연구 내용 요약 및 계획

|  |  |
| --- | --- |
| 날짜 | 2020년 2월 7일 |
| **1. 이 번주 연구 관련 학습 내용 정리**   * **데이터 사이언스 기초 강의 듣기(chapter1완료)** * **Tensorflow와 python을 이용한 실습(위에 강의에 대한 실습)** | |
| **2. 다음 주 학습 계획**   * **지금듣는 강의가 좀 어려워서 유데미에서 강의를 찾아서 들을 예정.**   **강의는 현재 검색 중. (data science 기초 강의 예정)**   * **강의 실습 내용 정리해서 git hub에 업로드** | |
| **3. 이번주 학습 내용 정리**  **linear함수를 데이터가 여러 개일 경우에 사용하는 방법. linear함수를 여러 번 사용는 방법도 있으나 그 보다는 matrix를 사용하면 더 편리하고 정확하게 결과를 추출해 낼 수 있다.**  **linear함수를 이용한 데이터 분류 알고리즘에 대해 학습했다. (logical regression)**  **이는 0과 1로 데이터를 분류하여 1이되는 원하는 값을 찾아내는 알고리즘이다. 데이터 값을 linear 알고리즘을 통해 sigmod를 구하고 이 값들로 logical을 이용하여 확률값으로 바꿔준다. 그 중 확률값이 가장 높은 값이 원하는 결과가 된다.** | |

20.02.04

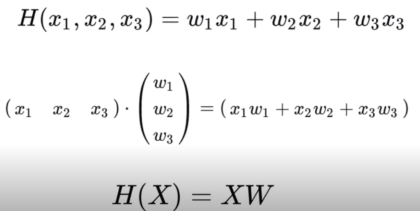
Multi-variable Linear Regression

Hypothesis : 가설함수

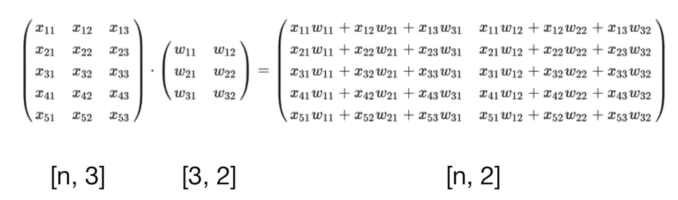
Cost function : 우리의 예측과 실제값의 차이

Gradient descent: : w가 최소가 되는 값을 찾아주는 알고리즘

데이터가 하나의 변수가 아닐경우 matrix를 사용하면 간편하게 표기 가능.



입력 데이터의 건수(인스턴스)가 많아도 동일하게 표현가능하는 장점이 있다.



출력 데이터가 여러 개인 경우는 입력데이터의 열수 출력데이터의 개수로 matrix의 크기가 결정된다.

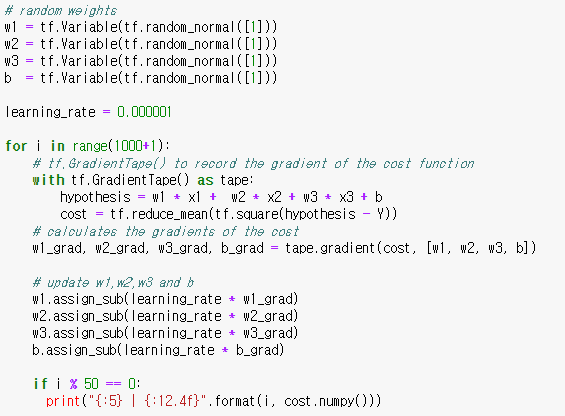
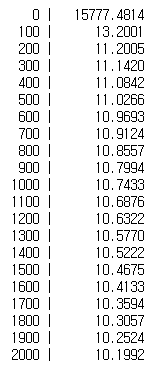
주어진 데이터로 학습을 시킨 후 새로운 3개의 데이터 값을 주었을 때 정확한 답을 추출해 내도록 함. 또한 비용은 어느정도인지 측정.

numpy의 slicing을 활용하여 데이터를 잘라냄.

X는 :밖에 없으므로 모든 행에서 -1=마지막 열을 제외한 데이터

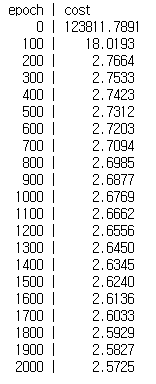
Y는 :는 모든 행을 의미. [-1]은 마지막 열을 의미하므로 마지막 열의 데이터.

데이터가 여러 개일 경우 matrix를 사용하면 식이 간단해지고 코드구현이 간단해진다.



