5주차 조별보고서 (Default)	
작성일: 2019년 10월 4일	작성자: 이재은
조 모임 일시: 2019년 10월 4일 9교시	모임장소: 학교 앞 카페
참석자: 위성조, 이충현, 최진성, 이재은, 김영연	조원: 위성조, 이충현, 최진성, 이재은, 김영연

구 분	내 용
학습 범위와 내용	2.1절: 선형대수 2.2절: 확률과 통계 2.3절: 최적화 이론
논의 내용	**<과제: 확률 문제 논의>** **Q1. 기계학습을 수강하는 37명의 학생들 중, 생일이 같은 사람의 확률은??** 1년이 365일이므로 생일의 가짓수는 365이다. 만약 366명 이상이 모이면 생일이 같은 두 사람이 반드시 존재한다. 즉, 366명 이상이면 생일이 같은 사람이 있을 확률은 무조건 100%이다. 확률을 구할 때, 여사건을 이용한 간접적인 방법으로 접근했다. 2명의 생일이 같지 않을 확률은 1명의 생일이 다른 1명의 생일을 제외한 다른 날짜이어야 하므로 364일 중 하루가 되어야 한다. 따라서, 1*(364/365)*(363/365)*100= 약 99.72%가 나온다. 이와 같은 방법으로 3명일 때는 99.17%, 4명일 때는 98.36%가 되면서 점점 감소하다가 37명일 때 같지 않을 확률이 15.13%이므로 37명중 생일이 같은 사람의 확률은 약 1-0.15= 85%이다.

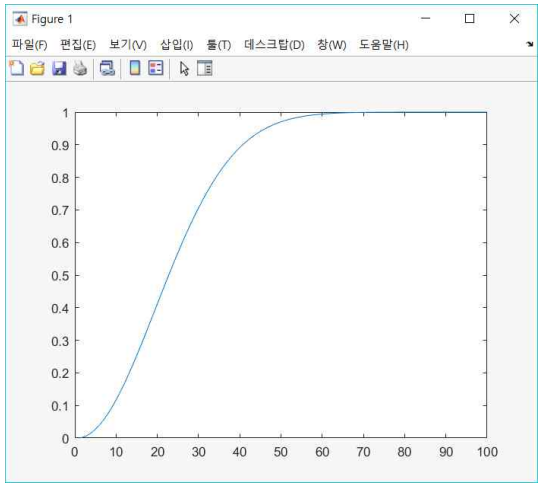
- MATLAB으로 만든 birthday 함수: 1에서 뺀 값을 넣어, 생일이 같을 확률을 계산

편집기 - C:₩Users₩이충현₩Documents₩MATLAB₩birthday.m

birthday.m

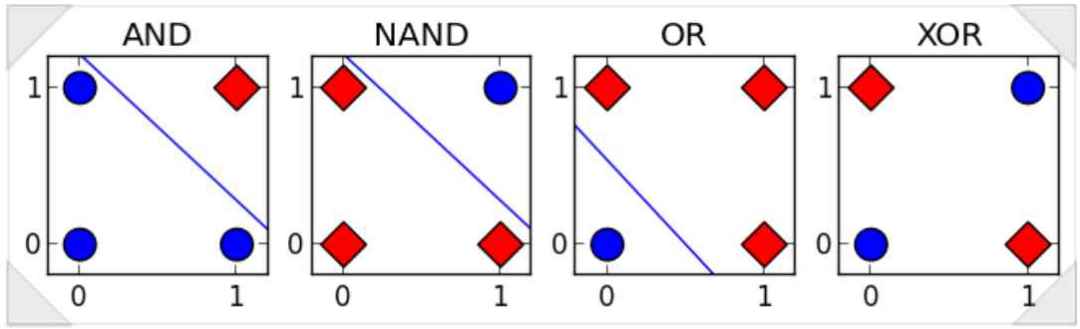
```
1   function [ A ] = birthday( n )
2
3 —   A = ones(n, 1);
4 —   p = 1;
5 —   for i=1:n
6 —       A(i) = 1 - p;
7 —       p = p * (365 - i)/365;
8 —   end
9
10 —  end
```

- 함수 호출 후, plot(A) 결과 값을 나타내는 그래프



X축은 사람의 수로 사람이 증가할수록 생일이 같을 확률이 높아짐을 알 수 있다.

