데이터 분석과 미적분

30515 원종우

목차

- 데이터 시각화
- 텐서플로우
- 경사하강법
- 응용학습

데이터 시각화란?

데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달되는 과정을 말한다. 이때, 시각적으로 표현한 표를 데이터 차트라고한다.

데이터 시각화를 왜 할까?

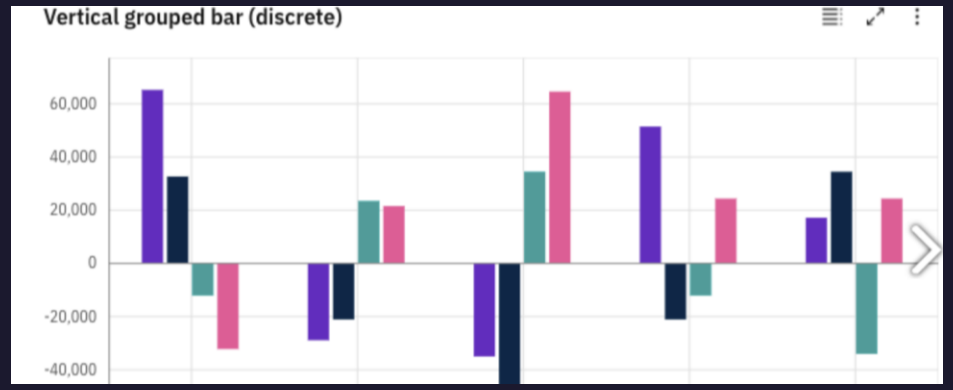
- 데이터 시각화는 데이터를 보는 방법을 바꿀 뿐만 아니라, 빠르고 효과적인 의사결정을 내리는 데도 결정적인 역할을 한다.
- 같은 정보라도 **시각화**를 잘 하면, 정보를 보다 쉽게 이해할 수 있다.
- 정보를 텍스트로 전달하는 것보다 그래프나 사진을 곁들여서 설명하면 우리는 보다 쉽게 정보를 이해한다.

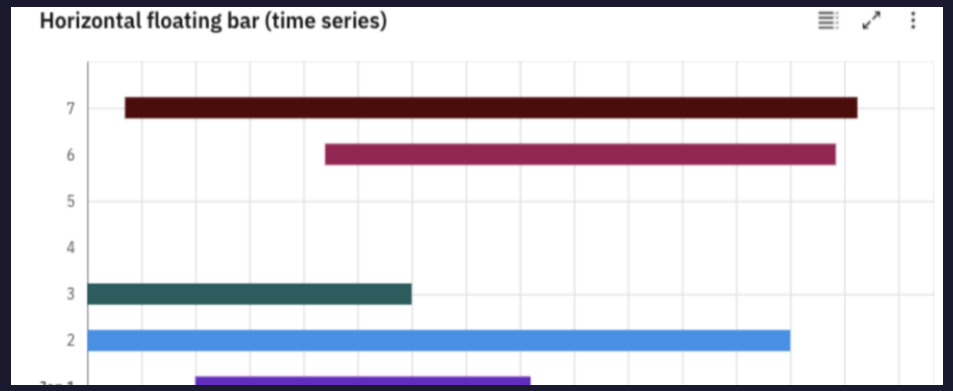
데이터 차트의 종류 (비교가 목적일 때)

