4 주차 조별보고서 (Default)	
작성일: 2019 년 9 월 27 일	작성자: 최진성
조 모임 일시: 2019 년 9 월 27 일 9 교시	모임장소: 학교 앞 카페
참석자: 위성조, 이충현, 최진성, 이재은, 김영연	조원: 위성조, 이충현, 최진성, 이재은, 김영연

구 분	내 용
학습 범위와 내용	Machine Learning 1, Introduction Machine Learning 2, Linear Regression Machine Learning 4, Multiple Features
논의 내용	저희 Default 조는 사전에 개인 레포트/질문을 작성하여, 금요일 9 교시에 만나 질문에 대하여 서로 대답하고 토론하는 방식으로 회의를 진행하였습니다.

논의 내용	**Q. 수업 중 공동으로 주어진 문제에 대한 논의. (Lecture 1. Introduction 17p, 30p)** **Q1. 당신은 회사를 운영하고 있습니다. 그리고 당신은 다음의 두 문제를 해결하기 위해 러닝 알고리즘을 개발하려고 합니다.** 1) 당신은 단일 품목으로 구성된 대량의 재고를 가지고 있습니다. 당신은 향후 3 개월동안 이 품목이 얼마나 팔릴지 예상하고 싶습니다. 2) 당신은 개별 고객의 계정을 검사하여, 각 계정이 해킹되었거나 문제가 발생했는지 여부를 알 수 있는 소프트웨어를 원하고 있습니다. **이 문제를 분류(Classification) 문제로 대해야 할까요, 회귀(Regression)문제로 대해야 할까요?** **A1.** 1)의 경우, 기존에 존재하는 데이터를 이용하여, 연속적인 값을 추측하는 문제이므로 회귀문제로 봐야 하며, 2)의 경우, 개별 데이터가 해킹되었는지 아닌지(1인지 0인지) 구분하는 이산 값을 추측하는 문제이므로 분류 문제로 보아야 한다. **Q2. 아래의 문제들에서, 어떤 문제들을 비지도학습을 이용하여 처리하면 좋을까요?** 1) 받은 이메일이 스팸인지 아닌지 분류하는 문제. 2) 인터넷에서 찾은 뉴스들을 같은 이야기를 주제로 하는 그룹으로 묶는 문제. 3) 고객 데이터가 담긴 데이터베이스를 활용하여, 자동적으로 마켓 세그먼트(market segment)를 찾고, 고객들을 다른

	세그먼트들로 분류하는 문제 4) 기존에 당뇨병이 있는 환자와 없는 환자들의 진단 데이터를 활용하여, 새로운 환자의 진단 데이터를 기반으로 환자가 당뇨병에 걸렸는지 안 걸렸는지 분류하는 문제 **A2.** 모든 문제를 비지도학습으로 처리할 수 있으나, 분류된 그룹이 무엇인지가 중요한 경우 지도학습을 사용하는 게 좋고, 분류된 그룹이 무엇인지보다 그룹화하는 것이 더 중요한 경우 비지도학습을 쓰면 좋다고 할 수 있다. 따라서 1) - 스팸 문제와 4) - 당뇨병 환자 진단의 경우 지도학습, 2) - 뉴스 그룹화와 3) - 고객 분류 문제는 비지도학습이 더 효과적이라고 생각할 수 있다. **Q. 수업 내용 중 개인이 제시한 문제에 대한 논의.** **Q3. 정규방정식을 보면 역행렬이 존재한다. 그러나 역행렬이 없다면 어떻게 해야 하는가? – 영연** **A3.** 특징들이 서로 선형 종속 관계, 다시 말해 중복된 특징이 존재하는 경우, 혹은 dataset의 크기에 비해 너무 많은 특징들을 사용하는 경우에 발생한다. 이를 해결하기 위해선 전자의 경우에는 중복된 특징을 찾아 하나를 삭제하고, 후자의 경우에는 열 n의 크기가 특징의 개수임으로 열의 개수를 줄이거나 정규화, 또는 데이터를 늘림으로써 해결할 수 있다. 어떤 경우에 해당하는지 모호한 경우에는 Octave라는 프로그램에서 제공하는 pinv함수를 이용하여 손쉽게 해결할 수 있

다. pinv 함수란 의사역행렬(pseudoinverse)을 반환하는 함수로 특이값해(SVD)를 사용하여 계산된다. 계산 방식에 대해서는 https://bskyvision.com/256에 자세한 설명이 되어있어 훑어보았다.

Q4. Cost function 에서 우리는 Linear Regression 에서 배웠고 Gradient Descent 를 써서 손실을 최소화했다.

그렇다면 볼록, 오목함수와 같은 Logistic Regression 은 어떻게 Cost Function 을 매겨야 할지 궁금하다. – 충현

A4.

