✈️ 준지도학습

* Reference> https://scikit-learn.org/stable/modules/semi_supervised.html

0. 개요

- **준지도학습(semi-supervised learning)**
  - train 데이터에서 일부 샘플에 레이블이 지정되지 않은 상황
  - `sklearn.semi_supervised`의 준지도학습 estimators는 레이블이 지정되지 않은 데이터를 추가적으로 활용하여 기존 데이터 분포의 모양을 더 잘 포착하고 새로운 샘플로 더 잘 일반화할 수 있음
  - 레이블이 지정된 데이터의 양이 매우 적고 레이블이 지정되지 않은 데이터의 양이 많을 때 활용할 수 있는 학습 방법
- y의 레이블이 지정되지 **않은** 항목
  - `fit()`을 사용하여 모형을 학습할 때 레이블이 지정되지 않은 곳에 레이블이 지정된 데이터와 함께 식별자를 할당하는 것이 중요
    - 사용하는 식별자: 정수 -1
  - 문자열(string) 레이블의 경우 문자열과 정수를 모두 포함할 수 있도록 y의 dtype이 `object`여야 함
- 데이터 세트의 **분포**에 대한 가정이 필요
  - 레이블이 지정되지 않은 데이터를 사용하려면 데이터의 기본 분포와 어떤 관계가 있어야 함
  - 준지도학습 알고리즘은 다음 가정 중 적어도 하나를 활용
- **준지도학습의 가정**

  a) 연속성/평활 가정(Continuity / smoothness assumption)
    - 서로 가까운 데이터들은 레이블(y)을 공유할 가능성이 높음
    - 일반적으로 지도학습에서도 가정되며 기하학적으로 간단한 의사 결정 경계에 대한 선호를 산출
    - 준지도학습의 경우, 평활 가정은 저밀도 영역의 결정 경계에 대한 선호도를 추가적으로 산출하기에 서로 가깝지만 다른 클래스에 속하는 소수의 포인트가 존재

b) 군집 가정(Cluster assumption)

- 데이터는 이산형(discrete) 군집을 형성하는 경향이 있으며, 동일한 군집에 있는 점들은 레이블을 공유할 가능성이 더 높음

  (라벨을 공유하는 데이터는 여러 군집에 분산될 수 있음)

- 평활 가정의 특이 케이스이며 클러스터링 알고리즘을 사용한 특징 학습을 발생

c) 매니폴드 가정(Manifold assumption)

- 데이터는 입력 공간보다 훨씬 낮은 차원의 매니폴드(다양체)에 대략적으로 놓여 있음
  - 이 경우 레이블이 지정된 데이터와 레이블이 지정되지 않은 데이터를 모두 사용하여 매니폴드를 학습하면 차원의 저주(curse of dimensionality)를 피할 수 있음
  - 이후 매니폴드에 정의된 거리와 밀도를 사용하여 학습을 진행할 수 있음
- 직접 모형화하기엔 힘들지만 자유도(degree of freedom)가 몇 개밖에 없는 일부 공정에서 고차원 데이터가 생성될 때 실용적
  - 예시) 사람의 목소리는 몇 개의 목소리 주름에 의해 조절되고, 다양한 얼굴 표정의 이미지는 몇 개의 근육에 의해 조절됨
  - 해당 경우 가능한 모든 음향파 또는 이미지의 공간이 아닌 생성 문제의 자연 공간(특정 공간)에서 거리와 부드러움을 각각 고려하는 것이 권장됨

1. 자가 학습(Self-Training)

- 야로프스키의 알고리즘을 기반으로 함
  - 주어진 지도 학습 분류기가 준지도 학습 분류기 역할을 할 수 있어 레이블이 지정되지 않은 데이터에 대해서도 학습할 수 있음
- `SelfTrainingClassifier`는 `predict_proba`를 구현하는 모든 분류기(classifier)와 함께 호출할 수 있으며, 매개 변수 `base_classifier`로 전달
  - 각 반복(iteration)에서 `base_classifier`는 레이블이 지정되지 않은 샘플의 레이블을 예측하고 레이블이 지정된 데이터 세트에 이러한 레이블의 하위 집합을 추가하는 작업을 수행
- 하위 집합의 선택은 선택 기준에 따라 결정됨
  - 예측 확률에 대한 `threshold`를 사용하거나 예측 확률에 따라 `k_best` 표본을 선택하여 수행할 수 있음