	Q2. 어떤 사건 현장에 용의자가 2명이 있다. 2명 중 하나는 무조건 범인으로 A는 혈액형이 AB형, B는 모른다. 이 때 A가 범인일 확률은? (단, 일반적으로 AB형일 확률은 10%이다.) 베이즈 정리를 이용하여 문제를 풀었다. $P(\text{A가 범인일 확률}) = P(\text{B가 AB형이 아닐 확률}) + P((\text{A가 범인일 확률}) \mid (\text{B가 AB형일 확률}))$ $P(\text{B가 AB형이 아닐 확률}) = 1 - P(\text{B가 AB형일 확률}) = 1 - 0.1 = 0.9$ $P((\text{A가 범인일 확률}) \mid (\text{B가 AB형일 확률})) = 0.5 * 0.1 = 0.05$ $\therefore P(\text{A가 범인일 확률}) = 0.9 + 0.05 = 0.95$
	<조원들의 질문> **Q3. 충현: 단층 퍼셉트론은 AND, OR문제를 해결할 수 있지만 XOR문제에 대해서는 해결하지 못한다고 들었다. XOR에 대한 배경과 이에 대한 해결책이 무엇인지 궁금하다.** **A3.**

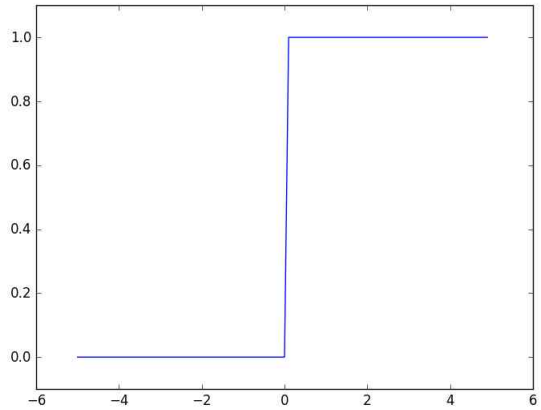
	단층 퍼셉트론은 다수의 신호(input)을 받아서 하나의 신호(output)으로 출력한다. 과거에는 이 퍼셉트론을 하드웨어를 이용하여 구현했다. 그 하드웨어만으로 당시에 중요하게 생각했던 AND와 OR문제를 해결할 수 있었다. 그 때문에 퍼셉트론을 제안했던 프랭크 로젠블래트 박사님은 1958년, 뉴욕 타임스에 위와 같은 이야기를 게재하였고, 이로 인해 인공지능은 그 당시 많은 사람들의 기대와 관심을 받았다. 이에 찬물을 끼얹은 문제가 XOR이다.  위의 그림과 같이 AND와 OR 문제는 적절한 경계선을 찾으면 입력에 대한 정답을 알 수 있었다. 그러나 XOR문제는 어떤 경계선을 찾든, 문제를 완벽하게 해결할 수 없었다. 실제로 1969년, MIT대학이 인공지능 연구실 Marvin Minskey교수는 '퍼셉트론즈'라는 책을 통해 XOR 문제 해결이 불가능하다는 것을 수학적으로 증명하며 단층 퍼셉트론 모델(SLP)의 취약점을 지적했다. 동시에 해결책 아닌 해결책을 제시했는데 퍼셉트론을 다층적으로 구현한 다층 퍼셉트론이 XOR 문제의 해결 방법이 될 것이라 이야기했다. 문제는 다층 퍼셉트론을 구성하는 퍼셉트론 개개별로 알맞은 weight와 bias값을 찾을 수 없다는 것이다. 이후로도 XOR문제는 풀리지 않았다.

Q4. 재은: 뉴런은 퍼셉트론과 다른 형태의 활성화 함수를 사용한다고 합니다. 그렇다면, 뉴런과 퍼셉트론이 어떤 차이가 있는지에 대해 자세히 알고 싶습니다.

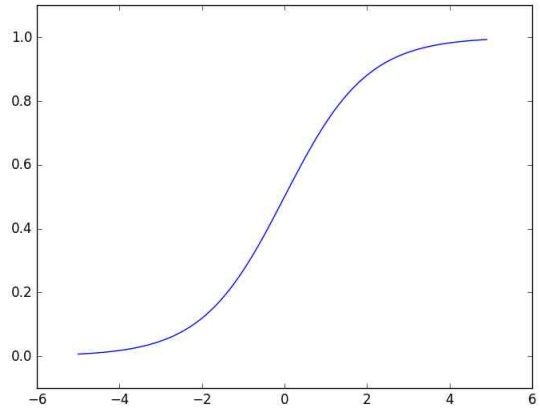
A4.