로지스틱 회귀 함수에서 사용하는 함수는 전형적인 비선형 함수이다.

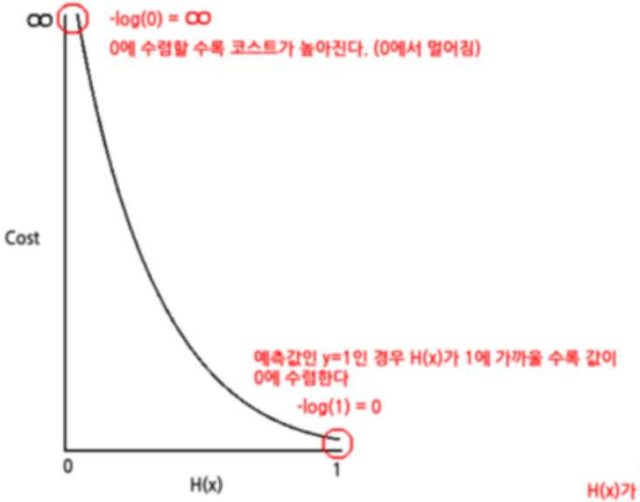
$$h(\theta) = \frac{1}{1 + e^{-\theta^T x}}$$

따라서 이러한 비선형함수를 이용해서 선형회귀의 Cost Function에 대입하면 울퉁불퉁한 그래프가 그려진다. 그 경우 사하강법을 이용할 때 문제가 생기므로 다른 종류의 Cost Function을 사용해야 한다.
우선 첫 번째로 Cost Function을 다른 식으로 변형한다.

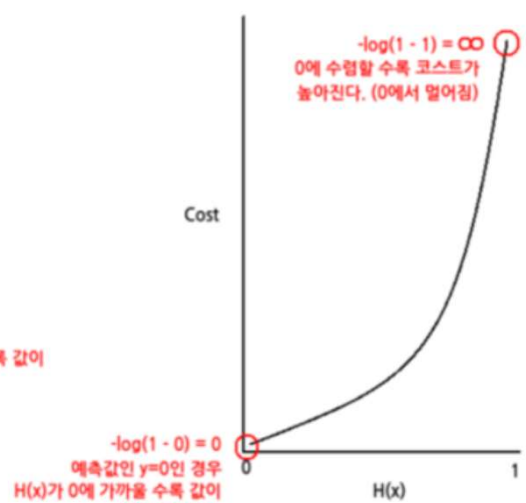
$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta x^{(i)} - y^{(i)}\right)^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta x^{(i)} - y^{(i)}\right)^2$$

$$\mathrm{COST}\left(h_\theta x^{(i)}, y^{(i)}\right) = \frac{1}{2}\left(h_\theta x^{(i)} - y^{(i)}\right)^2$$

	$$J(\theta) = \sum_{i=1}^{m} COST\left(h_\theta x^{(i)}, y^{(i)}\right)$$ Logistic Regression에서는 곡선을 보다 완만하게 만들기 위하여 변형된 cost function에 log를 취해준다. $$COST\left(h_\theta x^{(i)}, y^{(i)}\right)$$ Y가 1일 때 → - log($h_\theta(x)$) Y가 0일 때 → - log(1-$h_\theta(x)$) Logistic Regression의 Cost Function은 위의 수식처럼 y의 값이 0과 1인 경우 두가지로 분류된다. 이를 그래프로 그리면 다음과 같이 나타나고 cost의 값이 0에 수렴할수록 일치가 잘 된 것이다. (이항분류 이므로 y의 값의 범위는 0~1에 한정된다.)



$[-log(H(x))$ 그래프]



$[-log(1 - H(x))$ 그래프]

그러나 이런 식으로 y의 값이 고정된 것이 아닌 값에 따라 수식이 갈라지면 실제 프로그래밍을 할 때 코딩이 복잡해지기 때문에 식을 다음과 같이 압축하여 하나로 합치는 것이 좋다.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]$$

