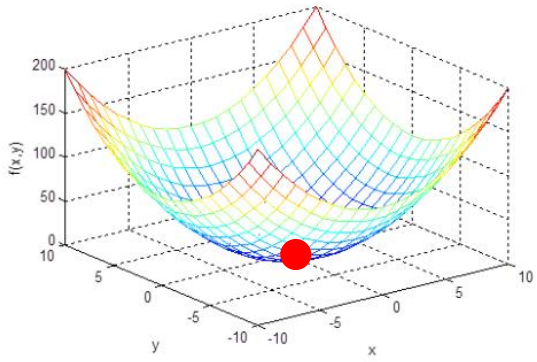
CH4 모델 훈련

목차

● 1. 서론

머신러닝이 어떻게 학습을 하고 높은 정확도를 구현 할 수 있을까? 모델의 비용 즉 예측값과 실제값의 차이가 가장 작아지는 점을 찾는 것이 학습의 목표이자 최상의 결과이다.



<그림 1> $f(x,y) = x^2 + y^2$ 그래프

좋은 학습 결과를 얻을 수 있는 기본적인 조건들이 있다.

1. 데이터의 양이 많을수록, 데이터에 잘못된 수치가 적을수록 적절하다.

   EX) 집값을 예측하는데 있어, 최대한 많은 집값이 있는 것이 학습에 유용하며, 수치가 잘못 표기된 데이터가 있으면 학습이 왜곡된다.

2. 모델이 더욱더 복잡하고 해당 데이터에 적합할 때 더 좋은 학습이 가능하다.

3. 결과에 영향을 줄 수 있는 변수가 많고 의미 없는 변수가 적을수록 좋을 것이다.

   EX) 껌 판매량과 범죄의 양 의 상관관계는 무의미하다. 범죄에 영향을 주는 변수들이 더 좋은 학습결과를 유도한다.

● 2. 회귀분석

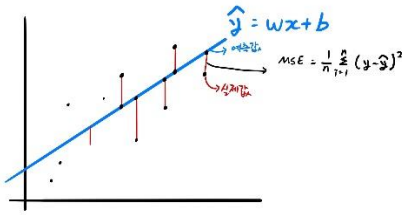
특징들의 수치를 바탕으로 목표값들을 가장 잘 설명하는 선형을 형성하는 것으로 단순선형회귀분석은 y= ax + b으로 되어있다.

$$y = \omega x + b$$

회귀분석의 오차를 측정하는 비용 함수는

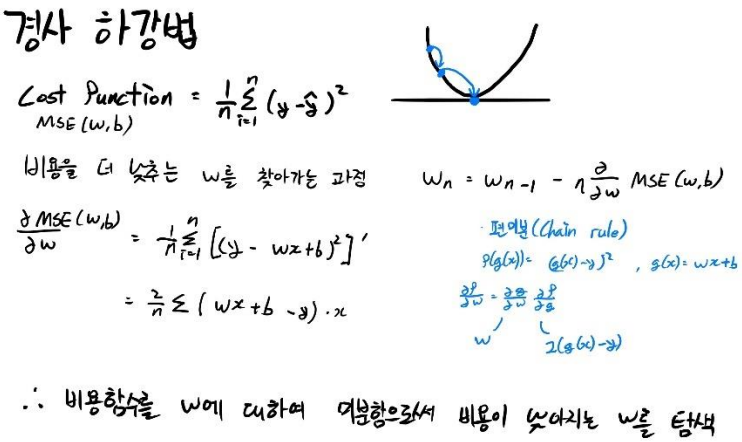
$$MSE(mean\ squared\ error) = \frac{\sum_{i=1}^{n}(y-\hat{y})2}{n}$$

