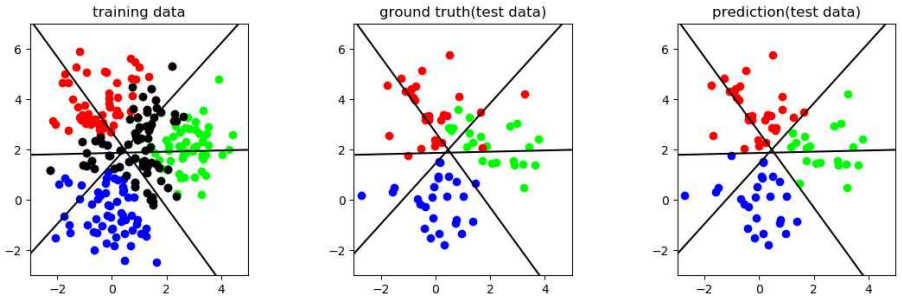


  클러스터링 알고리즘이기 때문에 ground truth에 나타난 군집의 색과 predict된 군집의 색이 다를 수 있으며, predict에서는 군집 간 색상만 다르면 됨.

  - Logistic regression

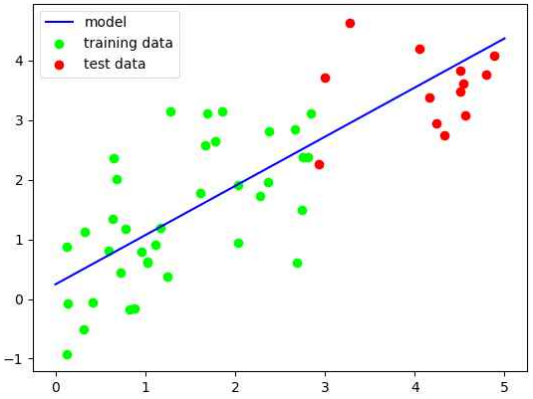


  - SVM



Linear SVM에 대한 실험을 수행한 결과를 플롯팅 시, 학습 데이터에 대한 플롯에서는 support vector를 찾아 표시하고(예시의 검은색 데이터), 각 클래스를 구분하는 decision surface를 표시할 것(예시의 검은색 직선).

- Linear regression



- **제출물**
  - 요구사항을 구현한 소스 코드
  - 플롯 결과를 포함한 한글 혹은 워드 문서
  - 위 두 가지를 압축한 .zip 파일(파일명: 학번_이름_homework2.zip)
- **기타**
  - Python 혹은 MATLAB 사용
  - Python 사용 시, numpy, matplotlib, scikit-learn 사용 권장
  - 이외의 어떠한 오픈소스 사용 가능