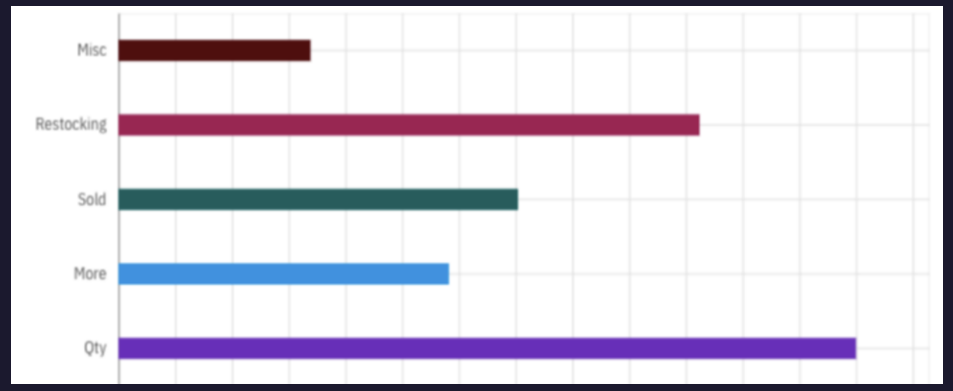
- 막대차트 사용

막대 차트의 종류에는 가로차트 그래프, 세로 막대 차트, 그룹형 막대 차트, 플로팅 막대 차트 등이 있음. 막대 차트는 각각의 카테고리 별 데이터를 비교하기 유리하며 시간 경과에 따른 추세를 표현하기 좋음

이외에도 롤리팝 차트, 버블형 차트 등 많은 차트들이 쓰임



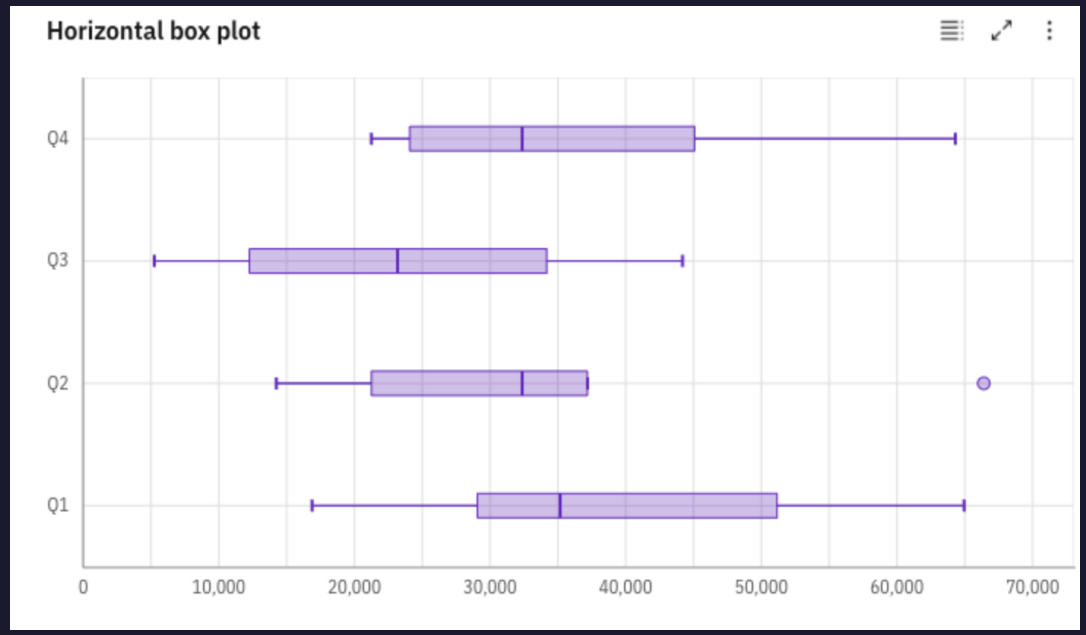


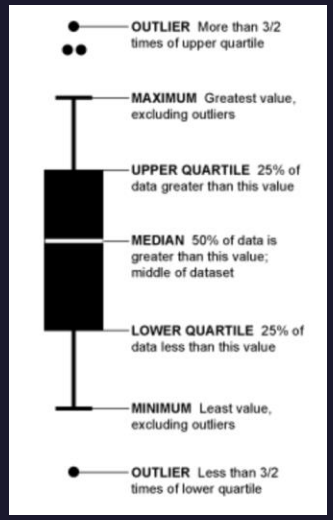


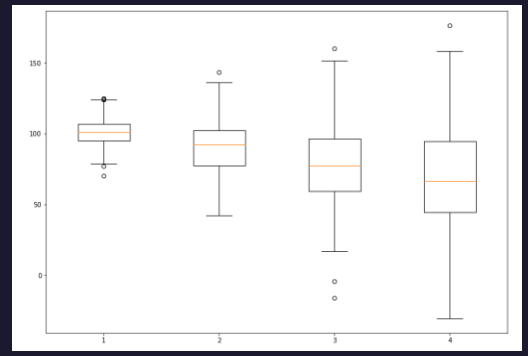
데이터 차트의 종류 (트렌드 및 추세를 표현할 때)

- 박스 플롯 차트 사용

박스 플롯은 많은 데이터를 눈으로 확인하기 어려울 때 그림을 이용해 데이터 집합의 범위와 중앙값을 빠르게 확인할 수 있는 목적으로 사용가능. 또한 통계적으로 비정상적인 이상점(outlier)이 있는지 확인 가능.







