

4주차 과제 김민섭

1. 수백만 개의 특성을 가진 훈련 세트에서는 어떤 선형 회귀 알고리즘을 사용할 수 있는가?

수백만 개의 특성을 가진 훈련 세트에서는 정규방정식과 특이값 분해의 계산 부담이 크므로 경사 하강법 계열의 알고리즘을 사용한다.

2. 훈련 세트에 있는 특성들이 각기 아주 다른 스케일을 가지고 있을 때, 이런 데이터에 잘 작동하지 않는 알고리즘은 무엇이며, 그 이유와 해결 방법은 무엇인가?

특성들의 스케일 차이가 크면 비용 함수의 등고선이 찌그러져 경사 하강법의 수렴이 불안정해진다. 이 문제는 특성 스케일링을 통해 해결할 수 있다.

3. 경사 하강법으로 로지스틱 회귀 모델을 훈련시킬 때 지역 최솟값에 갇힐 가능성이 있는가?

로지스틱 회귀의 비용 함수는 볼록하므로 경사 하강법을 사용할 때 지역 최솟값에 갇히지 않는다.

4. 충분히 오랫동안 실행하면 모든 경사 하강법 알고리즘이 같은 모델을 만들 것인가?

볼록 문제의 경우 충분한 학습 후에는 모든 경사 하강법 알고리즘이 유사한 전역 최적해에 수렴한다.

5. 검증 오차가 상승하면 미니배치 경사 하강법을 즉시 중단하는 것이 좋은가?

검증 오차가 일시적으로 상승하는 것은 노이즈일 수 있으므로, 바로 중단하지 않고 일정 기간 추세를 관찰한 후 조기 종료 기법을 적용한다.

6. 어떤 경사 하강법 알고리즘이 가장 빠르게 최적 솔루션의 주변에 도달하는가?

초기 단계에서는 확률적 경사 하강법이 빈번한 업데이트로 인해 최적점 주변에 빠르게 도달한다.

7. 다항 회귀를 사용했을 때 학습 곡선을 보니 훈련 오차와 검증 오차 사이에 간격이 크다. 이때 이 문제는 무엇이며, 해결 방법은 무엇인가?

훈련 오차는 낮으나 검증 오차가 높은 경우는 과대적합임으로, 모델의 복잡도를 낮추거나 규제 기법을 적용하여 개선할 수 있다.

8. 릿지 회귀를 사용했을 때 훈련 오차와 검증 오차가 거의 비슷하고 둘 다 높았다면, 이 모델에는 높은 편향이 문제인가? 아니면 높은 분산이 문제인가? 이때 규제 하이퍼파라미터 알파값은 어떻게 해야 하는가?

훈련 오차와 검증 오차가 모두 높고 비슷하면 높은 편향으로 인한 과소적합임으로, 규제 강도를 낮춰 알파값을 줄여야 한다.

9. 사진을 낮과 밤, 실내와 실외로 분류하려고 할 때, 두 개의 로지스틱 회귀 분류기를 만들어야 하는가? 하나의 소프트맥스 회귀 분류기를 만들어야 하는가?

사진 분류 문제는 낮과 밤, 실내와 실외가 상호 배타적이지 않으므로 각각의 이진 분류기로 구성하는 것이 적절하다.