비용 함수를 w에 대하여 미분하여 비용을 최소화 하는 w를 찾는 것이 학습의 목표이



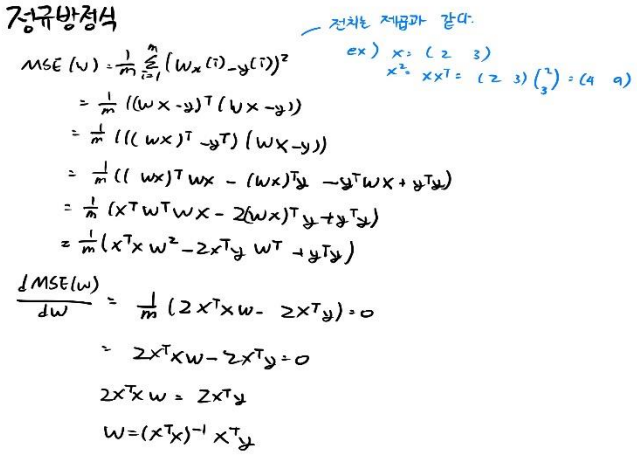
이를 위한 방법은 크게 경사하강법(Gradient Descent)와 정규방정식이 존재한다.

■ 1. 경사하강법(Gradient Descent)



■ 2. 정규방정식

앞선 과정을 정규방정식을 통해 구현할 수 있다. 이는 선형대수에 대한 이해가 부족하다면 어려울 수 있다. 이 방식에 대한 증명은 아래와 같다.



■ 3. 평가기준

회귀분석은 너무나 방대한 이론이며, 이에 대해서 실질적으로 다 설명할 능력이 없다. 여기서 회귀분석에 대해서 정확히 서술하지 않지만 기본적으로 유의해야 할 점과 평가방법 정도는 파악하는 것이 적절하다. 정확한 설명들은 검색하는 것을 추천한다.

✓ 회귀분석의 기본 가정

선형성 : 설명변수와 반응변수 간의 관계 분포가 선형의 관계를 가진다

독립성 : 설명변수와 다른 설명 변수 간에 상관관계가 적다.

잔차의 등분산성 : 잔차가 특정한 패턴을 보이지 않는다.

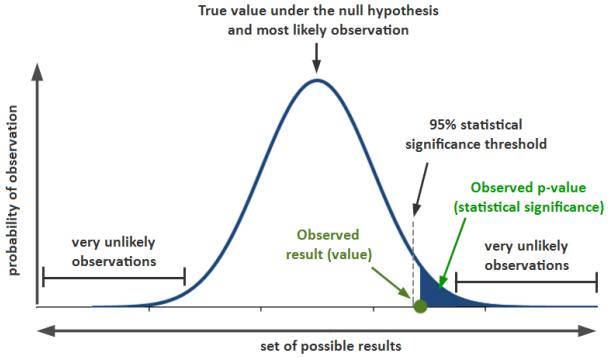
잔차의 정규성 : 잔차가 정규분포를 따른다.

이 가정들을 현실에서 모두 지키는 것은 어려운 경우가 많으며, 이를 파악하고 있다는 점은 필요하다. 이를 위해서 변수간 correlation을 계산하여, 같은 의미인 변수는 삭제해준다. 변수의 분포를 통해 정규분포를 가지고 있는지 확인한다.

평가기준은 크게 3가지 P-Value, RMSE, R Squared 가 존재한다.

이 외에도 많은 기준이 있으며, 본인이 생각하기에 이 3가지는 꼭 알아야한다.

1. P-Value(유의 확률) : **유의 확률**(有意 確率, 영어: significance probability, asymptotic significance) 또는 **p-값**(영어: *p*-value, probability value)은 귀무가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률이다



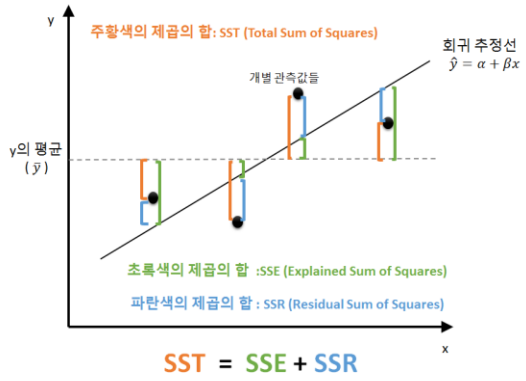
2. RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

이 값은 기존 MSE값에 루트를 씌워준 값이다. 루트를 통해, 예측값이 실제값과 얼마나 차이나는지 평가할 수 있다.

3. R Squared

R Squared을 이해하기 위해선 SST SSE SSR을 이해할 필요가 있다.