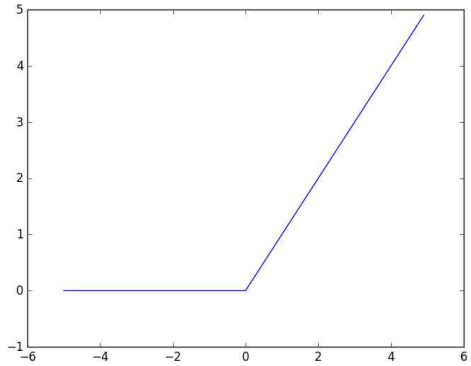
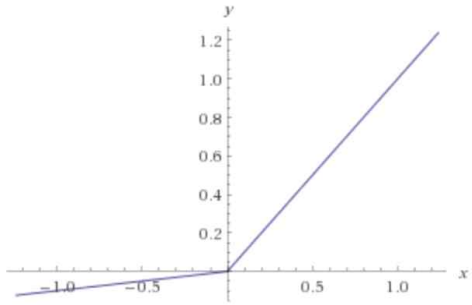
뉴런, 퍼셉트론 둘 다 활성화 함수를 사용한다. 퍼셉트론은 계단식의 단순한 함수를 사용하는 반면에, 뉴런은 연속된 값을 바탕으로 입력을 받아 연속적인 값의 출력이 가능한 활성화 함수를 이용한다.

우선, 퍼셉트론의 활성화 함수는 아래와 같은 계단식 활성화 함수를 사용한다.



이유는 입력된 값의 임계 값을 경계로 출력이 바뀌는 형태이기 때문이다. 그렇기 때문에 일

	정 구간마다의 출력 값이 변경되는 계단 형태의 함수를 이용한다. 그에 비해 뉴런과 같은 신경망의 경우 출력 값이 연속적으로 변화해야 하는 특징을 띠기 때문에 Sigmoid함수나 ReLU 함수 등, 출력 값이 연속적인 함수를 쓴다. Sigmoid 함수는 다항 회귀 출력을 확률로 매핑하여 0 ~ 1 사이의 값을 반환하는 함수이며 지금은 많이 쓰이지 않지만 한 때 가장 많이 사용이 된 활성화 함수이다. 뉴런의 시그모이드 함수의 계산식과 그래프는 아래와 같다. $$f(t) = \frac{1}{1+e^{-t}}$$ 

	ReLU 함수는 입력이 0 이하일 경우 0을, 0 이상일 경우 입력 값을 출력하는 함수로, 음수를 차단하는 역할을 한다. 참고로, 최근에 나온 Leaky ReLU는 x 값이 음수일 경우 입력값의 1/10만을 출력하는, 음수에서도 작동 가능하도록 설계된 ReLU 함수 또한 존재한다.  

질문 내용	이산 확률분포 $H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i)$ 연속 확률분포 $H(x) = -\int_{\mathbb{R}} P(x)\log_2 P(x)$ 엔트로피를 구하는 과정에서 연속확률분포든 이산확률분포든 항상 밑이 2인 로그를 취하는데, 단순히 계산 결과 밑이 2인 로그가 계산식에 가장 근접해서 사용하는 것인지, 아니면 다른 이유가 있는 것인지 궁금합니다.
기타	

<첨부 개인 레포트>

- 201402033 위성조

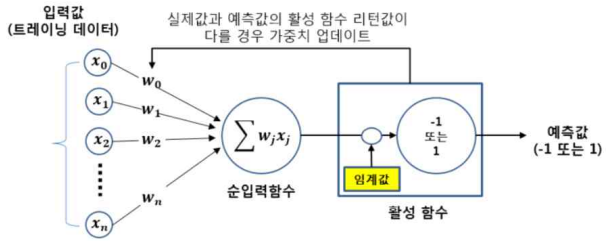
구분	내용
학습 범위	기계 학습 2장 2.1 선형대수 2.2 확률과 통계 2.3 최적화 이론
학습 내용	기계학습에서 수학의 역할 – 목적함수를 정의하고, 목적함수가 최저가 되는 점을 찾아주는 최적화 이론 제공 최적화 이론에 규제, 모멘텀, 학습률, 멈춤조건과 같은 제어를 추가하여 알고리즘 구축 벡터 – 샘플을 특징벡터(feature vector)로 표현 행렬 – 여러 개의 벡터를 담은 것. 훈련집합을 담은 행렬을 설계행렬이라 부른다 행렬 A의 전치행렬 는 A의 원소 를로변환한것. 다항식을 행렬을 이용하여 간단히 표현할 수 있다. 정사각행렬 – 행렬의 행과 열의 수가 같은 행렬 대각행렬 – i와j가 같은 위치에만 성분이 있는 행렬 단위행렬 – i와j가 같은 위치에만 1이 있는 행렬 대칭행렬 – i와j가 같은 위치를 라인으로 성분이 대칭인 행렬 벡터의 내적 텐서 - 3차원 이상의 구조를 가진 숫자 배열 ex)3차원 구조의 RGB 컬러 영상

