5주차 과제

2024170928 노연경

1. 서포트 벡터 머신의 근본적인 아이디어는 무엇인가?

- margin 을 최대화하는 결정초평면을 찾아 서포트 벡터만으로 분류기를 정의함으로써 일반화 성능을 얻는다.

margin : 각 클래스에서 가장 가까운 관측치 사이의 거리

2. 서포트 벡터가 무엇인가?

- 마진 plane 위의 데이터 포인트를 서포트 벡터라고 한다.

3. SVM을 사용할 때 입력값의 스케일이 왜 중요한가?

- SVM 같은 거리 기반 모델은 피처 간 거리를 계산할 때 값이 큰 피처에 훨씬 큰 가중치를 부여한다. 따라서 제대로 정규화하지 않으면 큰 스케일만 강조되어 학습이 왜곡될 수 있다.

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

- 기본적인 SVM은 하이퍼플레인으로부터의 거리(신뢰도 점수)는 출력하지만 직접적인 확률은 제공하지 않는다.

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

- 감마 값을 키워야한다. 감마값을 키우면 결정 경계가 더 복잡해지고 훈련 데이터에 더 잘 맞춰져 과소적합이 완화된다. 또한 C 값을 늘려 규제 강도를 낮춤으로써(복잡도를 제어하는 페널티를 줄임) 모델 복잡도를 높이는 방법도 있다.

\* RBF 커널 : 비선형 커널 중 한 종류로 입력 벡터 간의 유사도를 거리를 바탕으로 측정함

거리가 가까울수록 K(x,z)가 1에 가깝고 멀수록 0에 가까워진다. radial 모양의 기저 함수이므로 중심 x 주변에서만 크게 반응하고 멀어지면 급감함. 감마 파라미터는 반경에 영향을 주는데 감마 값이 작을 때는 반경이 넓어져 멀리 있는 포인트도 어느 정도 유사도를 반영하고 과소적합 위험이 있으며, 감마 값이 클 때는 반경이 좁아져 아주 가까운 포인트만 반영하고 따라서 과적합의 위험이 있음.