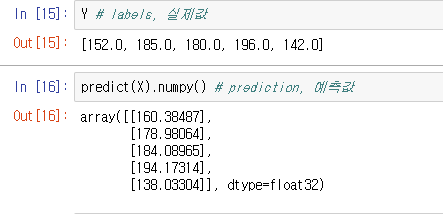
일일이 하나하나 변수할당과 값 할당을 해주어서 계산해야하지만

matrix사용시 하나의 식과 w가 하나로 표현이 가능하여 단순해진다.



또한 같은 데이터를 이용하여 비용을 비교하였을 때 matrix의 비용이 더 적게 측정된다.

이 학습된 알고리즘을 통해 값을 구한 결과



다음과 같이 결과를 구할 수 있지만 데이터가 적어 오차가 많이 발생한다.

Logistic Regression

분류기법중 하나. 서로 다른 특징을 가진 도형들을 분류하는 알고리즘

logistic데이터는 셀수있고 흩어져있으나 linear데이터는 데이터의 값이 이어져있다.

Logical은 데이터들이 구분되나 linear은 범위로 나타내어진다.

Ex)시간에 따른 합격률은 비례하는 상승 직선방정식을 나타내나 시간에 따라 합격하는가 안하는가로 나눈것인 logical 이다.

G(z) = 1/(1+) => sigmoid 함수

Linear 함수를 sigmoid함수를 통해 0/1의 값으로 나타낼 수 있다.

* 값을 구분하는 경계선을 decision boundary가 된다. Linear를 decision boundary를 거쳐서 logical로 만들 수 있다

Cost function

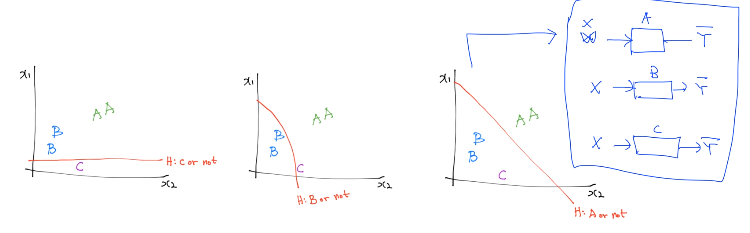
Loss\_fn함수를 이용하면 비용을 계산할 수 있다.(reduece\_min = 차원을 줄여줌)

Optimization : 비용을 최소화하는 방법.

Logistic regression은 0과 1을 잘 비교하기 위한 방법

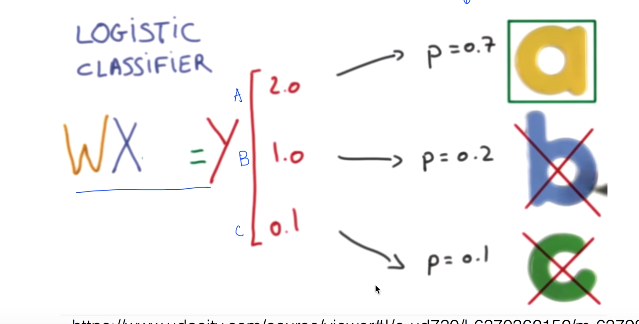
Multinomial classification

두개 이상의 데이터를 구별 방법



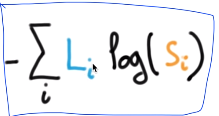
두 개 이상의 데이터를 구별하기 위해 두개의 데이터를 구별하는 방법을 여러 번 사용.

이보다 매트릭스를 사용한느 것이 구현하기 쉽고 편리하다.



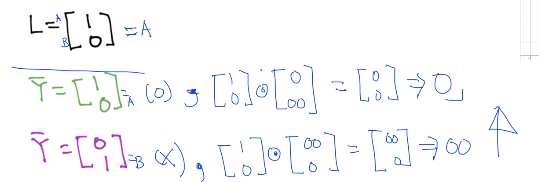
값을 0~1사이의 값으로 나타내며 그 값의 합이 1이 되도록 하는 것이 sigmoid

* 이를 구하기 위해 사용하는 것이 softmax

P를 확률로 봐서 가장 높은 값을 1로 하고 나머지를 0으로 만듬

* Cross-entropy

이 식을 이용하여 맞는값이면 0, 틀린값이면 매우 큰값을 출력하여 틀렸다는 사실을 알림..



L과 같이 A를 예측하고자 한다면 첫번째 Y^이 0으로 맞다는 것을 알 수 있다.

Softmax는 y값을 확률로 변경해 주는 함수. 함수이므로 모든 결과값의 합이 1이어야한다.\

캐글 현실세계에 있는 데이터 셋을 커널이라는 플랫폼을 이용하여 분석하고 예측하는 경진대회

->처음은 타이타닉으로 시작.