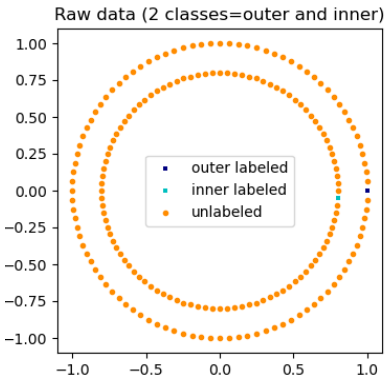
- 각 표본에 레이블이 지정된 iteration 뿐만 아니라 최종 적합(fit)에 사용되는 레이블 또한 속성으로 사용할 수 있음
    - `max_iter` 파라미터
        - 루프(반복)가 최대로 실행되는 횟수를 지정
        - `None`으로 설정하면 모든 샘플에 라벨이 있거나 해당 반복에서 새 샘플이 선택되지 않을 때까지 알고리즘이 반복됨

⚠️ Self-Training 분류기를 사용하는 경우 **분류기 보정(calibration)**이 중요

2. 라벨 전파(Label Propagation)

- 준지도 그래프 추론 알고리즘의 몇 가지 변형을 나타냄
- 해당 모델에서 가능한 기능들
    - 분류 작업에 활용
    - 대체(alternative) 차원 공간에 데이터를 투영하는 커널 방법(kernel method)
- 사이킷런이 제공하는 **라벨 전파 모델**
    - `LabelPropagation`
    - `LabelSpreading`

→ 둘 다 입력 데이터 세트의 모든 항목에 대한 **유사성 그래프(similarity graph)**를 구성하여 작동됨



레이블이 지정되지 않은 관측치의 구조는 클래스 구조와 일치 → 클래스 레이블은 레이블이 지정되지 않은 관측치의 훈련(train) 세트로 전파될 수 있음

- **차이점**

1) 레이블 분포에 대한 클램핑 효과에 대한 수정 사항

- 클램핑을 활용하면 알고리즘에서 실제 ground labeled data(우리가 정한 정답, 참값)가 지정된 데이터의 가중치를 어느 정도 변경할 수 있음
- `LabelPropagation` 알고리즘은 입력 라벨의 하드 클램핑을 수행 → **α = 0**을 의미
- 클램핑 인자는 완화될 수 있음
  - ex> **α = 0.2**: 원래 레이블 분포의 80%를 항상 유지하지만 알고리즘은 분포의 신뢰도를 20% 이내로 변경

2) 그래프의 유사도 행렬

- `LabelPropagation`: 수정을 거치지 않은 데이터로 구성된 원시 유사성 행렬 활용
- `LabelSpreading`**:** 정규화 특성이 있는 손실 함수를 최소화

  → 정규화 특성은 노이즈에 더 강함
  - 원본 그래프의 수정된 버전에서 반복됨
  - 정규화된 그래프 라플라스 행렬을 계산하여 edge 가중치를 정규화
  - Spectral clustering에도 활용됨

- 라벨 전파 모델의 **커널 함수(kernel method)**
  - `rbf`
    - $k(x1, x2) = exp(-\gamma||x1-x2||2, \gamma > 0)$
    - γ은 `gamma`라는 키워드로 설정 가능
    - 메모리에서 밀집행렬(dense matrix)로 표현되는 완전 연결 그래프(fully connected graph)를 생성

      → 행렬은 매우 클 수 있으며, 알고리즘의 각 반복에 대해 전체 행렬 곱셈 계산을 수행하는 비용(cost)과 결합하면 실행 시간이 엄청나게 길어질 수 있음
  - `knn`
    - $1[x' \in kNN(x)]$
    - k는 `n_neighbors` 라는 키워드로 설정 가능
    - 실행 시간을 획기적으로 줄일 수 있는 훨씬 더 메모리 친화적인 희소 행렬(sparse matrix)을 생성
  - 커널 선택은 알고리즘의 **확장성**과 **성능**에 모두 영향을 미침