Statistic

1. 확률모형과 확률변수는 무엇인가요?

확률 분포를 보다 단순하게 모사하기 위해 고안된 것이 **확률모형 (probability model)** 이다.

확률 모형은 분포함수 (distribution function) 또는 밀도함수 (density function) 라고 불리우는 미리 정해진 함수의 수식을 사용하여 분포의 모양을 정의 하는 방법이다. 이 때, 분포의 모양을 결정하는 함수의 계수를 분포의 **모수 (parameter)** 라고 부른다.

확률변수란?

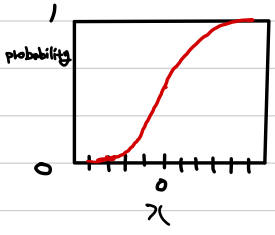
어떤 자료의 값이 분포가 특정한 확률 모형과 일치하는 경우 그 자료를 **확률 변수** (random variable) 라고 하고 해당 확률 모형을 따른다고 말한다.

2. 누적분포 함수와 확률밀도 함수는 무엇인가요? 수식과 함께 표현해주세요

누적 분포 함수 (cumulative distribution function) 란?

랜덤변수 X에 대하여 정의된 확률을 P_X 라고 할때 다음과 같이 정의되는 함수를 누적분포 함수(cdf) 라고 한다.

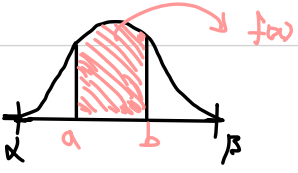
$$F_X(x) = P_X(X \leq x)$$



확률 밀도 함수란? (probability density function)

확률론에서 확률 밀도 함수는 확률 변수의 분포를 나타내는 함수로 확률 밀도함수 $f(x)$와 구간 $[a, b]$ 에 대해서 확률변수 X가 구간에 포함될 확률 $P(a \leq X \leq b)$는

$$\int_a^b f(x)\,dx$$ 가 된다.



Deep Learning

1. tensorflow와 pytorch의 특징과 차이?

구분	Tensorflow	PyTorch
패러다임	Define and Run	Define by Run
그래프 형태	Static graph(정적)	Dynamic graph(동적)
현재 사용자	많음	적음
자체 운영 포럼	없음	있음
한국 사용자 모임	Tensorflow Korea(TF-KR)	Pytorch Korea(Pytorch-KR)

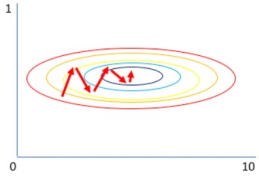
Define and Run은 코드를 직접 돌리는 환경인 세션을 만들고 placeholder를 선언하고 이것으로 계산 그래프를 만들고 (Define), 코드를 실행하는 시점에 데이터를 넣어 실행하는 방식 (Run), 이는 계산그래프를 명확히 보여주면서 실행 시점에 데이터만 바꿔줘도 되는 유연함을 갖지만, 그 자체로 비관적이다. 그래서 프레임워크 중 난이도가 높은 편

두 프레임 워크 모두 계산 그래프를 정의하고 자동으로 gradient를 계산해 주는 기능이 있다. 하지만 Tensorflow의 계산그래프는 정적이고 pytorch는 동적이다.

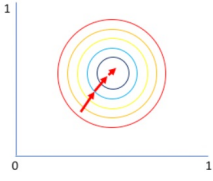
즉, tensorflow에서는 계산그래프를 한 번 정의하고 나면 그래프에 들어가는 입력 데이터만 다르게 할 수 있을뿐, 같은 그래프만 실행할 수 있다. 하지만 pytorch는 각 순전파마다 새로운 계산그래프를 정의하여 이용한다.

Pytorch 장점
1. 설치가 간편하다.
2. 이해와 디버깅이 쉬운 직관적이고 간결한 코드로 구성되었다.
3.Define by Run 방식을 기반으로 한 실시간 결과값을 시각화 한다.
4. 파이썬 라이브러리(Numpy, Scipy, Cython)와 높은 호환성을 가진다.
5. Winograd Convolution Algorithm 기본 적용을 통한 빠른 모델 훈련이 가능하다.
6. 모델 그래프를 만들 때 고정상태가 아니기 때문에 언제든지 데이터에 따라 조절이 가능하다(유연성).
7. Numpy스러운 Tensor연산이 GPU로도 가능하다.
8. 자동 미분 시스템을 이용해 쉽게 DDN(DataDirect Networks을 짤 수 있다.
9. 학습 및 추론 속도가 빠르고 다루기 쉽다.

Data Normalization은 무엇이고 왜 필요한가요?



Gradient of larger parameter dominates the update



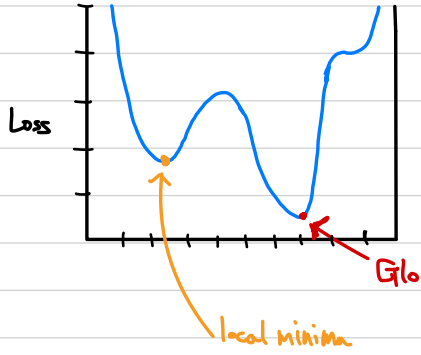
Both parameters can be updated in equal proportions

왼쪽과 같이 Unnormalized 상태에서는 Learning rate을 매우 작게 해야 학습이 된다. cost 그래프가 길쭉한 형태를 띄기 때문이다. input의 range가 서로 다르면 Gradient Descent Algorithm을 적용하는 것이 매우 까다로워진다. 하지만 Normalization을 적용하면 좀 더 구의 형태를 띄게 되고 최적화 지점을 쉽고 빠르게 찾을 수 있다.