SST(**total sum of squares**) = 실제값 – 평균의 제곱 : 총 변동

SSE(**explained sum of squareds**) = 예측값 – 실제값의 제곱 : 설명할 수 없는 변동

SSR(**residual sum of squares**) = 예측값 - 평균의 제곱 : 설명할 수 있는 변동

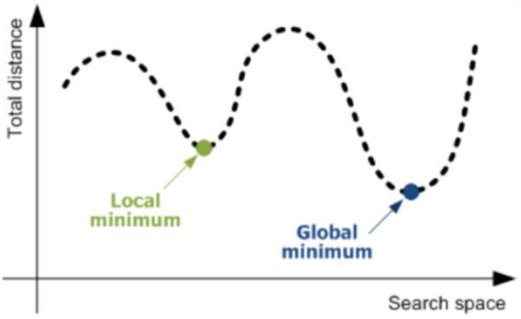
SST = SSR + SSE 이므로

R SQUARE = SSR/SST 이며 이는 설명할 수 있는 변동을 총 변동으로 나눈 것으로 총 변동 중 설명할 수 있는 비율을 나타낸다. 하지만 평가의 한 지표이므로 다른 부분과 함께 보아야한다.

● 3. 경사하강법

앞서 회귀분석의 경사하강법을 다뤘지만, 이 아이디어는 머신러닝 전반에서 쓰이는 최적화 방법이며, 조금 더 정확히 파악할 필요가 있어 추가한다.

경사하강법은 여러 종류의 문제에서 최적의 해법을 찾을 수 있는 매우 일반적인 최적화 알고리즘으로, 비용함수를 최소화하기 위해 반복해서 파라미터를 조정해가는 것이다.



이를 위해, 비용함수의 특징을 파악할 필요가 있다. 앞서 본 MSE를 보았을 때,

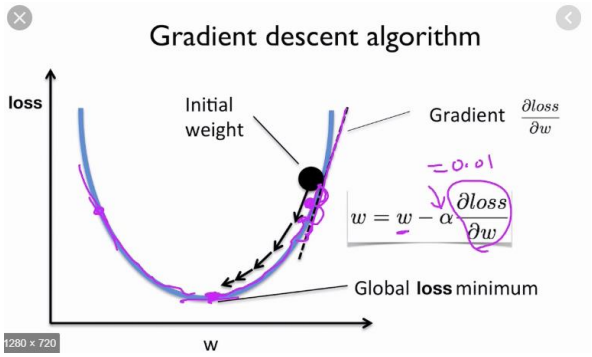
1. Convex하여, 미분값이 0이 되는 점이 존재해야한다.
2. 지역 최소값(Local Minimum)과 전역 최소값(Global Minimum) 중 전역최소값을 찾아가는 과정이다.
3. 비용을 최소화 해야 하므로, 비용은 양수값만 가진다.

   Ex) x^2-1 처럼, 비용의 최소가 음수라면 0인 값이 2곳이 존재하며, 미분이 0이 된 값이 -1이므로 오히려 비용이 늘어날 수 있다.

Cost 함수를 minimizing 하기 위한 알고리즘에 대해서 알아야하낟.
이 알고리즘의 이름은 Gradient descent algorithm 으로 앞서 봤던 y = wx + b 에서 w 와 b 를 찾는 것과 동일하며, 이를 업데이트 해주는 방식이다.
앞서 말한 파라미터가 최적화 된다는 것은, 앞서 봤던 비용이 작아지는 파라미터를 찾는 것이다.

Loss, cost function 은 회귀분석 기준으로 MSE이며 이는 (y실제값 - y예측값)^2 = (y실제값 – (wx+b))^2로, W에 대해서 2차 방정식이고, 이에 대해서 미분하여 하락하는 방향으로 이동하면 된

다.

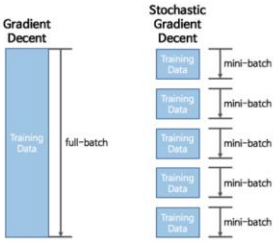


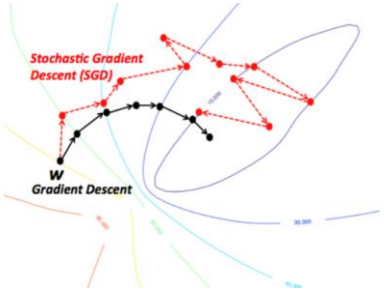
즉 새로운 W 는 기존 W – Learning rate * loss의 w 미분 값이다.