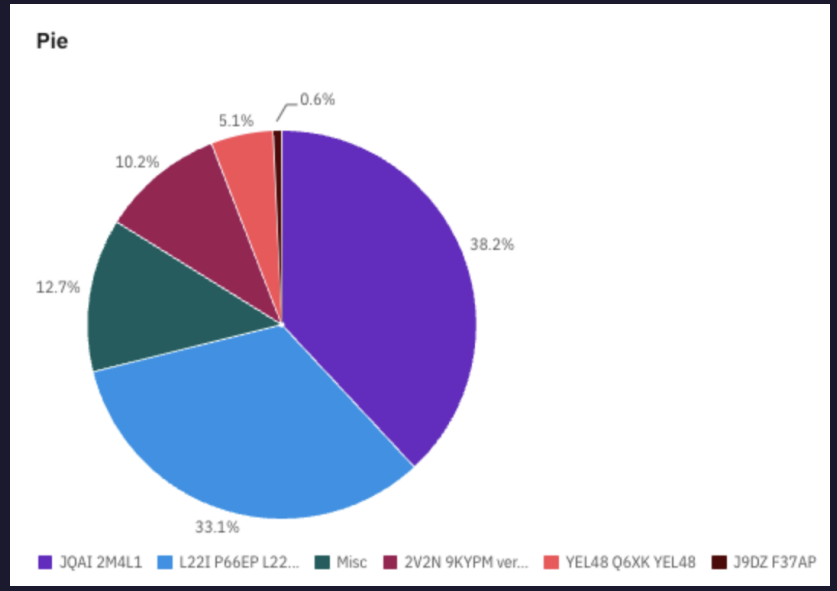
데이터 차트의 종류 (부분 대 전체 비교를 목적으로 한 데이터 시각화)

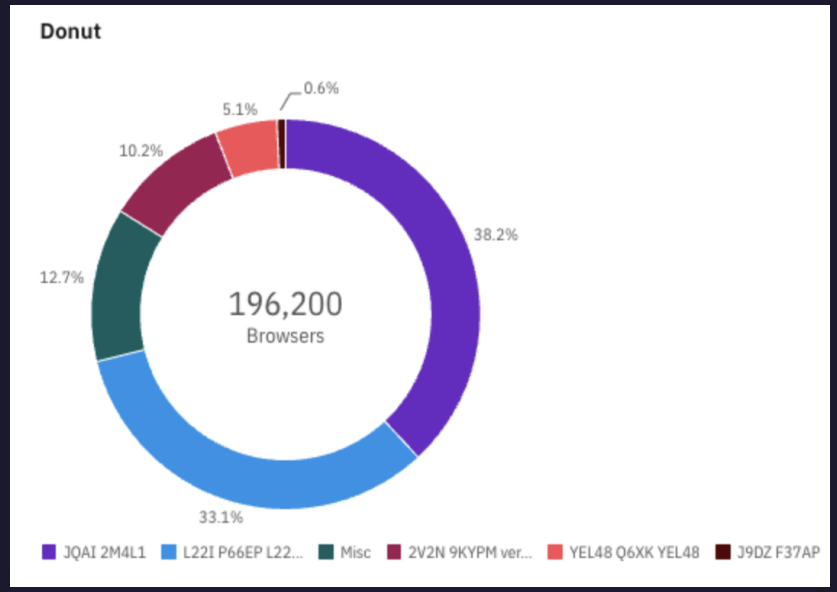
- 파이차트

파이 차트는 전체에 대한 각 부분의 비율을 부채꼴 모양으로 백분율로 나타낸 차트이다. 전체적인 비율을 쉽게 파악할 수 있어서 언론사에서 통계 수치를 공개할 때 자주 활용되지만 각 데이터 별로 크기의 차이가 없을 경우 시각적으로 비교하기가 어렵다는 단점이 있음.