	Q5. Gradient Descent를 이용하여 Cost Function의 최저값을 구할 때 아래의 두 상황의 경우에는 어떤 결과가 도출되는지, 만약 문제 상황이라면 어떻게 해결해야 하는지 궁금하다. – 진성 1) 최저점의 y 좌표가 모두 같을 경우 y = x^2 (-1 ≦ x ≦ 1) 2) 최저점이 끝도 없이 내려가는 형태일 경우 y = x^3 **A5.** 1)의 경우엔 기존의 Linear Regression의 Cost Function을 사용하여도 문제 없다. 다만 2)의 경우에는 Sigmoid 함수의문제점이 발생한다. 3차 함수와 같은 함수가 Cost Function의 Case로 들어오게 되면 Sigmoid 문제점이 발생하는데, 이는 기존 함수의 출력값을 받는 방법을 그대로 이용할 수 없기 때문에 발생하는 문제다. 일반적으로 함수의 처리 방법은 하나의 네트워크에 입력값을 줘서 Loss인 출력값을 받아 미분하여 Gradient를 계산하는 것, 즉 Back propagation작업을 거친다. 그러나 3차 함수와 같이 Sigmoid 함수의 경우 값이 커질수록 Gradient가 점점 0으로 소실되어 전달받을 Gradient가 없어지는 셈이다. 이를 방지하기 위해 Relu함수를 사용한다. Relu함수는 음수인 값의 최댓값은 무조건 0으로 처리하는 대신에 특정 변수 a를 곱한다. 이에 따라 Gradient가 0일 경우가 없어지게 되어 Back Propagation 작업이 원활하게 돌아가게 할 수 있다.
질문 내용	조원들이 질문하였던 것을 모두 해결하였기 때문에, 이번 주차는 질문 내용이 없습니다.

기타	

참고

조 운영 지침

1.매주 1 회 정해진 시간과 장소에 모여서 1 시간 정도의 조모임을 갖는다.
2.조장은 모임 전에 학습할 범위를 조원들에게 통보한다.
3.각 조원은 학습 범위 내의 교재와 강의 자료를 공부한 후에 이해한 내용과 이해하지 못한 내용을 각각 간단하게 정리하여 개별보고서를 작성한다. (1-2 쪽으로 충분함) 작성한 개인 보고서는 모임 전에 모든 조원에게 전송한다.
4.그 모임의 회의 진행은 순번을 정하여 돌아가면서 진행하고 해당 순번은 조별모임 한 후에 조별보고서를 작성하여 다음 수업시간 전에 과목 웹 페이지에 게시를 한다.
5.조별 모임에 참석하지 않는다든지 보고서를 작성하지 않는다든지 혹은 지각 등의 조의 단합을 저해하는 조원은 조원들 스스로 학기 초에 정한 규정에 의하여 처리할 수 있다. (벌금 부과나 조 퇴출 등) 이러한 규정들은 조가 결정된 후에 서로 조별로 협의하여 규정을 만들어 제출하며 규정은 계속 개정할 수도 있다. (규정을 소급적용할 수는 없다.)
6.조별모임을 원하지 않는 사람이나 퇴출된 학생은 다른 조에 동의를 얻어서 합류하거나 보고서 작업을 혼자 진행한다. (조원의 최대 숫자는 학기 초에 정해진다.)
7.개인 보고서와 조별 보고서 모두 "자료조사" 혹은 교재 내용을 요약 정리하는 것에 중점이 있는 것이 아니라 자신이 혹은 조원들이 잘 모르겠는 것들 이해되지 않는 것들이 무엇인지를 파악하는 데 중점을 둔다.
8.작성된 조별 보고서는 수업시간 혹은 과목 홈페이지 게시판을 통하여 설명될 것이다.

첨부 개인 레포트는 아래에 있음.

이름	김영연	학번	201500629

구분	내용
학습 범위	Lecture 1 introduction Lecture 2 linear regression Lecture 4 multiple features
학습 내용	● Supervised learning (지도 학습) 어떠한 기계학습 알고리즘에 데이터 집합을 제공할 때 이 데이터에 정답이 있는 상태.(사전에 데이터를 분류) ● Unsupervised learning (비지도 학습) ● Training set: supervised learning을 수행할 때 주어진 데이터가 있어야 하는데 이것을 training set이라고 한다. ● Hypothesis $h_\theta(x) = \theta_0 + \theta_1 x$ (θ_i: parameter, x: input) //parameter에 따라 기울기가 변한다. Hypothesis를 통해 그려 놓은 직선을 이용해서 잘 예측하려면 직선이 자료들과 얼마나 가까운지를 봐야한다. 즉, Hypothesis의 값과 실제 output의 값, y와의 오차가 작으면 작을수록 정확히 예측할 수 있다. 직선과 주어진 데이터의 거리를 구하기 위한 식은 $(h_\theta(x) - y)^2$이고 이 값이 작으면 작을수록 좋은 직선이다. 주어진 training set은 여러 개이고 모든 training set의 결과 값과 직선의 차이를 모두 합하고 평균을 구하기 위해 training set의 크기를 나눠주면 cost function이 된다.

- Cost function

  $J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$(h: hypothesis, y: output, m: training set size, i: training set index)
- Gradient descent (경사하강법)

  경사하강법은 비용함수를 최소화하기 위한 기법이다. 즉, cost function의 최소 값을 찾는 것이다.