Learning rate가 있는 이유는 적절히 이동하여 로컬 미니멈을 피하고 글로벌 미니멈을 찾아가는 것을 조절하기 위한 것이다. 보통 0.001에서 0.0001 사이 값으로 설정한다.

이러한 경사하강법에는 어떤 데이터를 가지고 하느냐에 따라, 배치 경사하강법, 확률적 경사하강법, 미니배치 경사하강법이 존재한다.

- **배치 경사하강법**은 전체 훈련 세트 X에 대해서 계산하는 것으로, 특성 수에 무관하지만 매우 큰 데이터 셋에서 너무 느린 문제를 가지고 있다.
- **확률적 경사 하강법**은 매 스텝에서 딱 한 개의 샘플을 무작위로 선태해 그 하나의 샘플에 대한 그레디언트를 계산한다. 알고리즘 속도가 매우 빠르지만, 샘플에 따라 불안정 하다는 문제를 가지고 있다.
- **미니배치 경사 하강법**은 전체 데이터에서 미니배치라 부르는 임의의 작은 샘플에 대해 그레디언트를 계산한다. 미니배치 사이즈에 따라서, 확률적 경사 하강법의 불안정성이 어느정도 완화되며, 빠른 속도를 유지 할 수 있다,

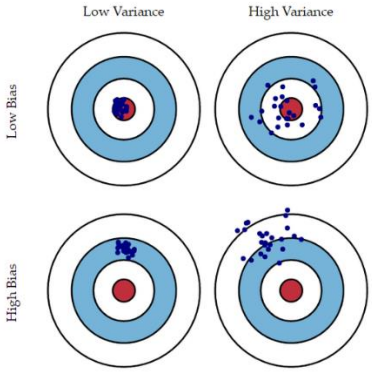




● 4. 학습곡선과 편향과 분산

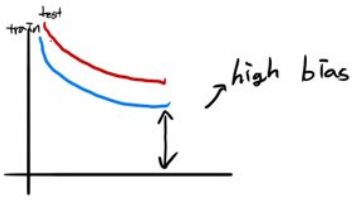
학습곡선과 편향과 분산을 알아야하는 이유는 모델의 훈련이 과소적합 혹은 과대적합 되었는지 의사결정을 할 수 있는 능력을 가질 수 있다. 예를 들어, 훈련 데이터에서 정확히 맞추더라도 실제 데이터에서는 예측력을 가지지 못하는 과대적합을 가질 수 있기 때문이다.

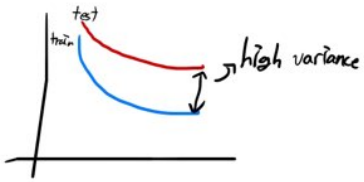
✓ 편향과 분산



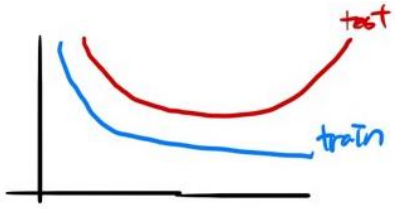
편향(bias) : 실제값 – 예측값의 평균의 제곱이므로 예측값이 실제값과 얼마나 다른지 판단