- 도넛차트

가운데가 도넛처럼 비어 있기 때문에 가운데 영역을 활용할 수 있는 장점이 있음.

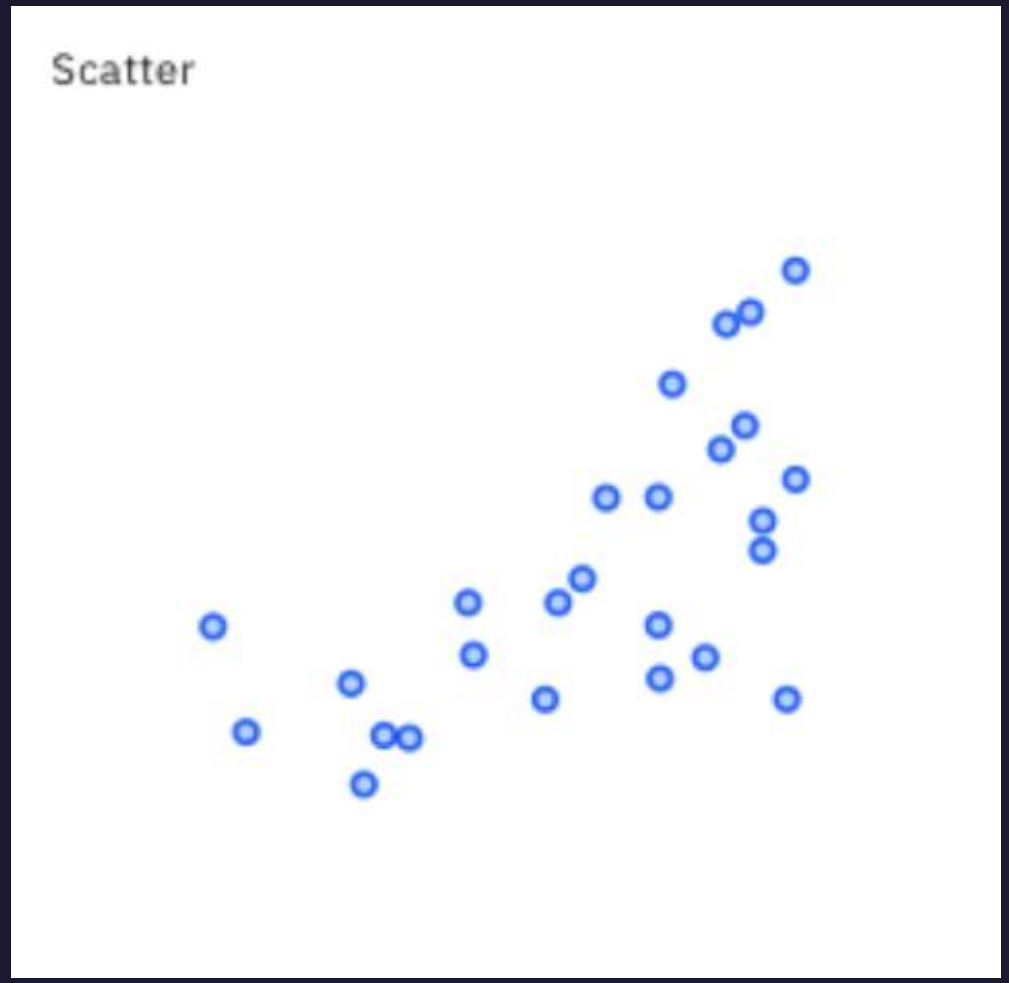




데이터 차트의 종류 (상관관계를 표현할 때)

- 스캐터 차트

좌표상의 점들을 표시함으로써 두 개 변수간의 관계를 나타내는 그래프 방법이다. 대부분의 차트에서는 카테고리가 축 중 하나에 표시되지만, 스캐터 차트에서는 카테고리는 점으로 표시되고 측정값은 두 축에 각각 표시된다.



