Machine Learning

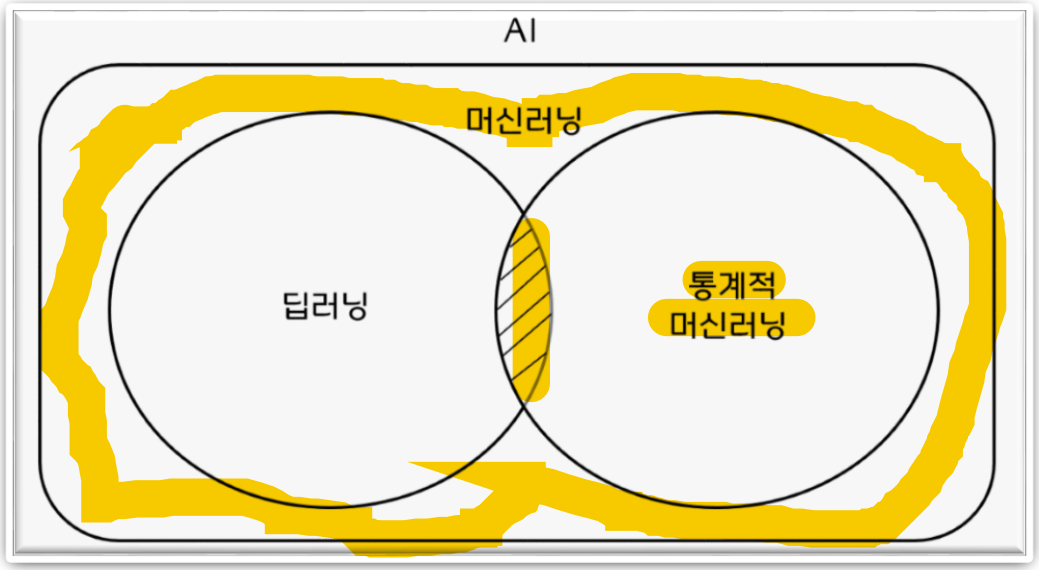
1장 Deep Learning이란?

고려대학교 통계학과
박유성

Contents

01 AI, Machine Learning, 그리고 Deep Learning

- AI (Artificial Intelligence)는 컴퓨터로 만들어진 지능으로 "사람에 의해 통상적으로 수행되는 지적인 작업을 자동화 하는 것을 목표"(Chollet, 2018)로 하는 학문 영역임.
- 또한 4차 산업혁명의 핵심인 machine learning과 deep learning을 포괄하는 보다 광범위한 개념임.



AI (Artificial Intelligence)

- 1950~1980년대에는 부호적 AI (Symbolic AI)에 치중함. 이는 명시적 규칙을 부여하여 사람이 하는 일을 자동화 하는 AI임.
- 1969년 perceptron 개념 도입했지만 가장 간단한 XOR 해결하지 못함 (AI 암흑기).
- 1980년에 activation function과 손실함수 도입함.
- 1980년대에는 다양한 인지적 오류와 차이가 존재하는 이미지 인식, 자연어 해석, 번역 등 인간의 인지 영역에서의 한계(AI의 암흑기)를 극복하기 위해 AI의 핵심분야인 Deep Learning이 탄생함.
- AI의 영역에는 Machine learning 이외에도 사물 인터넷(IoT)과 이로부터 발생하는 빅데이터 처리 및 저장하는 블록체인(Blockchain) 기술 등이 있음.

BIG DATA

다음과 같이 3V를 만족하는 데이터를 말한다.

Volume

수십 테라바이트의 크기이며 계속 증가함.