∴ 정규화의 목표는 값 범위의 차이를 왜곡시키지 않고 data set을 공통 scale로 변경하는 것!

Machine Learning

1. Local minima와 Global Minima에 대해 설명해주세요!

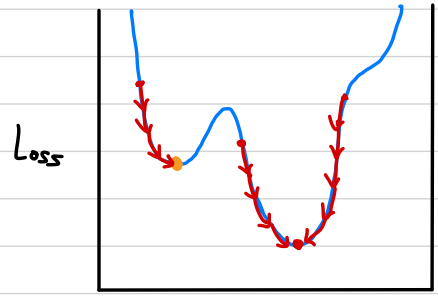


Local minima

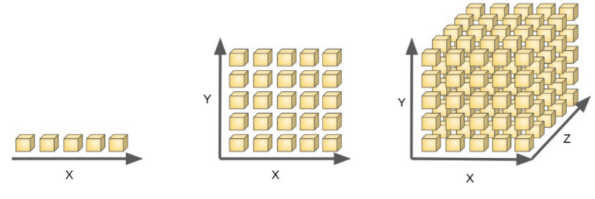
gradient descent를 사용하면서 마주치는 문제!

옆의 그림과 같이 학습을 진행할때 기울기가 local minima에서 cost function이 최소값을 얻는 것으로 착각하여 학습을 끝내는경우

학습시 가중치 최적화를 반복하여 최적 해를 찾아가므로 local minima에 수렴하여 Loss 값이 0 가까이 떨어지지 못한다 할지라도 시작위치가 다른 가중치에서 Global maximum에 수렴하여 Loss 값이 0에 수렴할 수있다.



2. 차원의 저주에 대해 설명해주세요!



특정 위치를 설명하는 상황에서 차원이 커질수록 설명공간이 지수적으로 늘어남.

→ feature가 많아질수록 충분한 데이터를 설명하는 빈공간이 증가함.

→ 차원의 저주로 인해 알고리즘 모델링 과정에서 저장공간과 처리시간이 비효율하게 증가됨!

1차원에 5개의 data	2차원의 25개의 data	3차원의 125개의 data
$5^1 = 5$	$5^2 = 25$	$5^3 = 125$

Solution

PCA, LDA, LLE, MDS, t-SNE

Database

DBMS를 정의하세요!

DBMS - Database Management System

다수의 사용자들이 데이터베이스 내에 데이터에게 접근, 사용할 수 있도록 해주는 소프트웨어를 말한다.

DBMS의 장점

1. 데이터 베이스 내 중복을 최소화할 수 있다.
   (완전히 허용하지 않는 것은 아님)
2. 같은 내용의 데이터를 여러 가지 구조로 지원해 줄 수 있는 DBMS의 정교한 기법 덕분에 데이터 베이스의 공용이 가능하다.
3. 데이터 베이스가 접근 처리 될때마다 제어 기능을 통해 그 유효성을 검사하기 때문에 데이터 무결성을 유지할 수 있다.
4. 철저한 보안 유지

DBMS의 단점

1. 컴퓨팅 시스템 자원의 소모량이 높음
2. 특정한 목적을 가지는 응용 시스템은 설계 시간이 길어지고 보다 전문적인 인원이 필요
3. DBMS는 통합된 시스템이기 때문에 일부가 장애를 일으키면 전체 시스템은 정지하며 시스템 신뢰성과 가용성을 저해할 수 있다.

데이터 중복 최소화, 데이터 공유, 일관성 무결성 보안성유지
최신의 데이터유지, 데이터 표준화 가능, 데이터의 논리적 물리적 독립성
융이한 데이터 접근, 데이터 저장공간 절약

DBMS의 종류

ORACLE DATABASE

MySQL

Microsoft SQL Server

RDBMS를 정의하고 장점에 대해 설명해주세요!

RDBMS는 관계형 데이터 베이스를 생성하고 수정하고 관리하는 소프트웨어라고 정의할 수 있다.

RDBMS의 장점

1. DB수준에서의 ACID 트랜잭션을 사용한 쉬운 개발 방식 사용
2. view를 사용한 column과 row에 대한 세밀한 보안 설정은 인가받지 않은 사용자들로부터의 조회나 변경을 막음.
3. 대부분의 SQL코드는 오픈소스로 포함한 다른 SQL 데이터 베이스로 포팅이 가능
4. 타입형 칼럼이나 제약들을 사용해서 데이터 베이스에 정보를 추가하기 전에 유효성 검사를 하여 데이터 품질을 향상함.

RDBMS의 단점.

1. 객체-관계형 매핑 레이어가 복잡해 질 수 있음
2. ER 모델링이 테스팅 전에 완료되어 있어야하며, 이는 개발을 더디게 한다.
3. RDBMS 시스템은 초기에 필요한 경우, 확장성을 제약함.
4. 여러 서버를 사용한 sharding 기술이 가능하지만, 별도의 어플리케이션 코드가 필요하며 비효율적
5. 테이블에 다양한 가변성이 있는 데이터를 저장하게 어렵