어디에 미적분이 쓰일까?

- 데이터 시각화는 데이터의 경향성, 분포, 패턴 등을 그래프나 차트 등의 시각적인 도구를 활용하여 나타내는 것이다. 이때, 그래프나 차트의 모양을 결정하는 것은 데이터의 변화량, 증감 정도 등을 나타내는 수치값이다.
- 미적분은 함수의 변화량을 나타내는 수학적 도구이므로, 데이터의 변화량, 증감 정도 등을 수학적으로 계산하고 이를 그래프로 나타내는 것이 가능하다. 미분은 함수의 기울기를 나타내므로, 데이터의 경향성을 파악하는 데 유용하게 활용된다. 또한, 적분은 곡선 아래 면적을 나타내므로, 데이터의 총량이나 증감 정도를 파악하는 데 유용하게 활용된다.
- 미적분을 활용하여 수학적으로 계산한 데이터의 변화량, 증감 정도 등을 그래프나 차트로 시각화하면, 데이터 분석에서 더욱 정확하고 깊이 있는 결과를 얻을 수 있다.

실습을 해봅시다

앞서 미적분으로 데이터의 변화량, 증감 정도 등을 수학적으로 계산하고 이를 그래프로 나타내는 것이 가능하다고 언급하였는데 이게 진짜 저와 함께 가능한지 알아봅시다

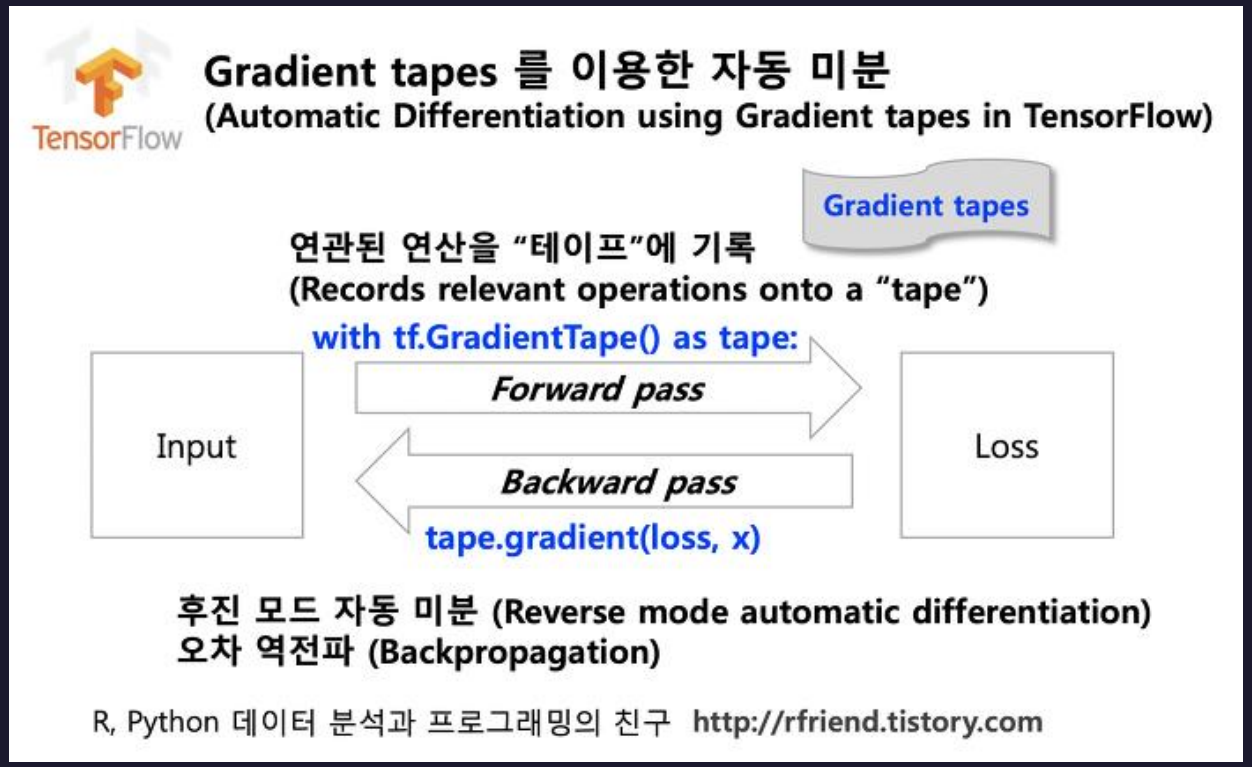
- 사용 언어 : Python
- 실습 환경 : Pycharm (파이썬 프로그래밍 언어에 특화된 통합 개발 환경)
- 실습 내용 : matplotlib를 이용하여 여러 그래프 그려보기, 간단한 이차함수 식을 이용하여 미분 함수 구현 후 기울기 값(미분값) 도출, 시각화
- 사용 라이브러리 : numpy(수치계산), matplotlib(그래프 그리기)