	벡터의 내적은 L2-norm과 같으며, 보통 특정 표기가 없는 경우 L2-norm으로 간주한다. 또한 유클리드 norm으로 불린다 행렬의 L2-norm은 프로베니우스 norm으로 불린다. 각 원소를 제곱하여 더한 값에 루트를 적용하여 구한다. 유사도와 거리 – 벡터를 기하학적으로 해석 코사인 유사도 cosine_similarity(a,b) = 퍼셉트론 – 활성함수로 계단함수를 사용했음. 딥 러닝은 퍼셉트론을 여러 층으로 확장하여 만듦 벡터 – 공간상의 한 점으로 화살표 끝이 벡터의 좌표에 해당 정부호 행렬 양의 정부호 행렬 : 0이 아닌 모든 벡터 X에 대해 양의 준정부호 행렬 : 0이 아닌 모든 벡터 X에 대해 음의 정부호 행렬 : 0이 아닌 모든 벡터 X에 대해 음의 준정부호 행렬 : 0이 아닌 모든 벡터 X에 대해 확률 곱 규칙 : = 합 규칙 : = 베이즈 정리 베이즈 정리의 해석

	사후확률을 지접 추정하는 일은 아주 단순한 경우를 빼고 불가능하므로 베이즈 정리를 이용하여 추정함. 사전확률은 를 이용하여 추정 우도 = likelihood, 뭐가 제일 그럴듯한가? 미분 – 기계학습에서 프로그램이 최저점을 찾기 위해 가장 많이 쓰이는 방식 랜덤한 시작점에서 미분값이 가장 큰 방향으로 전진해 나가며, 이를 경사하강법이라 한다.
질문 내용	

- 201402665 이충현

구분	내용
학습 범위	기계학습 2장 2.1 선형대수 2.2 확률과 통계 2.3 최적화 이론
학습 내용	<2.1 선형대수> ● 퍼셉트론 알고리즘은 지도학습이나 분류의 맥락에서 볼 때, 하나의 샘플이 어떤 클래스에 속해 있는지 예측하는데 사용될 수 있다. 순입력 함수의 결과값을 특정 임계값과 비교를 하고, 순입력 결과값이 크면 1, 그렇지 않으면 -1로 출력하는 함수를 정의한다(활성함수).



● 행렬 분해는 고유값과 고유 벡터가 있는데, 선형변환에 의한 변환 결과가 자기 자신의 상수배가 되는 0이 아닌 벡터를 고유벡터(eigen_vector)라고 한다. 이 상수배 값을 고유값(eigen_value)이라 한다.

$$A\mathbf{v}=\lambda\mathbf{v} \quad \text{---(1)}$$

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}=\lambda\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \quad \text{--- (2)}$$

<2.2 확률과 통계>

● 베이즈 정리(매개 변수를 모르는 상태에서 매개 변수를 추정하는 문제에 적합)

$$P(y,x) = P(x|y)P(y) = P(x,y) = P(y|x)P(x)$$

$$\longrightarrow P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (2.26)$$

● 유용한 확률 분포로는 가우시안 분포, 이항 분포, 베르누이 분포가 있다.

	가우시안 분포à평균과 분산으로 정의 이항 분포à성공 확률이 p인 베르누이 실험을 m번 수행할 때 성공할 횟수의 확률분포 베르누이 분포à성공(x=1)확률p이고 실패(x=0)확률이 1-p인 분포 ● 정보이론에서 동전 던지기를 예를 들면, 정상적으로 생각했을 때, 비정상적 동전 던지기(Head1/4, Tail3/4이라고 가정)가 불확실성이 더 적고 따라서 정보량도 적어야 한다. 그러나 실제 정보량을 계산하면 이벤트에 대한 전체 정보량은 단순합이 아니라는 것을 알게된다. 따라서 기대값이란 결과의 평균값을 생각하면서 정보량에 확률을 곱하는 방식이 앤트로피다. 이렇게 하면 비정상동전을 가지고 던질 경우, 정상동전보다 불확실성이 해소되는 양이 적다. 즉 앤트로피란 확률분포를 따르는 변수의 평균정보량이 된다. $H = -\sum_{i=1}^{n} p_i \log p_i$ <2.3 최적화 이론> ● 기계 학습은 적절한 모델을 선택하고, 목적함수를 정의하고, 모델의 매개변수 공간을 탐색하여 목적함수가 최저가 되는 최적점을 찾는 전략을 사용한다. 최적화는 예측 단계가 아니라 학습 단계에 필요하다.
질문 내용	단층 퍼셉트론은 AND, OR문제를 해결할 수 있지만 XOR문제에 대해서는 해결하지 못한다고 들었다. XOR에 대한 배경과 이에 대한 해결책이 무엇인지 궁금하다.