  $\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta_0, \theta_1)$ (for j=0 and j=1)

  경사하강법에서 중요한 것은 다음과 같다.

  Correct: Simultaneous update

  $$\text{temp0} := \theta_0 - \alpha\frac{\partial}{\partial\theta_0}J(\theta_0, \theta_1)$$
  $$\text{temp1} := \theta_1 - \alpha\frac{\partial}{\partial\theta_1}J(\theta_0, \theta_1)$$
  $$\theta_0 := \text{temp0}$$
  $$\theta_1 := \text{temp1}$$

  파라미터의 계산이 종료된 뒤 일괄적으로 θ값이 갱신되야 한다.

  α의 값이 작을 경우 경사하강법은 느릴 수 있다.

  반대로 값이 클 경우 overshoot할 수 있다
- Feature scaling

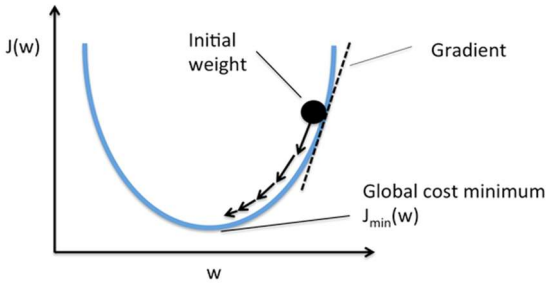
  모든 feature을 대략 -1<=x<=1사이로 만드는 것
- Mean normalization

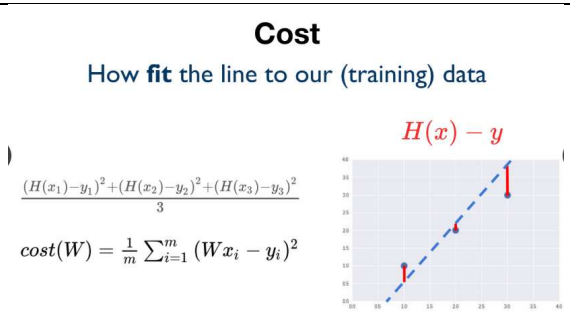
  x_i를 $x_i - \mu_i$로 대체하는 것
- Normal equation

	$\theta = (X^T X)^{-1} X^T y$. ● 경사 강하법의 장단점 장점: feature의 수가 많을수록 유리하다 단점: 적절한 알파 값을 결정해야 한다. 계속 미분을 해야 하므로 수많은 반복 수행이 필요하다. ● 정규 방정식의 장단점 장점: 알파 값을 결정할 필요가 없다. 반복하여 수행하지 않는다. Feature scaling 작업이 필요하지 않다. 단점: feature의 수가 많을수록 불리하다.
질문 내용	정규방정식을 보면 역행렬이 존재한다. 그러나 역행렬이 없다면 어떻게 해야 하는가?

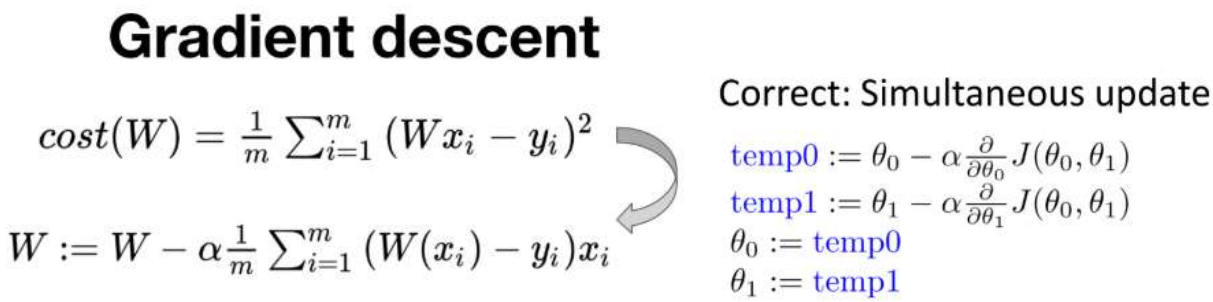
이름	**이충현**	**학번**	**201402665**
구분	**내용**		