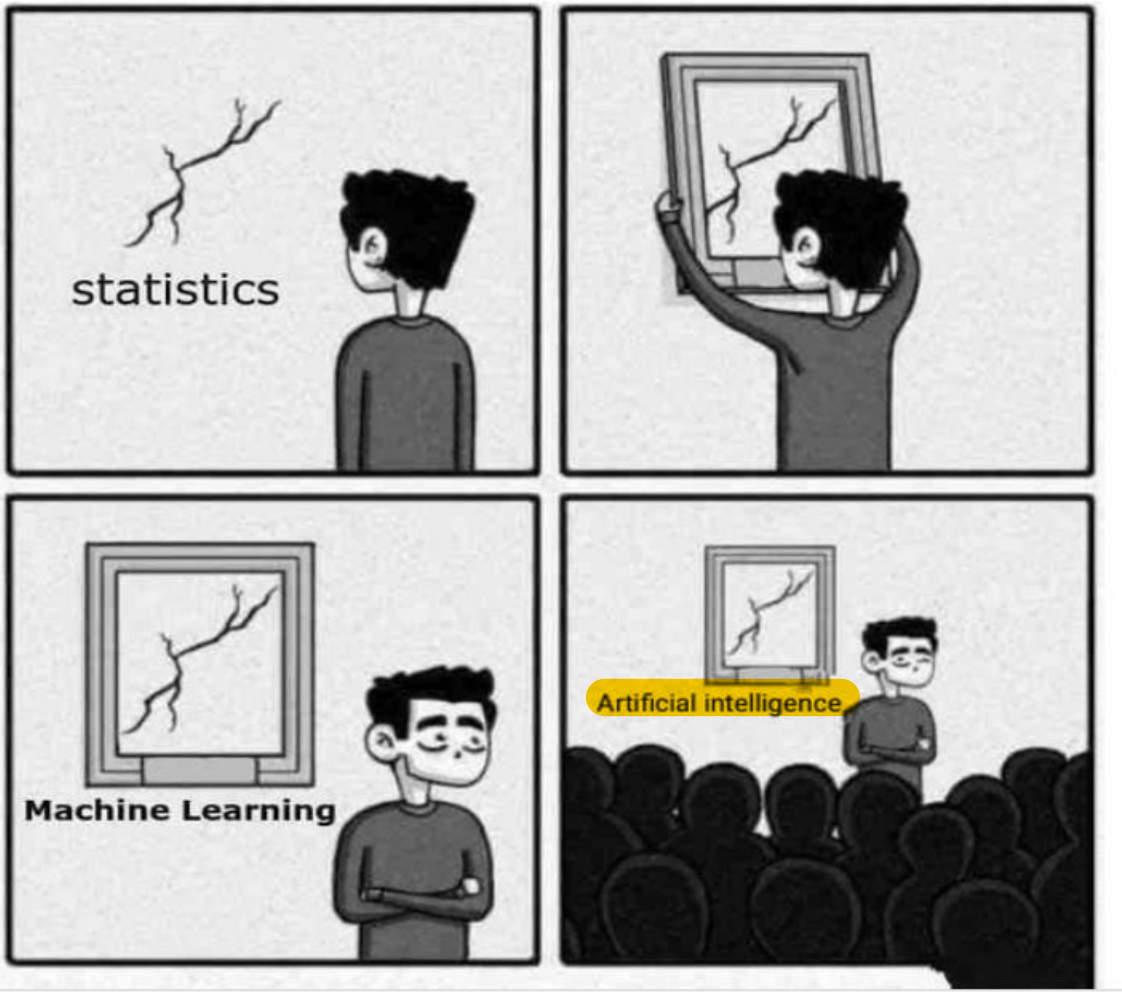
Variety

실수 형태의 정형데이터와 이미지, 언어, 글씨, 소리, 신호 등 비정형데이터

Velocity

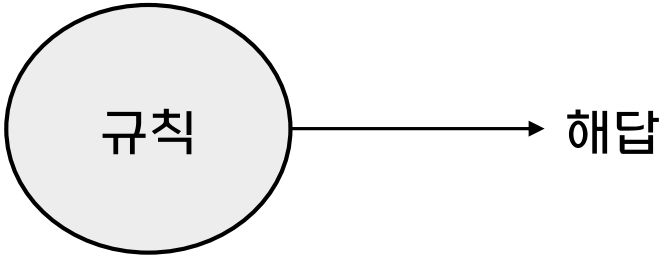
데이터의 전송속도(5G)

AI (artificial intelligence)는?



Machine Learning

- 부호적 AI



- ML (Machine learning)

– 통계적 규칙의 의미는 허용 가능한 오차가 존재한다는 의미임(전통적 통계학과 개념적으로 일치).

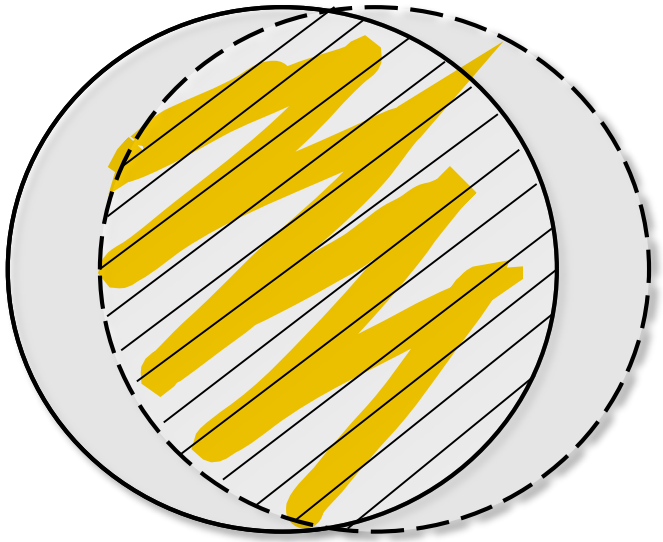
Machine Learning

전통적인 통계학

- 규칙의 통계적 추론에 중점(전문적인 통계적, 수학적 지식)
- 자료의 특성(다변량, 시계열, 범주형 등)에 따라 분석.