- 201403474 최진성

구분	내용
학습 범위	기계학습 2장 기계 학습과 수학 2.1 선형대수 2.2 확률과 통계
학습 내용	기계학습 2장, 기계 학습과 수학 2.1 선형대수 2.1.1 벡터와 행렬 벡터 : 크기랑 방향을 가지는 어떠한 물리량을 의미 특징 벡터 : '인식 대상'이 되는 객체를 특징으로 하여 차원을 가진 벡터로 표현한 것. 기계학습에서의 사용은 '샘플'을 특징 벡터로 표현하며, 여러 개의 특징 벡터를 나열할 경우 첨자로 구분한다. 행렬 : 1개 이상의 수나 식을 사각형의 배열로 나타낸 것. 벡터 행렬 : 여러 개의 벡터를 담은 행렬을 의미하며, 훈련 집합을 담은 행렬을 '설계 행렬'이라고 칭한다. 전치 행렬 : 행렬 내의 원소를 대각선 축을 기본으로 하여 서로 위치를 바꾼 것. M * N 행렬을 전치하면 N * M이 된다. $$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}^T = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{bmatrix}$$ - 전치 행렬 그 외의 행렬 -정사각행렬 – 행렬이 N*N의 형태로 정사각형을 이루는 행렬 -대각행렬 – M*N의 행렬에서 값이 대각선을 이루어 존재하는 행렬 -단위행렬 – M*N의 행렬에서 값이 1, 또는 0만 존재하는 행렬 -대칭행렬 – 대각선을 기준으로 하여 양 쪽의 값이 동일한 행렬

행렬의 계산

$$c_{ij} = \sum_{k=1,s} a_{ik} b_{kj}$$

C = AB(j)이며 로 표현
행렬의 계산은 분배법칙은 성립하나 교환 법칙은 성립하지 않는다.

AB != BA

A(BC) => AB = BC

텐서
2차원 행렬이 쌓여서 3차원 형태를 이루는 것을 말한다.

놈
'크기'를 일반화 한 것으로, 벡터와 행렬의 크기를 놈 형태로 바꾸어 측정할 수 있다.
*거리를 일반화 한 것은 거리함수

$$p\text{차 놈: } \|\mathbf{x}\|_p = \left(\sum_{i=1,d} |x_i|^p \right)^{\frac{1}{p}}$$

$$\text{최대 놈: } \|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \cdots, |x_d|)$$

프로베니우스 놈
놈은 p를 1, 2, 혹은 무한대를 많이 사용하는데, 여기에서 p = 2가 되는 놈을 프로베니우스 놈이라고 하며 ||A||(F)로 표기하기도 한다.
계산식은 다음과 같다

$$\left(\sum_{i=1,n} \sum_{j=1,m} a_{ij}^2 \right)^{\frac{1}{2}}$$

베이즈 정리
본래 역확률을 구하고자 하는 정리로, P(BIA) 주어진 상태에서 P(AIB)를 구하고자 하는 목적으로 사용되었다. 하지만 알고리

	즘과 기계학습의 대두로 인해서 베이즈정리는 이전의 경험과 현재의 증거를 토대로 '어떤 사건의 확률을 추론'하는 알고리즘으로서의 가치를 가지기 시작했다. 특히 기계학습에서의 베이즈 정리는 '사람이 생각하고 판단하는 근본적인 방식'이라고 여겨지기도 한다. 즉 처음에는 아무 정보가 없는 상태에서 새로운 정보들을 모아 자신이 가지고 있는 사건 확률 체계를 갱신하여 환경을 해석하거나 판단하는 방향으로 발전한다고 여기기도 한다. '사건 B가 일어난 것을 전제로 한 사건 A의 조건부 확률' $$P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$$ 베이즈 정리의 식에 대한 해석 $$\overbrace{P(y|x)}^{\text{사후확률}} = \frac{\overbrace{P(x|y)}^{\text{우도}}\ \overbrace{P(y)}^{\text{사전확률}}}{P(x)}$$ # 용어 정리 ● 사전 확률 : 이미 알고 있는 사건(들)의 확률 ● 사후 확률 : 사전 확률과 우도 확률을 통해서 알게 되는 새로운 확률 ● 우도 : 이미 알고 있는 사건(들)이 발생했다는 조건 하에 다른 사건이 발생할 확률 최대 우도 데이터 X가 주어졌을 때 X를 발생시켰을 가능성을 최대로 하는 매개변수 Θ = {g(3)}을 찾는 것. 기계학습에서는 보통 조건부 우도를 최대화 하는 방식으로 학습을 하기 때문에, 모델 X를 넣었을 때 실제 Y에 가깝게 반환하는 Θ를 찾는 것이 관건이라고 할 수 있다. 최대우도추정 기법으로 추정한 모수는 일치성과 효율성을 가진다고 할 수 있다. 일치성이란 추정에 사용하는 표본의 크기 X가 커질 수록 진짜 모수값 Y에 수렴하는 특성을 가리키고, 효율성이란 일치성 등에서 같은 추정량 가운데서도 분산이 작은 특성을 나타낸다. 추정량의 효율성을 따질 때는 보통 평균제곱오차를 기준으로 하는데, 일치성을 가진 추정량 가운데 최대우도추정량보다 낮은 평균제곱오차를 지닌 추정량이 존재하지 않는다고 한다.