학습 범위	앤드류 머신러닝 Lecture-1 앤드류 머신러닝 Lecture-2 앤드류 머신러닝 Lecture-4
학습 내용	<Lecture-1 머신러닝이란??> ● 인공지능의 한 분야로, 컴퓨터가 학습할 수 있는 알고리즘과 개술을 개발하는 분야를 말한다. ● Supervised Learning -> 직역하면 지도된 학습이다. 레이블링을 통한 학습(Labeled). Input으로 데이터를 주면 Output으로 학습된 결과물을 보여준다. 따라서 학습을 하기위한 Training Dataset가 필요하다. Unsupervised Learning ->말 그대로 지도되지 않은 학습. 레이블링 없이 데이터를 가까운 기준으로 묶을 때 사용된다. 딱히 답은 없다. 그래서 Input만 주고 Output은 그 상황이나 환경에 반영돼서 나온다. ● Cocktail Party Problem -> 각종 소음이 난무하는 곳에서도 자신이 관심있는 단어가 또렷이 들리는 현상이다. 주변 환경에 개의치 않고 자신에게 의미 있는 정보만을 선택적으로 받아들이는 것을 말한다. 인공 신경망을 수행할 때, 전체가 섞인 스펙트로그램을 Input으로 주어지면 거기에 사람의 목소리에 해당하는 부분의 스펙트로그램을 Output으로 내는 작업이므로 이는 Supervised Learning이라고 할 수가 있다. <Lecture-2 Linear Regression>





- 손실을 최소화하기 위한 방법에는 경사하강법이 있다. 경사하강법이란 쉽게 말해 Cost Function의 기울기가 최소가 되는 값을 찾는 알고리즘으로 임의의 한점에서 기울기를 구한 Cost function, 다음 기울어진 방향으로 계속 이동하여 이 최소가 되는 최적의 Parameter를 찾으면 된다. 경사도를 구하기 위해서는 미분이 필요하게 되는데 아래와 같이 변경한다. 그리고 경사 하강법의 수식을 가져온 후에 Cost Function의 식을 대입한다. 따라서 아래 그림과 같은 최종 식이 도출되게 된다. 해당 식을 여러 번 실행시키는 것이 경사 하강법의 핵심이다. 이를 통해서 Cost가 최소화된다. 주의할 점은 기울기 변화량이 동시에 업데이트가 되어야 한다. (a는 Learning Rate이다.)

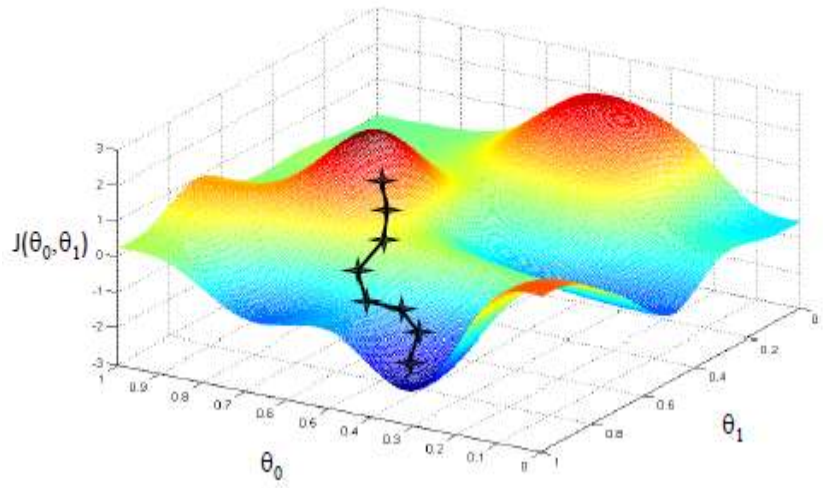


	● Cross Entropy(교차 앤트로피)는 주로 범주형 데이터에 널리 사용되는 또 다른 손실 함수이다. 교차 앤트로피는 두 분포 사이의 유사성을 재므로서 딥러닝에서 사용되는 모델은 보통 각 클래스의 확률값을 계산하므로 실제 클래스와 모델에서 제시한 클래스를 비교할 수 있다. 두 분포가 가까울수록 교차 앤트로피의 값은 더 작아진다. • 통신 이론의 엔트로피: $H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$ <Lecture-4 Multiple Features> ● Feature Scaling은 Raw Data를 전처리하는 과정이다. 주로 표준화, 정규화 방식으로 전처리한다. ● Learning Rate는 한 번 학습 시 얼마만큼 학습해야 하는 양(a알파)을 의미한다. 학습률이 너무 크면 큰 값을 반환하고, 너무 작으면 거의 갱신되지 않고 학습이 끝나버린다. $\alpha(t) = \alpha(0)(1.0 - t/rlen)$ ● 경사 하강법과 정규 방정식의 차이점은 정규 방정식은 반복하여 수행하지 않고 learning rate를 매길 필요가 없다. 반면에 경사 하강법은 적절한 learning rate를 설정하는 대신, features수가 많을수록 좋다.
질문 내용	Cost function에서 우리는 Linear Regression에서 배웠고 Gradient Descent를 써서 손실을 최소화했다. 그렇다면 볼록, 오목함수와 같은 Logistic Regression은 어떻게 Cost function을 매겨야 할지 궁금하다.