통계적 머신러닝

- 규칙의 상식적 일반화에 중점
- 목적변수의 관측여부에 따라 지도학습, 비지도학습으로 분석



——— 통계

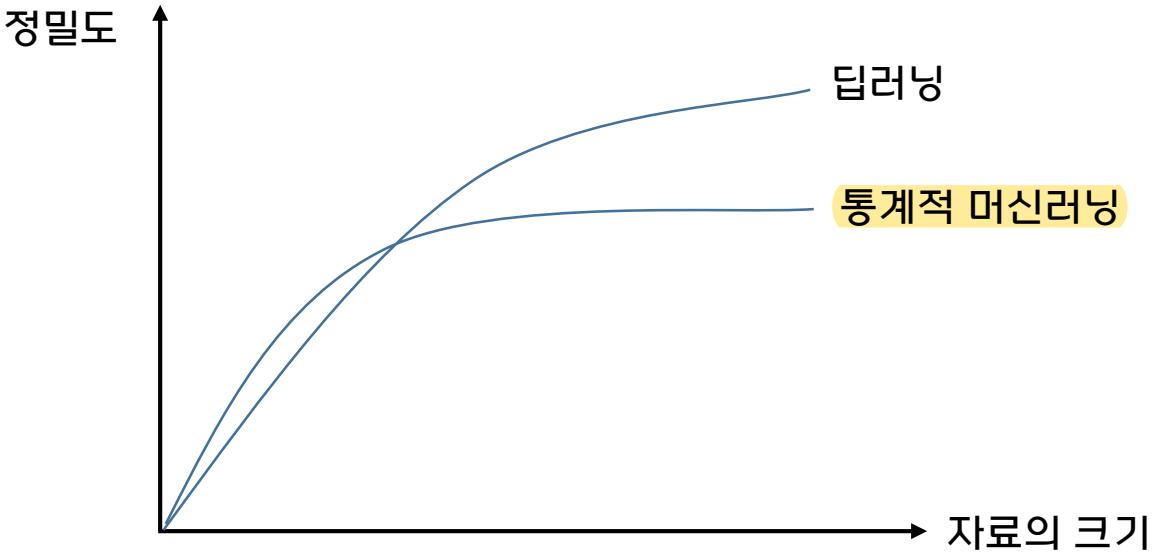
– – – 통계적 머신러닝

통계적 머신러닝과 딥러닝

구분	통계적머신러닝	딥러닝
데이터 크기	중/소 크기	빅데이터
분석자료 형태	2차원 텐서	2차원 텐서 이상
강점을 갖는 자료	정형화된 자료	비정형자료
특성변수	특성변수를 만들어야 함	특성변수가 만들어짐
특성변수의 정규화 및 표준화	선택	필요
모형	매우 많음	기본적으로 3 개의 모형
최적화	일반적으로 전체 데이터 사용	배치데이터
해석여부	해석이 쉬움(단, SVM과 boosting 제외)	어렵거나 불가능
하드웨어	중급	고성능(GPU 요구)
실행요구시간	최대 시간 단위	최대 주단위 시간

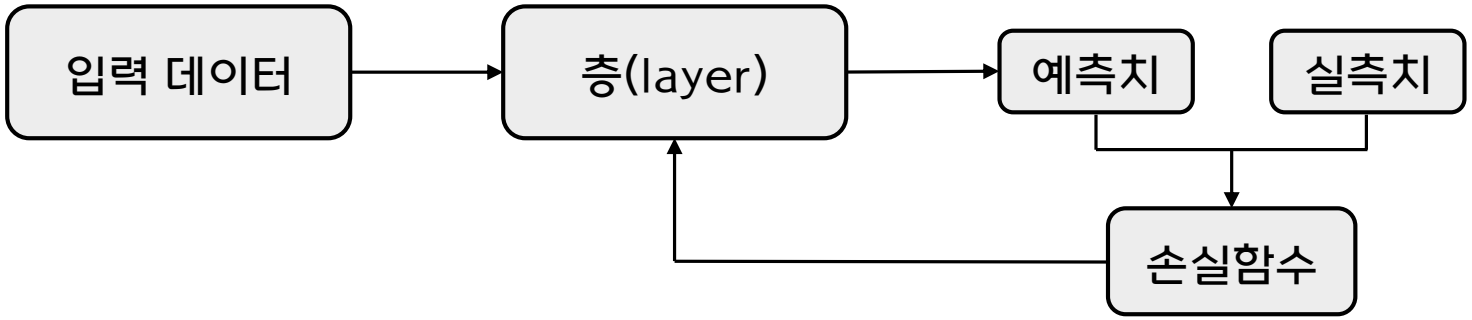
통계적 머신러닝과 딥러닝

자료의 크기와 정밀도



Machine Learning

- Machine learning의 구조



<Machine learning의 구조>

- k 차원 tensor로 이루어진 n개의 입력데이터가 층에 입력되면 특성변수의 결합(선형 또는 비선형)을 통해 또 다른 특성변수로 변환하거나 생성한 후 이를 예측함수에 입력하여 예측치를 출력하게 됨. 이후 손실함수의 값을 줄이기 위해 특성함수 결합의 가중치를 변경하는 과정을 반복함.

Deep Learning의 발전

- Convolution Neural Network (CNN, 2012)

  140만개의 images로부터 1,000의 객체를 구별을 기존의 74.3%에서 83.6%

  → 2015년 96.4%로 우승

- RNN에 의한 자연어처리(2016년) →LSTM, GRU.
- 재현 learning (2018년)

→ Transfer learning, ResNet, Sequence-to-Sequence learning, autoencoder, GANs, VAE 등