평균 벡터
평균 벡터 E(X)는 각 확률 변수들의 평균들을 담아 놓은 것을 의미한다.

공분산 행렬
공분산 – 서로 다른 두 확률 변수의 상관관계를 나타내며 두 확률변수의 편차의 곱으로 구해진다.
즉, 공분산 행렬은 각 정방 행렬의 값을 각 변수의 분산과 공분산으로 채워 넣은 것을 의미한다.

가우시안 분포
정규 분포라고도 하며, 현재 모든 자연 현상을 매우 잘 표현하고 있는 이상적인 확률 모형이다.
정의 :

$$N(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

베르누이 분포
매 시행마다 오직 두 가지의 결과만 나온다고 가정할 때 쓰는 확률 분포로, 성공을 x = 1, 실패를 x = 0으로 가정할 경우 다음과 같은 공식을 따른다.

$$Ber(x;p)=p^x(1-p)^{1-x}=\begin{cases}p, & x=1\text{일 때}\\ 1-p, & x=0\text{일 때}\end{cases}$$

이항 분포
베르누이 분포를 이용한 것으로, 성공 확률이 p인 실험을 m번 수행할 때 성공하는 횟수의 확률 분포이다.
주요한 특징으로는 m이 어느정도까지 커지면 다시 가우시안 분포를 따른다는 점이다.

$$B(x;m,p)=C_m^x p^x(1-p)^{m-x}=\frac{m!}{x!\,(m-x)!}p^x(1-p)^{m-x}$$

정보 이론
정보 이론의 기본 원리는 '확률이 작을수록 많은 정보'라는 점이다. 예시로 '알래스카에 눈이 왔다'와 '캘리포니아에 눈이 왔다'를 비교할 경우 눈이 내리는 빈도가 알래스카가 더 많기 때문에 상대적으로 확률이 작은 '캘리포니아에 눈이 왔다'가

	더 많은 정보라고 할 수 있다. 엔트로피 분포 p를 나타내는 H(X)의 불확실성을 측정한 것을 의미하며, 0에 가까울수록 불확실성이 적어지지만, 반대의 경우 불확실성이 매우 높아진다. 확률 번수 y가 이산확률분포를 따를 경우엔 다음의 공식을 따른다. $$H[Y] = -\sum_{k=1}^{K} p(y_k) \log_2 p(y_k)$$ 확률 번수 y가 정규분포와 같은 연속확률분포을 따를 경우엔 다음의 공식을 따른다. $$H[Y] = -\int_{-\infty}^{\infty} p(y) \log_2 p(y)\, dy$$ * 단, p(y)는 밀도 함수이다. 교차 엔트로피 두 확률 분포 P와 Q에 대하여, H(P, Q)와 H(Q, P)는 다른 결과를 가진다. 그러므로 두 개의 엔트로피를 비교하여 오차를 최소한으로 하는 것을 교차 엔트로피 오차라고 한다. 기본적으로 로그의 밑이 e인 자연로그를 예측값에 씌워서 실제 값과 곱한 후 전체 값을 합한 후 음수로 변환한다. 값의 출력이 1일 때 0이 되며 x가 커질수록 0에 가까워지고 x가 작아질수록 (0에 가까워질수록) 값이 작아진다. KL-Divergence 상대 엔트로피라고도 불리며, 두 확률 분포의 차이를 비교하기 위해 사용된다.

	$$KL(P \parallel Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$ P는 주로 참 분포, 실제 관찰 데이터를 의미하지만 Q는 가설, 모델, P의 근사 등으로 사용된다. 그러므로 항상 0 이상의 값을 가지며, 두 확률 분포가 동일할 경우 0을 가지게 된다. KL-Divergence를 분해하게 되면 정보 엔트로피와 교차 엔트로피로 나눌 수 있게 된다. $$D_{KL}(P\|\|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$ $$= \sum_i P(i) log P(i) - \sum_i P(i) log Q(i)$$ $$= -H(P) + H(P, Q)$$
질문 내용	엔트로피를 구하는 과정에서 연속확률분포든 이산확률분포든 항상 밑이 2인 로그를 취하는데, 단순히 계산 결과 밑이 2인 로그가 계산식에 가장 근접해서 사용하는 것인지, 아니면 다른 이유가 있는 것인지 궁금합니다.