이름	위성조	학번	201402033

구분	내용

학습 범위	Lecture 1 introduction Lecture 2 linear regression Lecture 4 multiple features
학습 내용	Supervised learning(지도학습) – 정답이 주어져 있는 기계학습이며, Regression과 Classification이 있다. Regression(회귀) – 연속적인 변수의 값을 예측하는 것 Classification(분류) – 이산 변수의 값을 예측하는 것 Unsupervised learning(비지도학습) – 정답이 주어져 있지 않으며, 데이터를 분류하기 위한 기계학습 Cocktail party problem은 녹음 데이터만을 제공하여, 그 속에서 패턴을 추출하는 것이므로 비지도학습의 일종이다. 1차원 함수로 이루어진 가설 $\mathrm{h\theta(x)} = \theta 0 + \theta 1 * x$ 머신 러닝 : $\theta 0, \theta 1$을 선택하여 $\mathrm{y} - \mathrm{h\theta(x)}$가 0에 가깝도록 하는 것 이를 위해 Cost function $\mathrm{J(\theta 0, \theta 1)} = \frac{1}{2m}\sum_{i=1}^{m}(h\theta(x^{(i)}) - y^{(i)})^2$의 값을 최소화 하는 $\theta 0, \theta 1$을 찾는다 Gradient descent – Cost function의 값을 최소화시키는 $\theta 0, \theta 1$의 값을 찾기 위해 임의의 $\theta 0, \theta 1$값에서 시작하여, Cost 함수가 작아지는 방향으로(수치해석적 방법으로) $\theta 0, \theta 1$을 변화시킨다.



연산의 끝을 설정하기 위해 업데이트량이 일정 수준 이하로 떨어지는 경우 연산을 중지하거나, 최대 연산횟수를 설정한다.

$\theta 0, \theta 1$의 변화량 $\propto$를 너무 작게 잡을 경우, 경사하강이 너무 늦게 진행될 우려가 있고, 반대로 너무 크게 잡을 경우, 최소값에 수렴하지 못하고 발산할 가능성이 있다.

경사하강법의 경우, Local minimum은 찾을 수 있으나, 그것이 Global minimum인지 보장하는 것이 불가능하므로, 여러 랜덤 위치에서 시작하여, 가장 낮은 값을 쓰는 형태로 운용한다.

Feacher Scaling – 모든 특징들을 $-1 \leq x_i \leq 1$ 의 범위에 오도록 만드는 것.

집값을 예측할 때, 변수를 면적과 침실의 개수로 놓으면, 면적은 0-2000 feet^2, 침실의 개수는 1-5개의 범위를 가지게 되는데,

(범위는 가변적) 2000과 5를 숫자 그대로 처리하기에는 문제가 있으므로, 면적은 2000, 침실의 개수는 5로 나누어 0에서1까지의 범위로 단위를 비슷하게 만들어 준다

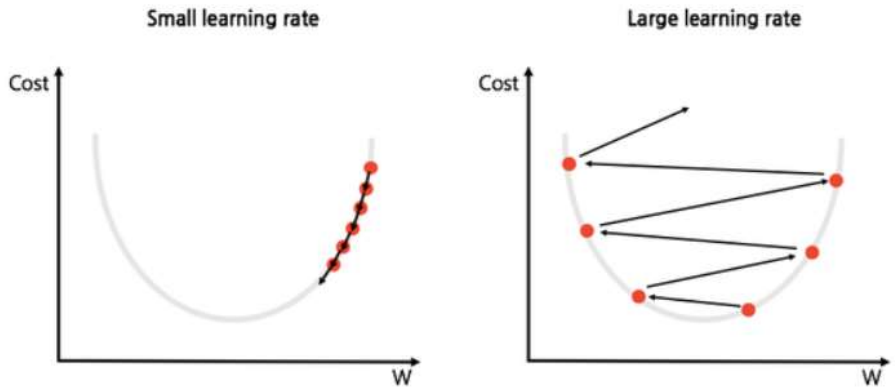
Normal equation(정규방정식) – 분석적으로 $\theta 0, \theta 1$값을 찾는 것

	$\theta = (X^T X)^{-1} X^T y$ 공식을 이용하여 구한다. 각 θ에 대하여 편미분을 하여 값을 구하기 때문에 계산을 여러 번 반복할 필요 없이 한번에 해답을 구할 수 있으나, n이 매우 큰 경우 너무 느려서 사용이 거의 불가능하다 //이에 반해 경사하강법은 n이 크더라도 준수한 성능을 보인다.
질문 내용	