분산(Varaince) : 예측값 – 예측값의 평균의 제곱이므로 예측값의 분포를 파악





High bias는 오류가 큰 것을 나타내며, High Variance는 훈련과 검증 데이터의 오차의 차이가 크다

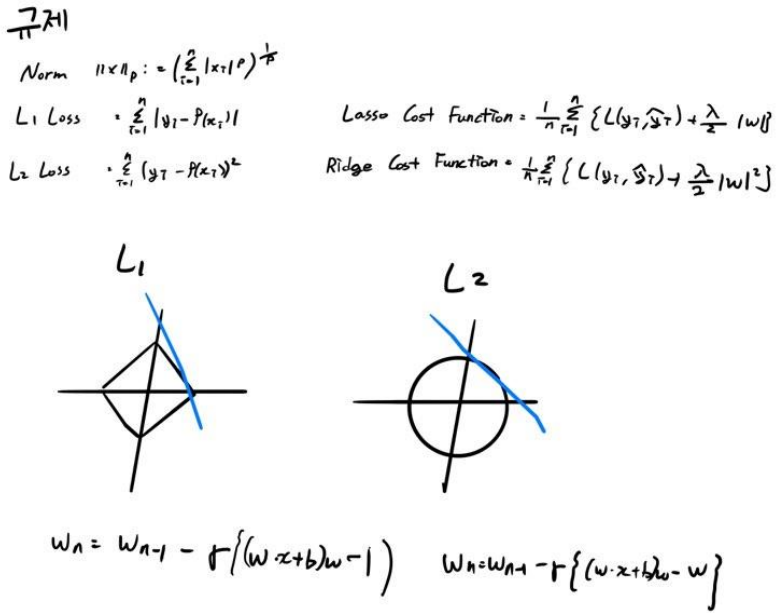


이에 대하여, 학습 곡선을 그렸을 때 변곡점에서 학습을 멈추는 것이 overfitting을 방지하는 가장 대표적인 방법이다.

● 5. 규제(Regularization)

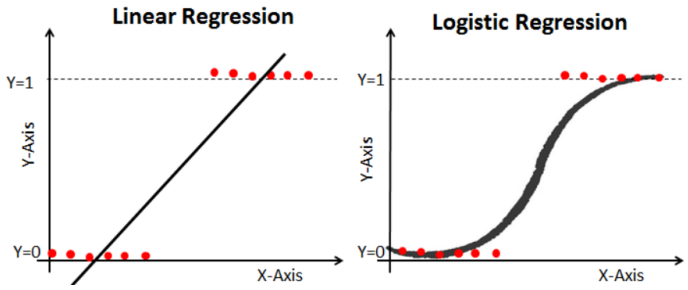
규제는 모델이 복잡해지지 않도록 모델 복잡도에 패널티를 부여하여, 특정 가중치가 너무 과도하게 커지는 것을 방지한다. 이는 Overfitting을 예방하고 일반화 성능을 높이는 도움을 주는 것이다. 이러한 방법은 비용함수에 규제 항을 추가하면 가능하다.

Lasso와 Ridge가 존재하며, Lasso는 L1 규제, Ridge는 L2 규제를 사용한다.



주목할 점은 Lasso의 경우 가중치를 0으로 만들 수 있기 때문에, 변수선택의 효과도 가지고 있다는 것이다. 이에 반해 Ridge는 가중치가 0에 가까워 질지라도 실제로 0이 되지는 않는다.

● 6. 로지스틱 회귀



종속변수 Y가 **범주형(categorical)** 변수일 때는 다중선형회귀 모델을 그대로 적용할 수 없다. 이러한 문제에 대해 로지스틱 회귀 모델은 범주형 종속변수에 대해서 적절한 알고리즘이다.

기존 회귀분석에 대해서 이를 유도하는 과정에서는 odds와 sigmoid function에 대한 이해가 필요하다.

