3장 분류 답지

3.1 훈련 세트를 섞어 모든 교차 검증 폴드를 비슷하게 만들어야 성능이 좋게 나타나는데, 훈련 세트를 섞음으로 인해서 역효과가 발생하는 경우는?

ㄴ시계열 데이터(주식 가격, 날씨 예보)

3.2 SGD가 온라인 학습에 적합한 이유를 온라인 학습의 의미와 함께 서술.

ㄴ온라인 학습은 데이터를 순차적으로 한개씩 또는 미니배치 단위로 주입하여 시스템을 훈련시키는 방법으로 확률적 경사 하강법 또한 한번에 하나씩 훈련 샘플을 독립적으로 처리하기 때문에 적합하다고 할 수 있다.

3.3 결정 임곗값에 대한 정밀도와 재현율의 그래프에서 임곗값이 높아질 때 정밀도의 울퉁불퉁한 구간이 나타나는 이유.

ㄴ임계값이 올라갈 때 올바르게 판단하는 샘플 수가 올라가지만 분모가 되는 전체 샘플 수도 함께 줄어드므로.

3.4 다중 분류기를 구성하는 과정에서 SGD 분류기와 SVM 분류기의 차이점

ㄴSGD 분류기는 직접 샘플을 다중 클래스로 분류할 수 있기 때문에 다중 분류를 위해 OvR이나 OvO를 별도로 적용할 필요가 없는 반면, SVM 분류기는 이진 분류만 가능하기 떄문에 OvO나 OvR을 적용하여 이진 분류기를 여러 개 사용해 다중 분류기를 구성해야 한다.

3.5 8처럼 보이는 숫자의 훈련 데이터를 실제 8과 구분되도록 성능을 향상하는 방안 3가지

1. 8처럼 보이는 숫자의 훈련 데이터를 더 많이 모아서 실제 8과 구분하도록 분류기를 학습시킨다.

2. 숫자의 동심원의 수를 세는 알고리즘을 통해 도움 될 만한 특성을 추가한다.

3. 동심원 같은 어떤 패턴이 드러나도록 이미지를 전처리한다.

3.6 다중 레이블 분류를 지원하는 분류 모델 네가지

ㄴ결정 트리, 렌덤 포레스트, OneVsRest분류기, KNN 분류기.

3.7 픽셀 강도에 잡음을 추가하기 위해 사용하는 함수

ㄴnumpy의 randint() 함수