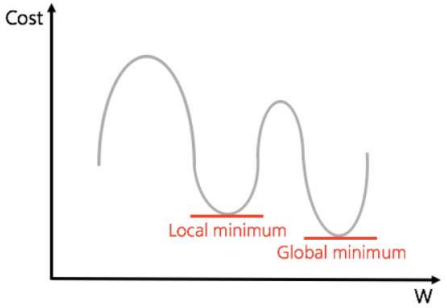
이름	**최진성**	**학번**	**201403474**
구분	**내용**		

학습 범위	Machine Learning 1 ~ 4
학습 내용	Machine Learning 1, Introduction Supervised Learning – Labeled Learning, Regression과 Classification의 방법이 있다. Regression - 이전까지 입력된 연속적인 트레이닝 데이터로 이후에 나올 변수의 값을 예측 Clasification – 불연속적인 값을 특징에 따라 분류하여 데이터를 판별 Unsupervised Learning – Unlabeled Learning, 데이터를 가까운 기준으로 모아서 사용 Ex : 검색어를 통한 기사 검색, 유전자 검사를 통한 질병 검사 등 Cocktail Party Problem – 여러 Speaker들 속에서도 자신이 원하는 이야기만 골라서 들을 수 있는 효과. 많은 데이터들 속에서 특정 데이터만과 연관된 데이터를 집합으로 구분하는 것이므로 Unsupervised Learning의 예시로 볼 수 있음. Machine Learning 2, Linear Regression Model Regression – Supervised Learning 중 Regression은 Traning set에서 Learning Algorithm을 통해 Hypothesis를 만드는 것인데, Hypothesis란 기존의 데이터를 토대로 예측 가능한 데이터를 만드는 최종 함수를 의미한다.

	Cost Function – Hypothesis를 만들 때 반복되는 학습을 통해 오차를 줄여야 하는데, 이 때 사용하는 것이 Cost Function이다. 기본 식은 hθ(x)=θ0+θ1*x이며 최종 목표는 yi = hθ(xi)를 줄이는 J(θ0, θ1) 값이 최소가 되는 최적의 기울기, 즉 θ0, θ1값을 찾는 것이다. Gradient Descent – Cost Function의 기울기가 최소가 되는 값을 찾는 알고리즘으로, 임의의 한 점에서 기울기를 구한 다음 기울어진 방향으로 계속 이동하여 Cost Function의 기울기가 최소가 되는 Parameter를 찾는 함수이다. Gradient Descent는 ∝(Learning Rate)의 값을 조절함에 따라 효과가 달라지는데, ∝가 작으면 작을수록 Cost Function의 기울기가 최소가 되는 값을 찾는데 너무 많은 시간과 자원이 들어가서 효율이 급격히 나빠지고, 반대로 ∝가 커지면 커질수록 경사면을 타고 올라가 발산해버리는 문제가 발생한다.



보통 적합한 Learning Rate를 정하는 방법은 어느 한 수치를 정해두고 귀납적으로 추론하여 범위를 좁혀나가는 방법을 이용한다. Gradient Descent에는 몇 가지 문제점이 있는데, 대표적으로 Local Minimum의 값과 Global Minimum의 값이 다를 수 있다는 점이다.



이 경우 특정 위치에서 시작하는게 아닌, 무작위 위치에서 시작하여 경사하강을 통해 최저점을 찾은 뒤, Local Minimum 값들 중에서 가장 낮은 구간에 있는 Global Minimum 값을 찾아 출력하는 방법을 이용하면 된다.

Machine Learning 4, Multiple Features

Multivariate Linear Regression –
Hypothesis를 구하는 기본 형태는 hθ(x)=θ0+θ1*x이다. 하지만 이는 Feature가 1개일 경우이며, Data Set의 Feature가 여러 개일 경우 hθ(x)=θ0+θ1x1+θ2x2+θ3x3+…+θnxn와 같이 길게 표현되게 된다.(x0 = 1)

Multivariate Linear Regression은 위의 경우를 벡터를 이용하여 식을 간소화 시킨 것이다. 이렇게 간소화 한 공식은 Cost Function과 Gradient Descent 공식에도 적용할 수 있다.

$$X = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} \quad \Theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$$

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n$$
$$= \Theta^T X$$
$$= \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix} \times \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix}$$

$$\textit{Hypothesis: } h_\theta(x) = \Theta^T X$$

$$\textit{Cost function: } J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2$$

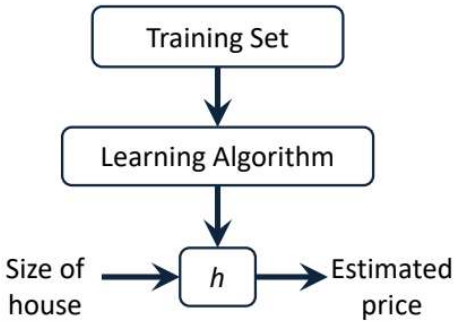
$$\textit{Gradient descent: } \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\Theta)$$