텐서플로우란?

- 구글이 2011년에 개발을 시작하여 2015년에 오픈 소스로 공개한 기계학습 라이브러리.
- 딥 러닝과 머신러닝 분야를 일반인들도 사용하기 쉽도록 다양한 기능들을 제공한다.
  2016년 알파고와 함께 한국에서도 관심이 높아진 추세이며 관련 컨퍼런스들도 개최되고 있다.

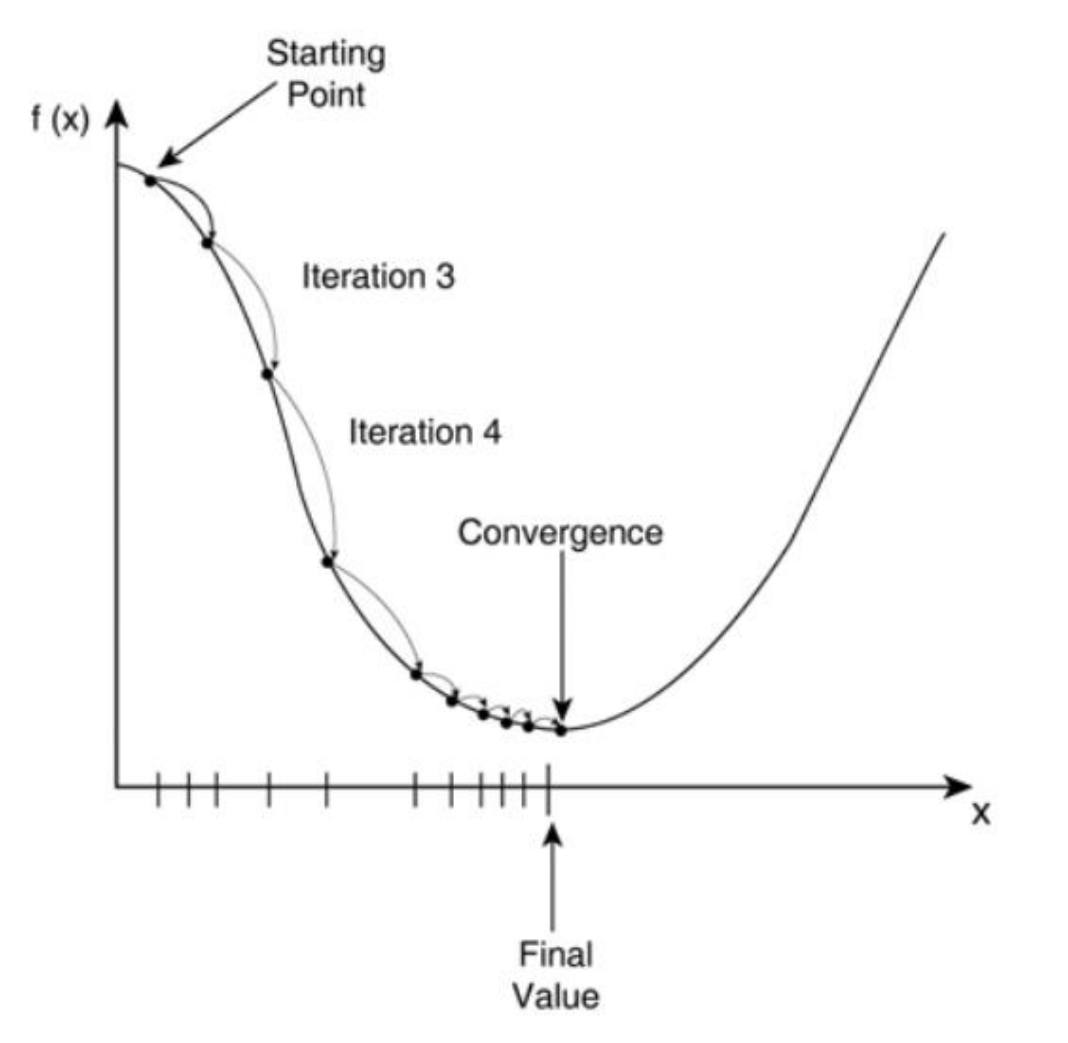
자동미분 소개

- 자동미분 : 신경망처럼 수만 개의 파라미터를 가진 복잡한 함수의 도함수(미분, 그래디언트)를 쉽게 계산할 수 있도록 해주는 도구.
- 텐서플로우는 자동 미분(주어진 입력 변수에 대한 연산의 그래디언트(gradient)를 계산하는 것)을 위한 API를 제공한다.
- tf.GradientTape는 컨텍스트(context) 안에서 실행된 모든 연산을 테이프(tape)에 "기록"합니다. 그 다음 텐서플로우는 후진 방식 자동 미분(reverse mode differentiation)을 사용해 테이프에 "기록된" 연산의 그래디언트를 계산합니다.
- 여기서 그래디언트란 다변수 함수의 모든 입력값에서 모든 방향으로의 순간변화율이다.



실습을 해봅시다

- 실습 내용 : 텐서플로우의 GradientTape를 이용하여 자동미분 구현
- 사용 언어 : Python
- 실습 환경 : 구글 코랩(colab)
- 사용 라이브러리 : tensorflow
- 소스코드 : https://colab.research.google.com/drive/1BeudFlx0CCt_aSHs36yOpBXZ9dmc32lR#scrollTo=oNacy-XunAv3



경사하강법이란?