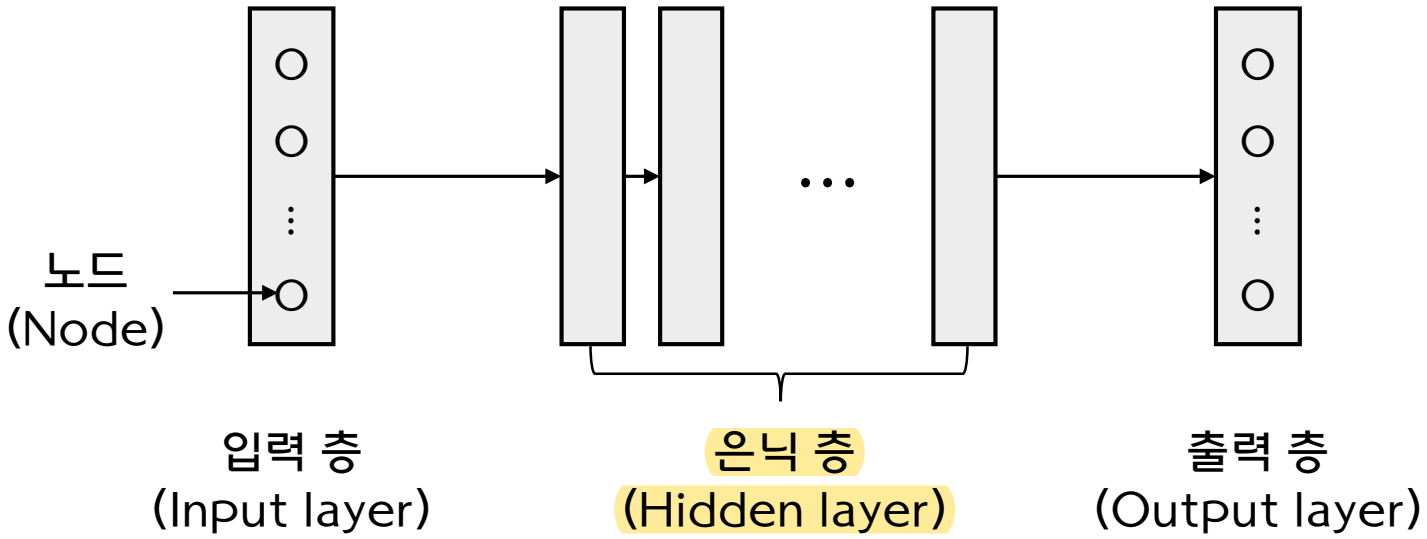
Deep Learning

- 딥러닝은 특성변수들의 선형 및 비선형 결합을 통해 목적변수를 확률적으로 맞추는 전형적인 통계적 모형이다.
- 여기에서 특성변수는 통계학에서 독립변수 x와 동일하며 목적변수는 종속변수 y와 동일하다.
- 딥러닝은 수많은 데이터를 통해 사람의 인지능력에 대한 통계적 규칙을 찾아 구현하는 통계모형의 일종이라고 할 수 있다.
- 예를 들어, 사람은 글을 쓸 때 상호간에 의미가 통하도록 의미론적 관점에서 다음 단어를 선택하지만, 딥러닝은 확률적으로 나타날 가능성이 가장 큰 단어를 다음 단어로 선택한다.
- 물론 확률은 빅데이터를 통해 미리 계산된 것이다.

Deep Learning

- 이러한 이유로 비교적 확률화 모형이 용이한 이미지인식을 기반으로 하는 자율자동차는 실용화 단계에 이르렀지만,
- 신뢰할만한 대화기술, 사람수준의 통역 특히, 보통사람 수준의 일반적인 지능을 갖는 딥러닝의 개발은 많은 시간이 소요될 수 밖에 없다.

Deep learning



- 머신러닝은 은닉층이 없다.
- 딥러닝은 MLP (multilayer perceptron), CNN (convolutional neural networks), RNN (recurrent neural networks) 등 세 가지 기본모형으로 구성되어 있으며 딥러닝 모형은 이 세 모형의 다양한 조합으로 구축된다.

Deep learning

- 그러므로 딥러닝의 출발은 이 세 개의 기본모형을 기능적인 측면에서의 이해 보다는 개념적인 측면에서의 이해가 필수적이다.
- 기능적인 측면에서의 이해란 CNN은 이미지분석에 사용하고 RNN은 머신번역 등 언어나 텍스트 자료분석에 사용하는 딥러닝 모형과 같은 이해를 말한다.
- 한편, 개념적 측면에서의 이해는 자료의 구조적인 측면에서의 이해를 말하는 것으로 행렬의 일반화인 텐서(tensor)를 이해하고, 데이터를 구성하는 특성변수와 목적변수의 역할을 이해하고 특성변수와 모수와의 관계를 이해하여야 한다.

Deep Learning

- 0차원 텐서(0D텐서): 스칼라, 1차원 텐서(1D텐서): 벡터,

  2차원 텐서(2D텐서):행렬, 3차원 텐서(3D텐서): 3개의 축으로 구성된 데이터

  4차원 텐서(4D텐서): 4개의 축으로 구성된 데이터

- MLP는 2D텐서 데이터를, CNN은 4D텐서 데이터를, 그리고 RNN은 3D텐서 데이터를 사용할 수 있도록 특화되어 있다.

- 딥러닝의 기본 모형에 사용되는 텐서 데이터의 제 1축은 항상 데이터 또는 배치의 크기를 나타내며 나머지 축은 특성변수 또는 목적변수를 나타낸다.

- 그러므로 MLP의 특성변수는 1D텐서, CNN은 3D텐서, 그리고 RNN은 2D텐서의 특성변수가 정의된다.

Deep learning

- 딥러닝은 입력층, 은닉층, 그리고 출력층의 기본구조를 가지고 있다.
- 은닉층은 목적변수를 좀 더 잘 설명하기 위한 특성변수를 잘 가공하는 층이라고 이해하면 된다.
- 특성변수가 은닉층에 전달되면 특성변수를 선형 및 비선형결합을 하고 선형결합에 사용되는 가중치를 모수라고 한다.
- 그러므로 이 모수를 추정하여 목적변수를 예측하게 된다.
- 여기에서 중요한 것은 이 모수들이 표본에 의존하지 않는다는 사실이다.
- 이 의미는 각 표본은 특성변수의 선형결합에 동일하게 기여한다는 것을 말하며, 각 표본은 통계적으로 같은 분포에서 나온 정보이며 서로 간에 독립이라는 의미를 갖는다.

Deep learning

- 특성변수와 목적변수는 하나의 쌍으로 관측되므로 표본의 독립성에 의해 동일한 특성변수에 서로 다른 목적변수가 관측되어야 하고 반대로 동일한 목적변수에 서로 다른 특성변수도 관측되어야 한다.
- 표본의 독립성은 딥러닝모형을 실제 응용할 수 있도록 하는 일반화에 매우 중요한 요소이다.
- 딥러닝 모형의 입력층에는 특성변수와 목적변수가 텐서형태로 입력되고 이 텐서가 은닉층에 전달되어 목적변수를 잘 설명하도록 특성변수를 선형 및 비선형결합을 통해 새로운 특성변수를 만들어 낸다.
- 그러므로 은닉층에서는 일종의 매개변수를 만들어 낸다고 생각하면 된다.