- **201502469 이재은**

<table>
<tr><td>구분</td><td>내용</td></tr>
<tr><td>학습 범위</td><td>2.1절: 선형대수
2.2절: 확률과 통계
2.3절: 최적화 이론</td></tr>
<tr><td>학습 내용</td><td>-기계학습에서 수학은 목적 함수를 정의하고, 목적 함수가 최저가 되는 점을 찾아주는 최적화 이론을 제공함에 있어 중요한 역할을 하므로 이 절에서는 수학에 대한 기초를 배운다.

-Iris 데이터는 영국 통계 학자이자 생물학자 Ronald Fisher가 1936년 논문에서 발표한 다 변수 데이터 세트이다. 이는 데이터를 여러 변수들의 선형결합으로 표현하였을 때 서로 다른 그룹을 잘 구분할 수 있게 해 주는 coefficient 를 찾는 방법인 Fisher's linear discriminant(FLD)를 만든 사람과 동일한 사람이 맞다.

-딥러닝에서 텐서의 개념

<table>
<tr><th>rank</th><th>type</th><th>example</th></tr>
<tr><td>0</td><td>scalar</td><td>[1]</td></tr>
<tr><td>1</td><td>vector</td><td>[1,1]</td></tr>
<tr><td>2</td><td>matrix</td><td>[[1,1], [1,1]]</td></tr>
<tr><td>3</td><td>3 tensor</td><td>[[[1,1], [1,1]], [[1,1], [1,1]]]</td></tr>
<tr><td>n</td><td>N tensor</td><td></td></tr>
</table>


Rank가 0인 tensor는 scalar, Rank가 1인 tensor는 vector, rank 1 tensor를 item으로 갖는 tensor를 matrix라 부른다.이 matrix를 item으로 갖고 있는 tensor를 3 tensor라 부른다. 즉, 이런 식으로 tensor가 n-tensor로 증가하며 n차원 배열을 의미한다.</td></tr>
</table>

	-신경망은 연결주의(connectionist) 계산 모형이다. 인공신경망의 종류 중 하나인 퍼셉트론은 출력 값이 1 또는 0(또는 -1)이기 때문에 선형 분류 모형으로도 볼 수 있다. -베이즈 정리는 이전의 경험과 현재의 증거를 토대로 어떤 사건의 확률을 추론하는 과정이다. 이 과정에서 사전 확률과 사후 확률 사이의 관계를 조건부 확률을 이용하여 계산한다. (conditional probability)이란 사건 B가 일어나는 경우에 사건 A가 일어날 확률을 말한다. 사건 B가 일어나는 경우에 사건 A가 일어날 확률은 P(A|B)로 표기한다. -확률이 낮을수록 가치가 높은데, 여기서 그것의 기댓값을 엔트로피라 한다. 예를들어 앞뒤가 다른 동전이 있는데 $H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$ 그것이 나올 확률이 각각 [50%, 50%] , [100%, 0%], [90%, 10%]라 하면 엔트로피공식 으로 계산한 결과는 1, 0, 0.47이 나온다. 높은 엔트로피는 불확실성이 높은 것을 의미하고, 낮은 엔트로피는 불확실성이 낮은 것을 의미하는 것을 알 수 있다. -배치란 모델을 학습 할 때 반복 1회당 사용되는 예(data)의 총 개수이다. 여기서 반복은 정해진 배치 크기를 이용하여 학습을 반복하는 횟수를 말한다. 배치 경사 하강법은 모든 예를 다 계산해야 하므로 시간이 많이 소요된다. 반면에, 스토캐스틱()은 배치와 trade-off 관계로서 반복 당 하나의 예(배치 크기 1)만을 사용한다. 반복이 충분하면 스토캐스틱 경사 하강법이 효과적일 수 있지만 노이즈가 심하다. Stochastic은 확률적 이라는 의미로 각 배치를 포함하는 하나의 예가 무작위로 선택된다는 것이다.
질문 내용	뉴런은 퍼셉트론과 다른 형태의 활성화 함수를 사용한다고 합니다. 그렇다면, 뉴런과 퍼셉트론이 어떤 차이가 있는지에 대해 자세히 알고 싶습니다.