경사하강법(Gradient Descent)은 함수의 최솟값을 찾기 위한 최적화 알고리즘 중 하나다. 경사하강법은 함수의 기울기(Gradient)를 이용하여 함수의 최솟값을 찾는 방법이다.

- 초기값 설정: 최솟값을 찾고자 하는 함수의 입력 변수(x)의 초기값을 설정한다.
- 기울기 계산: 입력 변수의 현재 위치에서의 기울기를 계산한다
- 이동: 입력 변수를 기울기의 반대 방향으로 이동시킵니다. 이때 이동 거리는 학습률(learning rate)이라는 하이퍼파라미터에 의해 결정된다
- 종료 조건 확인: 미리 정한 종료 조건(예를 들어, 일정한 반복 횟수, 혹은 기울기가 충분히 작아졌을 때)을 만족하는지 확인한다 만약 종료 조건을 만족한다면 알고리즘을 종료한다 그렇지 않다면 2번 단계로 돌아가 반복한다.

어디에 활용이 될까?

- 머신러닝: 선형 회귀(Linear Regression), 로지스틱 회귀(Logistic Regression), 신경망(Neural Network) 등의 모델에서 파라미터 학습에 이용됩니다.
- 딥 러닝: 심층 신경망(Deep Neural Network)에서 역전파(Backpropagation) 알고리즘과 함께 사용되어 모델의 가중치를 학습합니다.
- 자연어 처리: 자연어 처리에서는 최대우도추정(Maximum Likelihood Estimation)을 이용하여 모델을 학습하는데, 이때 경사하강법이 사용됩니다.
- 데이터 분석: 데이터 분석은 주어진 데이터를 통해 어떤 패턴이나 인사이트를 도출하는 것이 목적이므로, 데이터와 모델(혹은 함수)의 관계를 설명하는 파라미터를 경사하강법을 통해 찾아야 한다