Deep learning

- 출력층 직전의 은닉층에서 만들어진 매개변수는 출력층에 입력되고 출력층에서는 입력된 매개변수를 선형 및 비선형 결합을 통해 목적변수와 동일한 크기의 텐서를 출력하게 된다.
- 출력층의 출력은 목적변수가 클래스의 분류이면 조건부확률이고 목적변수가 주식가격과 같은 연속형이면 조건부 기대값이므로
- 결국 출력층의 예측치는 확률적 의사결정 문제와 동일하며 이는 전통적인 통계적 예측치와 동일하다.

Deep learning

- 은닉층은 2개 이상이며 입력층과 출력층은 연구목적에 따라 2개 이상일 수 있다.
- 딥러닝 모형에서의 텐서의 입출력을 설명할 때, 배치크기를 나타내는 제 1축을 제외한 텐서만을 설명한다. 근본적인 이유는 딥러닝 모형의 최적화를 위한 배치의 크기를 자유롭게 조정하기 위함이다.
- 입력된 텐서가 1D텐서이면 MLP 모형을 사용하고 2D텐서이면 RNN 모형을 사용하며, 3D텐서이면 CNN모형을 사용하여 새로운 텐서를 출력하고 이를 다음 층에 입력한다.
- 입력받은 데이터가 2D텐서이지만 MLP모형을 사용할 수 있다. 이 경우, 전달받은 텐서의 제 1축(실제로는 제 2축)을 표본으로 인식하여 처리하게 된다.

Deep learning

- 출력층에 의해 예측된 목적변수 값은 실제 목적변수 값과 비교하여 이들 간의 일종의 거리인 손실함수를 계산하고 이 손실함수를 최소화하기 위한 모수 최적화를 하게 된다.
- 손실함수는 입력층 텐서의 제 1축인 배치크기 단위로 계산되며 모수의 최적화는 소위 역전파(backpropagation)을 통해 구현된다.

Deep Learning의 성과

- 통계적 머신러닝은 전문적 지식이 요구되는 영역에 주로 적용
- 딥러닝은 우리의 일상생활이 주 적용 영역임
- 사람과 비슷한 수준의 이미지분류, 음성인식, 필기인식.
- 향상된 머신번역, Text-to-Speech 변환
- Google now, Amazon Alexa와 같은 초보적 디지털 비서
- Google, Bing에서 사용하는 향상된 광고 targeting, 향상된 웹 검색.
- 자연어 질문에 대답하는 능력, 사람을 능가하는 바둑실력.
- 자율자동차 실용화 단계.

Deep Learning의 오해

- 딥러닝은 인간의 뇌를 모형화하지 않음. 단지 자료의 대표특성을 찾는 알고리즘임(Neural Network이라는 이름으로부터의 오해).
- 신뢰할만한 대화기술, 사람수준의 언어 통역, 사람수준의 자연어 이해 특히,보통사람 수준의 일반적인 지능은 많은 시간이 필요함.
- 딥러닝은 사람과 달리 입력데이터를 전혀 이해하지 못함.
- 딥러닝은 훈련데이터와 완전히 다른 자료를 입력하면 황당한 결과를 보임.
- 예를 들면, 사람은 복권이 당첨되었다고 가정하면 추상적 관념(abstraction)과 추론(reasoning)으로 "어떠할 것이다"가 예상되지만

Deep Learning의 오해

- 딥러닝은 불가능. 즉 데이터가 거의 없거나 전혀없는 "기발하고 전혀 경험없는 것"에 대해 적응력이 없음.
- 보통 사람의 지능을 갖는 딥러닝이 성공하더라도 과학이나 소프트웨어 개발 등에 사람의 보조원으로 역할을 함.
- 인간을 조정하거나 심지어 지배한다는 것은 불가능함(딥러닝의 훈련은 사람의 세세한 진단이 요구되기 때문임)

왜 Deep Learning은 이제야 빛을 보는가?

- CNN (1989), backpropagation (1990), LSTM(1997)이 이미 개발되었으나 2012년부터 딥러닝은 비약적 발전을 함.
- GPU의 개발(딥러닝은 수많은 행렬연산임)
- Big data: 인터넷, 유튜브 등으로 부터의 데이터(딥러닝의 연료)
- 최적화 알고리즘의 개발
- C++나 CUDA와 같은 전문적 지식을 요구하는 프로그램언어 대신 배우기 쉬운 Python
- 인터넷만 있으면 무료로 사용(colab). 휴대전화도 가능, 아이패드도 가능(Carnets)
- LEGO와 같은 Keras(2015년)등 user-friendly 라이브러리(딥러닝의 대중화)

Deep learning의 미래

- 딥러닝은 팬과 종이로 풀어내는 수학· 통계학이 아닌 데이터로부터 해답을 찾는 데이터과학(data science)임.
- 딥러닝의 원리는 간단하며, 특히 통계적 머신러닝에서 가장 어려운 특성변수의 추출을 모형화함.
- 빅데이터로 개발된 딥러닝모형은 새로운 작은 데이터로 연계해서 사용할 수 있기 때문에
- 기존 딥러닝모형을 축적하여 좀더 복잡하고 강력한 딥러닝 모형을 개발할 수 있음.
- 딥러닝의 하드웨어인 GPU보다 연산 능력이 10배 정도 빠른 tensor processing unit (TPU)가 개발완료 단계에 있음.