- 201500629 김영연

구분	내용
학습 범위	Chapter 2 기계학습과 수학
학습 내용	1. 기계학습에서의 수학 1) 선형대수: 이 분야의 개념을 이용하면 학습모델의 매개변수 집합, 데이터, 선형연산의 결합 등을 행렬 또는 텐서로 간결하게 표현할 수 있다. 데이터를 분석하여 유용한 정보를 알아내거나 특징 공간을 변환하는 등의 과업을 수행하는데 핵심 역할을 한다. 2) 확률과 통계: 데이터에 포함된 불확실성을 표현하고 처리하는데 활용한다. 베이즈 이론과 최대 우도 기법을 이용하여 확률 추론을 수행한다. 3) 최적화: 목적함수를 최소화하는 최적해를 찾는데 사용하며, 주로 미분을 활용한 방법을 사용한다. 수학자들이 개발한 최적화 방법을 기계학습이라는 도메인에 어떻게 효율적으로 적용할지가 주요 관심사이다. -설계행렬: 훈련집합을 담은 행렬 -행렬은 교환법칙이 성립하지 않는다. -행렬은 분배법칙과 결합법칙은 성립한다. -벡터의 내적: -텐서: 3차원 이상의 구조를 가진 숫자 배열 -벡터의 p차 놈: -최대 놈: -퍼셉트론

	- -양의 정부호 행렬: 0이 아닌 모든 벡터x에 대해 양의 준정부호 행렬: 0이 아닌 모든 벡터 x에 대해 음의 정부호 행렬: 0이 아닌 모든 벡터 x에 대해 음의 준정부호 행렬: 0이 아닌 모든 벡터 x에 대해 -고유 값과 고유 벡터: 각 고유 값에는 그에 대응하는 고유벡터가 있고 하나의 행렬은 고유치와 고유벡터에 의해 분해될 수 있다. AV = -고윳값 분해: Q는 A의 고유 벡터를 열에 배치한 행렬이고 고유값 분해는 정사각행렬에만 적용 가능한데, 기계학습에서는 정사각행렬이 아닌 경우의 분해도 필요하므로 고윳값 분해는 한계를 가진다. -특잇값 분해(SVD): SVD는 대각화에 유용한 방법이다. U는 로 나온 행렬을 고윳값 분해를 해서 얻은 고유벡터들을 열벡터로 처리하여 행렬로 묶은, 직교행렬이다. V는 로 나온 행렬을 고윳값 분해를 해서 얻은 고유벡터들을 열벡터로 처리하여 행렬로 묶은, 직교행렬이다. U나 v의 고윳값들을 루트를 취해서 나온 특이값들을 대각요소로 넣은 대각행렬이다. -SVD의 장점: sudo 역행렬 구하는 것에 있어서는 SVD가 최고로 안정적이고, 선형연립방정식을 쉽게 풀 수 있습니다. 또, 모든 SVD 분해할 수 있다. -베이즈 정리: 두 확률 변수의 사전확률과 사후확률 사이의 간계를 나타내는 정리. 사전확률로부터 사후확률을 도출할 수 있다. p(A)는 사건 A의 사전확률, p(A|B)는 B값이 주어졌을 때 나타나는 A의 사후확률, P(B|A)는 A가 주어졌을 때 나타나는 B의 조건부 확률 P(B)는 사건 B의 사건확률을 사전확률을 나타낸다. -최대 우도: 매개변수를 모르는 상황에서 매개변수를 추정하는 문제

	$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log P(\mathbb{X}\|\Theta) = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(\mathbf{x}_i\|\Theta)$ -공분산 행렬: 대칭행렬을 가짐 -가우시안 분포: 도수 분포 곡선이 평균 값을 중심으로 좌우 대칭을 이루는 것. 모든 측정에서 중복으로 실험했을 경우 결과 값이 똑같이 이루어지는 경우는 없으며, 약간의 오차를 수반하게 된다. -베르누이 분포: 성공 확률을 p라고 한다면 성공하지 못할 확률은 1-p이다. 어떤 일을 한번 수행할 때 p의 확률로 성공하면 1이라는 숫자를 부여하고 1-p의 확률로 실패를 하면 0이라는 숫자를 부여하는 것 -이항 분포: 성공 확률이 p인 베르누이 실험을 m번 수행할 때 성공할 횟수의 확률분포 -엔트로피: 확률변수 x의 불확실성을 나타내는 엔트로피 따라서 불확실성이 높을 때 우리는 엔트로피가 높다고 말한다. (주사위가 윷보다 엔트로피가 높음) -교차 엔트로피: 두 확률분포 p와 Q사이의 교차 엔트로피 -KL다이버전스: 두 확률분포의 거리를 계산할 때 사용 KL(P||Q) =
질문 내용	