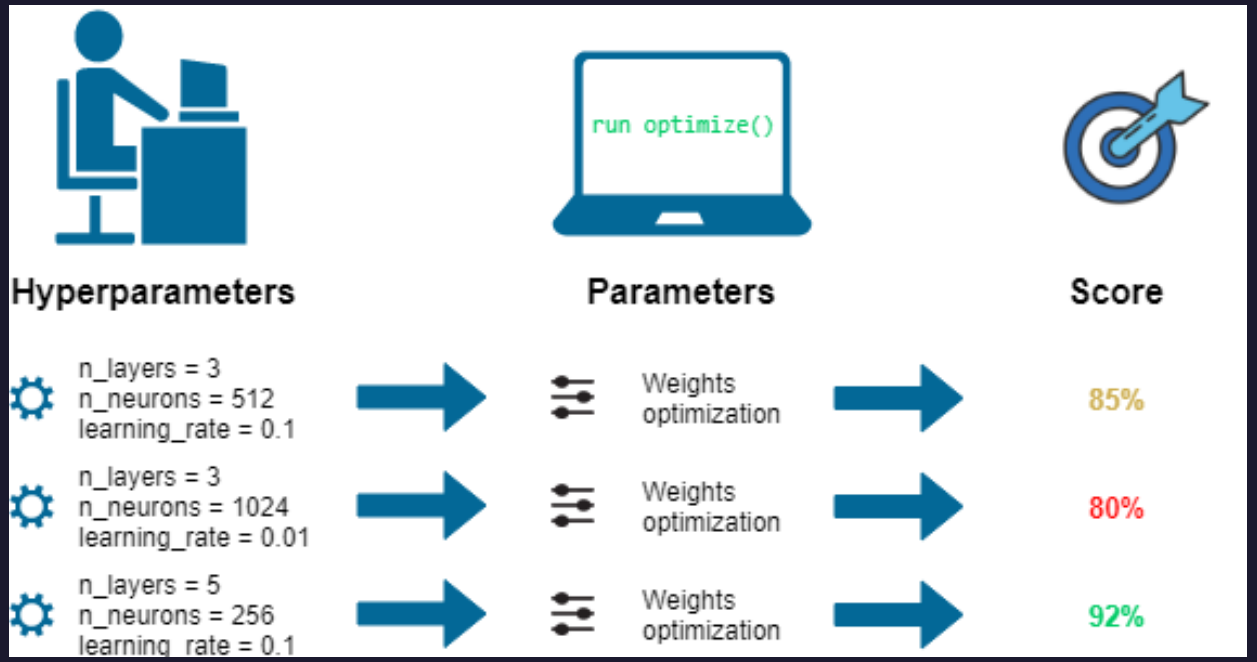
파라미터와 하이퍼 파라미터

- 파라미터

파라미터는 한국어로 매개변수이다. 가중치나 편향처럼 모델이 학습을 통해 최적의 값을 찾는 변수이다.

- 하이퍼 파라미터

하이퍼 파라미터란, 모델이 학습하면서 최적의 값을 자동으로 찾는 것이 아니라 사람이 직접 지정해 주어야 하는 변수다.



실습을 해봅시다

- 실습 내용 : **gradient_descent()** 함수에 경사하강법을 구현함. 함수 내부에서는 모델 파라미터 a 와 b 를 초기화하고, 반복문을 통해 모델 파라미터를 업데이트한다. 업데이트 과정에서는 손실 함수를 모델 파라미터로 미분한 결과를 이용하여 모델 파라미터를 조정한다. 이 과정을 통해 최적화된 모델 파라미터를 추정하고, 추정된 모델 파라미터를 이용하여 데이터와 모델의 관계를 시각화한다.
- 사용 언어 : Python
- 실습 환경 : Pycharm
- 사용 라이브러리 : numpy, matplotlib

응용학습

- 학습 내용 : 당뇨병 환자의 데이터셋을 이용하여 경사하강법(Gradient Descent)을 구현하고, 학습된 결과를 시각화
- 사용 언어 : Python
- 학습 환경 : Pycharm
- 사용 패키지 : scikit-learn (당뇨병 환자의 데이터셋을 포함)
- 사용 라이브러리 : matplotlib

감사합니다