AI의 미래

- 보다 향상된 최적화 알고리즘의 개발
- 사물인터넷, Blockchain등으로 부터의 신뢰도 높은 빅데이터의 생산
- 미분 불가능한 시스템의 개발(generalized backpropagation)
- 3살 정도의 AI이지만 멀지 않은 미래에 AI시대가 도래할 것임.
- AI는 복잡하고 정보집약적 시대에 Interface를 제공하여
- 사람의 비서가 되고 친구가 되고, 일상의 정보에 대한 질문에 대답하고, 아이들의 교육을 돕고, 식료품을 배달하고, 사람대신 운전해줄 것이다.
- AI는 유전체학으로부터 수학까지 획기적인 과학발전의 보조자가 될 것임.
- AI는 현재의 internet처럼 우리 일상과 삶의 거의 모든 과정에 응용될 것임.

02 Deep Learning의 내용

- 딥러닝의 기본모형을 근본적으로 이해하기 위해서는 추정해야 할 모수의 수를 계산할 수 있어야 한다. 제 2장 딥러닝 데이터와 세 가지 기본 신경망은 지금까지 논의한 내용을 다루고 있으며 가장 중요한 핵심내용 중의 하나이다.
- 제 3장에서는 딥러닝의 최적화 문제를 다루고 있다. 최적화는 손실함수를 최소화하는 모수를 추정하는 것을 말한다.
- 목적변수가 클래스를 나타내는 변수일 때 예를 들어, 감성분석, 숫자인식, 이미지 인식, 머신번역 등의 분류(classification)가 목적일 때, 손실함수는 이항 또는 다항분포 우도함수(likelihood function)의 음수로 정의되고,
- 목적변수가 주식가격과 같은 연속형일 때는 평균제곱합(mean squared error) 또는 평균절대오차(mean absolute error)등이 손실함수이다.

02 Deep Learning의 내용

- 입력변수의 분포를 재생하는 GAN (generative adversarial network)의 경우 (제 11~13장), Kullback-Leibler divergence가 손실함수로 사용될 수 있다. 이 역시 전통적으로 통계학에서 사용해오고 있는 손실함수이다.
- 손실함수를 최소화하는 역전파는 출력층으로부터 입력층 방향으로(즉, 역방향으로), 출력층의 모수로부터 차례대로 은닉층의 모수 그리고 최종적으로 입력층에 연결된 은닉층의 모수 순서로 모수를 최신화하는 기법을 말한다.
- 역전파는 미분의 체인룰(chain rule)과 동일하다. 역전파 과정에서 중요한 것은 극복해야 할 과제인 손실함수의 국소최소값(local minimum), 안장점(saddle point), 그리고 최소값 근방에서의 진동문제를 해결하는 것이다. 이러한 문제를 해결하기 위한 일곱가지 최신화 알고리즘의 원리와 특성을 제 3 장에서 논의하게 될 것이다.

02 Deep Learning의 내용

- 손실함수의 최소화과정에서 필연적으로 수반되는 문제는 과대적합(overfitting)문제이다.
- 학습데이터를 이용하여 모형의 모수를 최적화하고 검증데이터를 이용하여 초모수(hyper-parameter)를 조정하고 모형을 최적화하게 된다.
- 최적화된 모형은 시험데이터에도 잘 작동하여야 한다.
- 학습데이터에서의 모형 성능만큼 시험데이터에서도 보이지 못한다면 이는 학습된 모형이 학습데이터에 과대적합되었으며 학습된 모형의 일반화(generalization)가 실패했다고 말한다.
- 이러한 머신러닝과 딥러닝에서의 일반화 진단은 통계학과 가장 큰 차이점을 보여주고 있다.

02 Deep Learning의 내용

- 머신러닝과 딥러닝에서는 통계학의 추론 중 검정(testing)대신 위와 같은 간단하고 직관적인 일반화 진단을 사용하기 때문에 통계학에서 가장 어려운 분야 중의 하나인 검정통계량과 이에 관련된 표본분포(sampling distribution)의 유도가 불필요하다.
- 과대적합은 모형이 학습데이터의 목적변수를 과대하게 잘 설명하기 때문에 일어난 현상이며 이는 모형의 임의성을 지나치게 제한하였기 때문에 발생하는 현상이다.
- 모형의 임의성을 증가시키는 가장 간단하고 좋은 방법은 데이터를 증가시키는 것이다. 차선책으로는 중요하지 않은 모수를 규제화(regularization, 일부 모수를 0으로 놓거나 거의 0으로 놓는 것)하거나 학습과정에서 일부 특성변수를 빼고(dropout) 학습시키는 방법 등이 있다.

02 Deep Learning의 내용

- 제 4장에서는 딥러닝 모형을 실제적으로 적용하기 위한 프로그램 언어인 keras를 설명한다.
- 딥러닝은 배우기 쉬운 python을 기반으로 만들어진 keras의 등장(2015년)과 함께 대중화되었다.
- keras에 의한 딥러닝은 Sequential API와 function API로 구현할 수 있다.
- Sequential API는 입력층→은닉층→출력층 순으로 일방통행식 딥러닝 아키텍처에 사용하는 API이고
- function API는 다중입력, 다중출력, 병렬형 은닉층, 비순환적 딥러닝 아키텍처 등 유연한 딥러닝 모형을 처리할 수 있는 API이다.

