6주차 정리노트

2021840026 서동현
2019040025 김성욱

SVM 정리

- 서포트 벡터
  서포트 벡터 사이의 간격, 즉 도로의 폭이 최대가 되도록 학습
- 하드 마진 분류
  모든 훈련 샘플이 도로 바깥쪽에 올바르게 분류되도록 하는 마진 분류
- 소프트 마진 분류
  도로의 폭을 가능한 한 넓게 유지하는 것과 마진 오류 사이에 적절한 균형 잡기
- Iris-Virginica 품종 여부 판단
  사이킷런의 SVM 분류기 LinearSVC 활용

- 비선형 SVM 분류
  <선형 SVM에 특성 추가>
  특성 $x1$ 하나만 갖는 모델에 새로운 특성 을 추가한 후 선형 SVM 분류 적용
  <SVC + 커널 트릭>

SVM

- 두 클래스로부터 최대한 멀리 떨어져 있는 결정 경계를 이용한 분류기

- 선형, 비선형 분류, 회귀, 이상치 탐색

- 작거나 중간 크기의 데이터셋에 적합

선형 SVM

- 라지 마진: 클래스를 구분하는 가장 넓은 도로

- 분류 대상 클래스들 사이의 가장 큰 도로, 즉 라지 마진을 계산하여 클래스 분류

- 서포트 벡터
- 하드 마진 분류
- 소프트 마진 분류

비선형 SVM

- 선형 SVM에 특성 추가

1. 선형 SVM + 다항 특성 추가

: 특성 x1 하나만 갖는 모델에 새로운 특성 (x1^2)을 추가하여 1차원 데이터셋을 2차원 데이터셋으로 바꾸어 데이터셋을 선형적으로 구분할 수 있게 해준다.

Pipeline (PolynomialFeatures, StandardScaler, LinearSVC)

LinearSVC: C(마진 폭 조절), loss=”hinge”, random_state

2. 선형 SVM + 유사도 특성

- 유사도 함수: 각 샘플에 대해 특정 랜드마크(landmark)와의 유사도를 측정하는 함수

장점: 차원이 커지면서 선형적으로 구분될 가능성이 높음

단점: 훈련 세트가 매우 클 경우 동일한 크기의 아주 많은 특성이 생성

- SVC + 커널 트릭

커널 트릭 사용: 특성을 추가하지 않으면서, 다항식 특성을 많이 추가한 것과 같은 결과를 얻음. 특성을 추가하지 않으므로, 많은 수의 특성 조합이 생기지 않음.

1. 다항식 커널 :

▪ 장점 – 간단하고, 머신러닝 알고리즘에 잘 작동함.

▪ 단점 – 낮은 차수의 다항식은 매우 복잡한 데이터셋을 잘 표현하지 못함. 높은 차수의 다항식은 굉장히 많은 특성을 추가하므로 모델을 느리게 만듬.

2. 가우시안 RBF 커널 : 유사도 특성을 많이 추가하는 것과 같은 비슷한 결과를 얻을 수 있는 커널 트릭

질문:

1. 하드 마진 분류랑 소프트 마진 분류는 서로 어떤 상황에 사용하는게 좋아? 김성욱 서동현

하드마진 = 데이터가 완벽한 상황이 주어졌을 때

소프트 마진 = 데이터가 완벽한 상황이 아닐 때

2. SVM을 사용할 때 큰 크기의 데이터셋을 사용하면 안되는 이유가 뭐야? 김성욱

데이터 셋이 크면 느려질 수 있고, 훈련시간이 길어질 수 있다.