$$odds = P(A)/P(A_c) = P(A)/1-P(A)$$

$$\frac{p(x)}{1-p(x)} = e^a$$

$$p(x) = e^a \{1-p(x)\}$$
$$= e^a - e^a p(x)$$

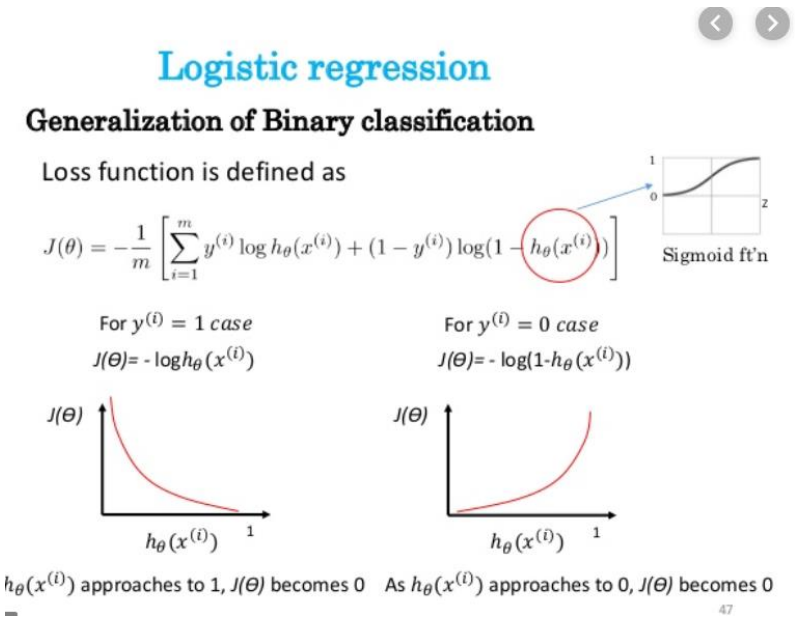
$$p(x)\left(1+e^a\right) = e^a$$

$$p(x) = \frac{e^a}{1+e^a} = \frac{1}{1+e^{-a}}$$

$$\therefore P(Y=1|X=\vec{x}) = \frac{1}{1+e^{-\vec{\beta}^T\vec{x}}}$$

그렇다면 분류의 비용함수 Cross entropy를 파악할 필요가 있다.

- 1. Cross entropy



로지스틱 회귀분석의 비용함수인 cross entropy 이다.

1. Non-negative 하다. : 시그마 내부의 모든 항은 항상 음수이며, 시그마 바깥의 음의 부호로 인해 수식 전체의 값은 언제나 0 이상이다.

2. 모든 training data input 들에 대해서 만약 h(x)의 값이 y 에 가깝다면, cross-entropy 의 값은 0 에 가까워질 것이다.

4. convex 하여 최적최소값이 존재한다.

이러한 이유로 cross entropy 를 자주 이용하며, 이를 최소화 하기 위해선 이전 mse 를 미분하듯 미분하여야 한다.

chain rule 을 사용해서 w 에 관해서 미분해보자.

ln(x)의 미분 값이 1/x 이므로

$$\frac{\partial C}{\partial w_j} = -\frac{1}{n}\sum_x \left(\frac{y}{\sigma(z)} - \frac{(1-y)}{1-\sigma(z)}\right)\frac{\partial \sigma}{\partial w_j}$$
$$= -\frac{1}{n}\sum_x \left(\frac{y}{\sigma(z)} - \frac{(1-y)}{1-\sigma(z)}\right)\sigma'(z)x_j$$

여기서 sigmoid function을 미분하면

$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$

이기 때문에 식이 매우 아름답게 정리된다!

$$\frac{\partial C}{\partial w_j} = \frac{1}{n}\sum_x x_j(\sigma(z) - y).$$

둘의 차이가 클수록 변화의 기울기도 커진다는 의미가 된다.

■ 2. 소프트맥스 회귀

소프트맥스 회귀는 다분류의 경우 그 분류에 해당할 확률을 나타내는 것이다.

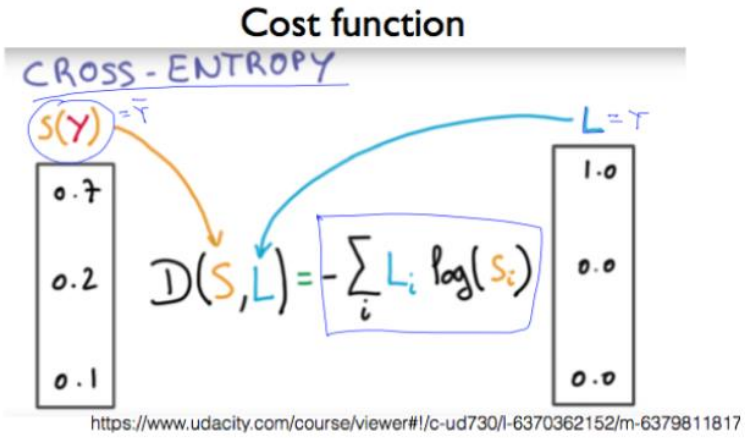
$$f(\vec{x})_i = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \quad \text{for } i = 1, \ldots, K$$

다중분류에서 우수한 성능으로 분류한다. Sigmoid에 대비하여 평가하자면

분류시 100%안으로 예측값들을 정규화 해준다.