Feature Scaling – Feature의 값이 너무 작거나 클 경우 Gradient Descent의 값 도출에 많은 시간이 걸리게 된다.(너무 커서 복잡한 경로를 통해 최소값에 도달하거나 너무 업데이트 속도가 느리기 때문에 최종 결과값을 도출하는데 오래 걸리거나)

	그럴 경우 변수를 조정하여 결과값을 도출하는 시간을 조정하는 방법이다. 이는 어느정도의 오차를 동반하나 너무 큰 오차만 발생하지 않게 한다. Ex : x = size(0 – 2000 feet^2) -> x = size(feet ^ 2)) / 2000 Polynomial Regression – 값의 연속성이나 환경에 따라서 기존의 Traning Data를 통한 예측이 일반적이지 않을 경우 기존의 Hypothesis Function의 식에서 기인한 새로운 식을 만들 수 있다. Ex : 전 세계 IQ 별 인원의 경우 hθ(x)=θ0+θ1x1+θ2x2 처럼 위로 볼록한 형태의 그래프가 나올 것이지만 사이즈에 따른 집 값의 통계의 경우엔 hθ(x)=θ0+θ1x1+θ2√x2의 형태를 띠는 식을 사용하는 것이 더 적합하다. Normal Equation -
질문 내용	Gradient Descent를 이용하여 Cost Function의 최저값을 구할 때 아래의 두 상황의 경우에는 어떤 결과가 도출되는지, 만약 문제 상황이라면 어떻게 해결해야 하는지 궁금하다. 1. 최저점의 y 좌표가 모두 같을 경우 y = -x^2(x >= -1 && x <= 1) 2. 최저점이 끝도 없이 내려가는 형태일 경우.(y = x^3)

이름	이재은	학번	201502469

구분	내용
학습 범위	Lecture 1-Introduction Lecture 2-Linear regression Lecture 4-Multiple features
학습 내용	- 지도 학습(Supervised learning): 입력과 결과값(label)을 이용한 학습으로 분류(classification), 회귀(regression) 등 여러가지 방법에 쓰인다. - 비지도 학습(Unsupervised learning): 입력 값만을 이용하여 학습을 시키는 것으로 군집화(Clustering), 압축(Compression) 이나 변환 함수를 자동으로 알아내는 등의 문제에 쓰인다 - 지도 학습에서 문제는 2가지로 나뉜다. 먼저 기대되는 목적 값이 연속성을 가지고 있을 때는 Regression problem이다. 예로는 평수에 따른 집값의 변화가 있다. 반대로 목적 값이 연속성이 없어 몇가지 값으로 끊어지는 경우에는 Classification problem으로, 종양 크기에 따른 유방암 여부 그래프가 해당 된다. - 선형 회귀 문제(Linear Regression)이란 데이터를 분석했을 대 선형 그래프 형태로 정의 되는 문제이다.



위 그림의 h는 Hypothesis(추론)은 추론 알고리즘의 집합을 의미한다. Feature를 넣으면 Targeted value를 계산 해주는 일종의 공식이다. 추론 $h_\emptyset(x)=\emptyset_0+\emptyset_1 x$ 에 대해 Training Set의 입력 값(집의 평수)에 대해서 예측된 값과 Training Set에 정의된 Targeted Value 값과의 차이가 가장 적은 $(\emptyset_0, \emptyset_1)$ 를 구해내는 것이 최종 목적이다.

- 손실 함수의 기울기를 이용해 손실 함수의 극솟값을 찾는 알고리즘이 경사하강법이다. 즉, θ_0, θ_1 값으로 시작해서, $J(\theta_0, \theta_1)$ 가 최소가 될 때까지 θ_0, θ_1 의 값을 계속해서 바꾸는 것이다. θ_0, θ_1 값을 업데이트 하는 과정은 아래와 같다.

	$temp_0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} \cdot J(\theta_0, \theta_1)$ $temp_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} \cdot J(\theta_0, \theta_1)$ $\theta_0 := temp_0$ $\theta_1 := temp_1$ ※참고: **:=** : 수행을 하고 거짓이라면 Assertion (Assignment->Truth Assertion) **α** : learning rate $\partial/\partial\theta_j * J(\theta_0,\theta_1)$: 기울기 - Gradient Descent Algorithm은 반복하면서 수렴하는 값을 찾았다면, Normal Equation은 반복 없이 분석적으로 θ 값을 찾는 방법이다.
질문 내용	