

homework#5 2023170855 최정연

1. 서포트 벡터 머신의 근본적인 아이디어는 무엇인가?

클래스 간의 마진을 최대로 만드는 결정 경계를 찾아 일반화 성능을 극대화한다. 커널 트릭을 이용하면 선형적으로 분리되지 않는 데이터도 높은 차원으로 매핑해 같은 원리를 적용할 수 있음.

2. 서포트 벡터가 무엇인가?

최적 결정 경계(초평면)에 가장 가까이 위치해 마진을 결정하는 훈련 샘플들

이 점들만이 모델 파라미터(가중치)에 직접 영향을 주며, 다른 샘플은 영향을 주지 않음

3. SVM을 사용할 때 입력값의 스케일이 왜 중요한가?

SVM은 샘플 간 거리에 민감하기에 각 특성의 범위가 크게 다르면 큰 값 특성이 마진 계산을 지배해 성능이 떨어짐. 보통 StandardScaler나 MinMaxScaler로 모든 특성을 같은 스케일 (평균 0, 분산 1 등) 설정.

4. SVM 분류기가 샘플을 분류할 때 신뢰도 점수와 확률을 출력할 수 있는가?

가능함.

`decision_function(X)` → 초평면까지의 signed distance(신뢰도 점수).

`predict_proba(X)` → `SVC(probability=True)`로 학습 시 Platt scaling으로 추정된 클래스 확률.

5. RBF 커널을 사용해 SVM 분류기를 훈련시켰더니 훈련 세트에 과소적합되었다. 이때, 어떤 값을 어떻게 수정해야 하는가?

모델 복잡도를 높여야 한다.

C 값 \uparrow → 규제 완화, 마진을 줄이고 오차를 덜 허용.

γ 값 \uparrow → RBF 커널 반경 축소, 개별 샘플 영향력 확대.

두 하이퍼파라미터를 GridSearchCV로 조합하여 해결