02 Deep Learning의 내용

- 제 5장에서는 CNN의 응용과 이전학습(transfer learning)을 다룬다.
- CNN모형은 MLP모형과 비교하여 추정해야 할 모수를 획기적으로 줄여줄 뿐만 아니라
- 동물의 귀와 같은 패턴을 추출하면 위치에 관계없이 동일하게 인식하는 위치 이동불변(translation invariant) 특성을 가지고 있다.
- 또한 입력층에 가까운 CNN층은 이미지의 경계선, 모서리, 가장자리 등 기초적인 특성을 찾아내고 CNN층이 반복될수록 점차적으로 이들 특성들을 조합하여 귀, 눈, 입 등의 이미지로 구체화하는 계층적 패턴인식의 특성을 가지고 있다.

02 Deep Learning의 내용

- 제 6장에서는 텍스트 자료분석에 대해 논의하고자 한다.
- 텍스트 자료분석에서의 특성변수의 수는 매우 크다는 특성을 가지고 있다.
- 단어의 순서와 의미가 중요한 머신번역과 같은 경우는 단어의 의미론적 수량화가 요구된다.
- 여기에서 단어의 의미론적 수량화란 예를 들어, 개와 고양이가 있고 늑대와 호랑이가 있을 때 개와 늑대는 개과 동물로 고양이와 호랑이는 고양이과 동물로 나눌 수 있어야 하며 동시에 개와 고양이는 길들여진 동물이고 늑대와 호랑이는 야생이라는 개념으로 분류할 수 있는 수량화를 의미한다.

02 Deep Learning의 내용

- 딥러닝은 단어와 단어 간의 동시 출현관계와 동시 출현단어 간의 거리로 단어의 의미론적인 수량화를 구현하고자 하며
- 이를 word embedding이라고 한다.
- 단어의 존재여부나 출현빈도로 수량화를 하면 각 표본은 1D텐서의 특성변수로 정의되고,
- 단어의 순서가 중요한 역할을 하는 머신번역에서는 각 표본은 시간스탭과 특성변수로 구성된 2D텐서로 정의된다.

02 Deep Learning의 내용

- 제 7장은 자율자동차의 이미지인식을 CNN모형을 통해 구현하고자 한다.
- 자율자동차를 운행하기 위해 교통표식의 인식, 특정 이미지를 분리하여 인식, 여러 개의 이미지를 동시에 인식하는 것 등은 필수적이며 정밀도 역시 매우 높아야 한다.
- 제 7장은 제 5장에 이어서 CNN 응용의 두 번째 사례이다.

02 Deep Learning의 내용

- 제 8장은 RNN의 응용으로, 머신번역과 같이 단어의 의미상의 이해를 위해 단어의 순서가 필수적이거나 시계열 자료분석에서 주식가격 예측을 위해 최근의 주식가격이 중요한 특성변수가 될 때 사용하는 딥러닝 모형의 응용이다.
- RNN모형의 입력데이터는 3D텐서이며 각 표본은 일정 시간동안 반복해서 관측된 특성변수로 구성된 2D텐서이다. 여기에서 일정시간을 시간스탭이라고 한다.
- RNN모형에서도 다른 기본 딥러닝 모형과 동일하게 입력 특성변수를 선형 및 비선형결합하게 되고 선형결합에 사용되는 모수를 추정한다.
- 그런데 이 모수는 시간스탭에 의존하지 않고 오직 시간스탭의 거리에만 의존한다는 사실을 기억해야 한다. RNN모형에서는 현재 시점과 한 시점전의 시간스탭 거리에 있는 정보만 사용한다.

02 Deep Learning의 내용

- 이러한 모수구조가 가능하기 위해서는 시계열로 주어진 특성변수가 소위, 정상성(stationarity)을 만족하여야 한다.
- 시계열이 정상성이 되기 위한 최소한의 조건은 특성변수의 평균과 분산이 시간 또는 시간스탭에 의존하지 않아야 한다.
- 주식가격 예측과 같은 경우에는 특수한 경우를 제외하고 정상성조건을 충족하지 못하므로 딥러닝 모형을 적용하기 이전에 정상성조건을 만족하는지 우선 점검하여야 하며,
- 정상성조건을 만족하지 못하면 정상성조건을 만족하도록 자료변환을 하여야 한다.
- 제 8장에서는 이러한 정상성조건을 포함하여 simple RNN, LSTM, GRU, 그리고 Bidirectional RNN 모형을 논의하게 될 것이다.

02 Deep Learning의 내용

- 제 9장에서는 다중 입·출력, 병렬형, 비순환 딥러닝 모형을 논의한다. 승용차 가격을 예측하고자 할 때, 입력자료로 전문가의 평가, 승용차사진, 그리고 차령이 있으면 다중 입력모형이 된다.
- SNS상의 글을 보고 글쓴이의 성별, 나이, 직업을 예측하고자 하면 다중 출력 모형을 사용하여야 한다.
- 다중 입·출력을 모형에 반영하기 위해서 딥러닝 아키텍처는 CNN, RNN, 또는 MLP 은닉층을 병렬형으로 구성할 수밖에 없게 되고,
- 입력층, 은닉층, 출력층의 순서를 자유롭게 할 수 있는 비순환 딥러닝 아키텍처도 고려되어야 한다.
- 제 9장에서는 이러한 모형을 논의하고 이를 실행하기 위한 keras의 function API를 적용할 것이다.

