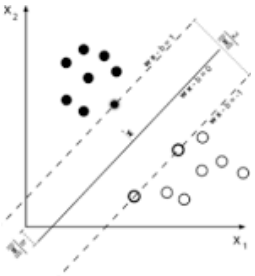
12week

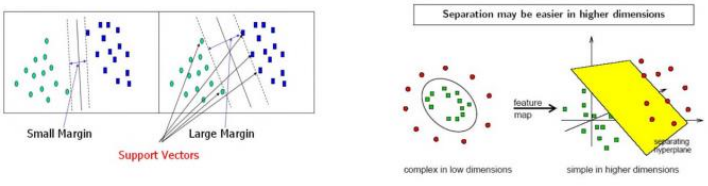
머신러닝 실습 과제

이연희

- **서포트 벡터 머신(support vector machine, SVM):** 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적이진 선형 분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상 된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. 주로 복잡한 분류 문제에 잘 맞으며 작거나 중간 크기의 데이터 셋에 적합하다.



- **SVM의 기본 원리:** 위와 같은 사진에서 흰색 점과 검은색 점이 학습 데이터로 주어졌을 때 두 그룹에서 각각의 데이터 간의 거리를 측정하여 두개의 데이터 사이의 중심을 구하고 그 가운데에서 최적의 초평면을 구함으로써 두 그룹을 나누는 방법을 학습한다.



(a) 선형 분류 (b) 비선형 분류

[그림 9.1] SVM을 이용한 분류

- [그림 9.1]의 (a), 직관적으로 자료를 군집별로 가장 잘 분리하는 초평면은 가장 가까운 훈련용 자료까지의 거리 (이를 마진(margin)이라 함)가 가장 큰 경우이며(마진이 가장 큰 초평면을 분류기(classifier)로 사용할 때, 새로운 자료에 대한 오분류가 가장 낮아진다.

- [그림 9.1]의 (b) 참고, SVM 모형은 선형분류 뿐 아니라, 커널 트릭(kernel trick)이라 불리는 다차원 공간상으로의 맵핑(mapping) 기법을 사용하여 비선형분류도 효율적으로 수행한다.

- **SVM에서 중요한 요소 3가지:**

  1) **마진(Margin):** 하나의 데이터 포인트(Support Vector)와 판별경계(Hyperplane)사이의 거리이다. 정확히는 각각의 클래스의 데이터 벡터들로부터 주어진 판별 경계까지의 거리 중 가장 짧은 것을 말한다. SVM에서는 마진이 클수록 분별을 잘하는 분류기로 판단한다. 일반적으로 학습 데이터에 과대적합 될수록 높은 복잡도의 비선형 분류기가 되는데 이렇게 학습데이터에 과대적합이 될수록 학습데이터의 노이즈까지 학습시켰기 때문에 오차가 커지는 현상이 발생한다. 따라서 학습데이터에 일정 정도의 오차를 내야 최적의 분류기가 된다. SMVM에서는 이 일정 정도의 오차를 Margin으로 둔 것이다. 이를 통해 일반화의 오류를 줄이면서 데이터 판별의 정확도를 높일 수 있다.

  2) **서포트벡터(Support Vector):** 위에서 언급한 데이터 포인트가 서포트벡터다. 판별경계까지의 거리가 가장 짧은 데이터 벡터를 서포트벡터라고 한다. 서포트벡터로 인해 SVM이 가지는 장점은 새로운 데이터 포인트가 들어왔을 때 전체 데이터포인트와의 내적거리를 보지 않고 서포트벡터와의 내적거리만 구하면 되므로 계산비용을 줄일 수 있다는 점이다.

  3) **커널(Kernel):** 선형분리가 불가능한 저차원의 데이터를 고차원의 공간 값으로 매핑시켜 선형평면으로 분류가 가능한 선형문제로 변환시켜 분류를 가능하게 할 수 있지만, 여기서 차원을 높임으로써 계산비용이 높아지는 문제가 있다. 이러한 문제를 해결할 수 있는 방법이 커널 방법이다. 커널 트릭은 SVM뿐 아니라 모든 비선형 문제를 가진 분류기에서 쓰일 수 있다.

- **선형 SVM 분류:**

  1) 하드마진 분류: 모든 샘플이 바깥쪽에 올바르게 분류되어 있는 경우

  문제: 데이터가 선형적으로 구분될 수 있어야 제대로 작동한다. 이상치에 민감하다.

  2) 마진 오류: 샘플이 도로 중간이나 반대쪽에 있는 경우

  소프트마진 분류: 도로의 폭을 가능한 넓게 유지하고, 마진 오류 사이에 적절한 균형을 잡아주어야 한다.

- **비선형 SVM 분류:**

  1) 다항식 커널: 다항식 특성을 추가하는 것은 간단하고 모든 머신러닝 알고리즘에서 잘 동작한다. 하지만 낮은 차수의 다항식은 매우 복잡한 데이터셋을 잘 표현하지 못하고 높은 차수의 다항식은 굉장히 많은 특성을 추가하므로 모델을 느리게 만든다. 커널 트릭이라는 수학적 기교를 적용하면 실제로는 특성을 추가하지 않으면서 다항식 특성을 많이 추가한 것과 같은 결과를 얻을 수 있다. 특징을 변환하는 대신 두

샘플 사이의 유사도를 의미하는 커널을 정의한다. 주로 다항 커널과 가우시안 커널이 사용된다.

- **선형 SVM 회귀:** 분류 문제와 달리 마진 안에 최대한 많은 샘플을 포함하는 것이 목적이다.
- **비선형 SVM 회귀:** 커널 SVM 모델을 사용한다.
- **결정 트리:** SVM처럼 분류, 회귀, 다중출력 작업이 가능한 알고리즘으로 복잡한 데이터셋도 처리가 가능하다. 데이터 전처리와 스케일링이 필요 없다.

  1) **확률 추정:** 결정 트리는 한 샘플이 특정 클래스 K에 속할 확률을 추정할 수 있다. 먼저 샘플에 대해 리프 노드를 찾기 위해 트리를 탐색하고 그 노드에 있는 클래스 K의 훈련 샘플의 비율을 반환한다.

  2) **CART:** 사이킷런이 결정 트리를 훈련시키기 위해 사용하는 알고리즘

  3) **계산 복잡도:** 예측을 위해 결정 트리를 루트 노드에서부터 리프 노드까지 탐색해야 한다.

  4) **회귀:** 결정 트리는 회귀 문제에서도 사용할 수 있다. 사이킷런의 DecisionTreeRegressor을 사용한다.

  5) **불안정성:** 결정 트리의 제한사항- 결정 트리는 계단 모양의 결정 경계를 만든다. 따라서 훈련 세트의 회전에 민감하다. 결정 트리의 주된 문제는 훈련 데이터에 있는 작은 변화에도 매우 민감하다는 것이다.