자연함수의 특징 상 차이를 더욱 더 크게 만들어 준다.

Cost function에 들어갔을 때 앞선 이유들로 차이를 더 뚜렷하게 만들어준다.



4장 연습문제 정답

1. 수백만 개의 특성이 있는 훈련세트는 확률적 경사하강법이나 미니배치 경사하강법이 적절하다.
2. 훈련 세트에 있는 스케일이 다르다면 비용함수는 길쭉한 타원 모양의 그릇 형태가 되고, 경사하강법 알고리즘이 수렴하는데 오랜 시간이 걸린다.
3. 로지스틱 회귀모델의 비용함수는 볼록 함수 이므로 경사하강법이 훈련될 때 지역최소값에 갇힐 가능성이 없다.
4. 최적화할 함수가 볼록함수이고 학습률이 크지 않다고 가정하면 모든 경사 하강법은 전역최소값에 도달한다. 학습률을 점진적으로 감소시키지 않으면 sgd와 미니배치 gd는 최적점에 수렴하지 못할 것이다.
5. 에포크마다 검증 에러가 지속적으로 상승한다면 한가지 가능성은 학습률이 너무 높고 알고리즘이 발산하는 것일지 모른다. 훈련 에러도 올라간다면 학습률을 낮추어야한다. 훈련 에러가 올라가지 않으면 과대적합으로 훈련을 정지
6. 무작위성 때문에 sgd나 미니배치 gd 모두 매 훈련 반복마다 학습의 진전을 보장하지 못한다. 검증에러가 상승시 바로 멈춘다면 최적점에 도달하기 전에 일찍 멈추게 될지도 모른다. 더 나은 방법은 정기적으로 모델을 저장하고 오랫동안 진전이 없을 때, 가장 좋은 모델을 복원하는 것이다.
7. 확률적 경사 하강법은 한번에 하나으 | 훈련 샘플만 사용하기 때문에 훈련 반복이 가장 빠릅니다. 그래서 가장 먼저 전역 최적점 근처에 도달합니다. 그러나 훈련시간이 충분하면 배치경사 하강법만 실제로 수렴할 것입니다.
8. 검증 오차파가 훈련 오찹보다 훨씬 더 높으면 모델이 훈련 세트에 과대적합되었기 때문일 가능성이 높습니다. 이를 해결하는 첫 번째 방법은 다항 차수를 낮추는 것입니다. 자유도를 줄이면 과대적합이 훨씬 줄어들 것입니다. 두번째 방법은 모델을 규제하는 것입니다. 비용함수에 l2패널티나 l1패널티를 추가합니다. 세번째 방법은 훈련세트의 크기를 증

가시는 것입니다.

9. 훈련 에러와 검증에러가 거의 비슷하고 매우 높다면 모델이 과소적합되었을 간으성이 높습니다. 즉 높은 편향을 가진 모델입니다. 이때는 규젯하이퍼파리미터 a를 감소시켜야 합니다.

10. 규제가 있는 모델이 일반적으로 규제가 없는 모델보다 성능이 좋습니다. 라쏘 회귀는 l1 패널티를 사용하여 가중치를 완전히 0으로 만드는 경향이 있습니다. 희소모델을 만든다. 또한 자도으로 특성 선택의 효과를 가지므로 의미있다. 라쏘가 특성이 너무 많으면 불규칙하게 행동하므로 엘라스틱넷이 라쏘보다 더욱 선호된다.

11. 실외와 실내, 낮과 밤에 사진을 구분하고 싶다면 이 둘은 배타적인 클래스가 아니기 때문에 두개의 로지스틱 회귀분류기를 훈련시켜야 합니다.

12.

13.

---

[i] https://taeoh-kim.github.io/blog/cross-entropy%EC%9D%98-%EC%A0%95%ED%99%95%ED%95%9C-%ED%99%95%EB%A5%A0%EC%A0%81-%EC%9D%98%EB%AF%B8