02 Deep Learning의 내용

- 제 10장에서는 제 9장의 논의를 바탕으로 머신번역을 위한 딥러닝 모형을 논의한다.
- 머신번역은 RNN 모형의 대표적인 응용분야이며 여러 개의 입력과 여러 개의 출력(many-to-many)을 가진 딥러닝 모형이다.
- 언어의 의미를 파악하여 같은 의미를 가진 다른 언어로 번역하는 인간의 번역 방식과 다르게,
- 머신번역은 특정언어의 단어가 주어졌을 때 단어가 가진 의미와 관계없이 다른 언어의 단어가 나올 확률이 가장 높은 단어로 번역하는 구조를 가지고 있다.
- 이러한 이유로 머신번역은 현재까지 딥러닝의 가장 큰 난제로 남아있다.

02 Deep Learning의 내용

- 입력변수가 입력되면 이로부터 잠재변수를 생성(encoder)하고 잠재변수가 주어진 조건에서 입력변수를 재생(decoder)하는 딥러닝 모형을 제 11장에서 논의할 것이다.
- encoder-decoder 모형은 잘 재생하는 것이 목적이지만,
- 또 다른 흥미로운 응용분야는 입력변수를 재생할 뿐만 아니라 창조적인 입력변수를 만들어 낼 수 있다면,
- 예를 들어 수많은 여성사진이 입력자료이면 이 여성사진을 그대로 생성하면서 동시에 원래 이미지에 머리칼, 피부색, 눈의 크기 등의 변화를 주어서 새로운 여성이미지를 생성하는 것도 흥미로운 분야일 뿐만 아니라 실제 상품으로도 사용할 수 있는 분야이다.

02 Deep Learning의 내용

- 제 12장는 제 11장과 동일하게 입력변수의 분포를 추정하는데 목적이 있지만 입력변수의 특성을 유지하면서 새롭고 창의적인 입력변수의 생성에 중점을 두고 있다.
- 예를 들어, 지폐위조범이 지폐를 위조하면 경찰은 위조지폐와 진짜지폐를 구분하게 되고, 지폐위조범은 경찰의 위조지폐 구별법을 통해 좀 더 정교한 지폐를 위조하여 궁극적으로 경찰이 위조지폐와 진짜 지폐를 구별하지 못하는 위조지폐를 만드는 과정과 유사한 딥러닝 모형을 구현하고자 한다.
- 여기에서 지폐위조범을 generator라고 하고 경찰을 discriminator라고 한다. generator와 discriminator가 서로 간에 적대적(adversarial)이기 때문에 generator와 discriminator로 구성된 딥러닝 모형을 GAN (generative adversarial networks)이라고 한다.

02 Deep Learning의 내용

- GAN모형의 손실함수는 입력변수분포와 GAN모형의 generator에서 재생한 입력변수분포의 차이를 측정하는 Kullback-Leibler divergence로 정의된다. 제 12장은 이에 대한 논의도 포함하고 있다.

02 Deep Learning의 내용

제 13장은 cross domian GAN모형을 다루고 있다. 도메인 A의 이미지를 도메인 B 이미지로 전환하고 반대로 도메인 B 이미지를 도메인 A 이미지로 전환하는 GAN 모형으로, 가장 큰 특징은 두 도메인의 이미지가 짝짓기(align)되어 있을 필요가 없다는 것이다.

화가의 그림을 사진이미지로, 위성사진을 지도로, 얼굴이미지를 이모티콘이나 커리컬처로, 흑색이미지를 컬러이미지로, 의료스캔 이미지자료를 실제 사진이미지로의 전환 등이 가능한 응용분야의 일부이다.

제 13장에서는 3가지의 응용사례를 다룰 것이며 가장 복잡하고 긴 프로그램이 제공될 것이다.

만약, 큰 어려움 없이 프로그램을 이해할 수 있다면 딥러닝 모형의 이론적 완성도는 어느 정도 갖추었다고 스스로를 평가해도 좋을 것이다.

02 Deep Learning의 내용

- 제 14장은 딥러닝 모형의 성능향상과 최적화를 위한 점검방법을 논의한다.
- 손실함수를 최소화하기 위한 역전파 과정에서 흔히 발생하는 미분값의 실종현상(vanishing gradients)을 방지하기 위한 여러 가지 방법을 논의하고
- 딥러닝 모형의 최적화를 점검하고 시각적 점검을 위한 keras 라이브러리 사용법을 논의할 것이다.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

03 Data Scientist

- 이러한 지식 중 통계학의 중요성은 2010년 1월 25일 Quora Digest의 "What should every data scientist master?"라는 질문에 대한 답변으로 "Statistics"라는 단답형 형태로 주어진 답변으로 대변할 수 있다.
- 통계적 지식 없이 오직 programming 지식만으로 만들어진 머신러닝이나 딥러닝 모형은 작동은 될 수 있을지라도, 데이터의 특성과 독립성이나 정상성이 결여된 잘못된 통계적 가정에서 구축된 모형은 없는 것보다 더 나쁜 결과를 초래할 수 있다.

03 Data Scientist

- 좋은 데이터 과학자가 되기 위해서는 세계에서 가장 큰 온라인 데이터과학 community인 Kaggle에 가입하여 경쟁에 참여해 보는 것도 좋은 방법 중의 하나이다.
- 새로운 연구결과 논문을 모아놓은 https://arxiv-sanity.com을 수시로 방문하여 최근의 연구동향을 파악하고,
- 특히 https://paperwithcode.com은 코드가 첨부된 논문이 제공되어 있어 다양한 분야의 최첨단 머신러닝/딥러닝 연구를 진행하는 데에 많은 도움을 받을 수 있다.
- https://colab.research.google.com에서 무료로 Jupyter Notebook 환경으로 딥러닝 모형(12시간 이내의 running time)을 실행할 수 있다.

Q & A