04. 기계학습과 수학

지난주까지는 기계학습을 위해 기본적으로 알아야 할 특징 공간, 데이터, 모델, 규제 등을 배웠습니다.
이번주는 다소 수학적인 내용이지만, 쉽게 이해하면 될 것 같습니다.

선형대수

벡터(Vector)

1개의 **샘플 데이터**는 특징 공간(feature space)에서 1개의 포인터로 표현됩니다.-->그래서 특징벡터라고 불립니다.

$$\mathbf{x} = (x_1, x_2, x_3, x_4)^T = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}$$

열벡터
행벡터 (5.1, 3.5, 1.4, 0.2)

여러 개의 특징벡터는 첨자로 구분합니다.

$$\mathbf{x}_1 = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 4.9 \\ 3.0 \\ 1.4 \\ 0.2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 4.7 \\ 3.2 \\ 1.3 \\ 0.2 \end{pmatrix}, \cdots, \mathbf{x}_{150} = \begin{pmatrix} 5.9 \\ 3.0 \\ 5.1 \\ 1.8 \end{pmatrix}$$

$X = \{x_1, x_2, x_3, \ldots, x_{150}\}$

행렬(Matrix)

벡터를 모아놓은 곳을 행렬이라고 말합니다.
이때 **학습데이터를 담은 행렬 --> 설계행렬**

(예) Iris 설계행렬(design matrix)

$$\mathbf{X} = \begin{pmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ \vdots & \vdots & \vdots & \vdots \\ 6.2 & 3.4 & 5.4 & 2.3 \\ 5.9 & 3.0 & 5.1 & 1.8 \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} \\ x_{4,1} & x_{4,2} & x_{4,3} & x_{4,4} \\ \vdots & \vdots & \vdots & \vdots \\ x_{149,1} & x_{149,2} & x_{149,3} & x_{149,4} \\ x_{150,1} & x_{150,2} & x_{150,3} & x_{150,4} \end{pmatrix} \leftarrow \text{행}^{row}$$

열column

행렬의 **행과 열을 교환**한 행렬 --> **전치 행렬**(transpose matrix)

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, \quad \mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}$$

(예) $\mathbf{A} = \begin{array}{c|c} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{pmatrix} 3 & 4 & 1 \\ 0 & 5 & 2 \end{pmatrix} \end{array} \quad \mathbf{A}^T = \begin{array}{c|c} & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} & \begin{pmatrix} 3 & 0 \\ 4 & 5 \\ 1 & 2 \end{pmatrix} \end{array}$

행렬을 이용하면 수학을 간결하게 표현할 수 있습니다.

$$\begin{aligned} f(\mathbf{x}) &= f(x_1, x_2, x_3) \\ &= 2x_1x_1 - 4x_1x_2 + 3x_1x_3 + x_2x_1 + 2x_2x_2 + 6x_2x_3 - 2x_3x_1 + 3x_3x_2 + 2x_3x_3 + 2x_1 + 3x_2 - 4x_3 + 5 \\ &= \boxed{\underbrace{(x_1 \quad x_2 \quad x_3)}_{x^T} \underbrace{\begin{pmatrix} 2 & -4 & 3 \\ 1 & 2 & 6 \\ -2 & 3 & 2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{x} + \underbrace{(2 \quad 3 \quad -4)}_{b^T} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{x} + 5} \\ &= \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c \end{aligned}$$

특수한 행렬 입니다. #몰라도 된다고 합니다. 단위행렬만 압시다.

정사각행렬 $\begin{pmatrix} 2 & 0 & 1 \\ 1 & 21 & 5 \\ 4 & 5 & 12 \end{pmatrix}$, 대각행렬 $\begin{pmatrix} 50 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 8 \end{pmatrix}$,

단위행렬 $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, 대칭행렬 $\begin{pmatrix} 1 & 2 & 11 \\ 2 & 21 & 5 \\ 11 & 5 & 1 \end{pmatrix}$

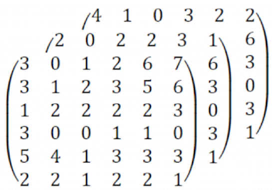
행렬은 교환법칙이 성립되지 않습니다. (결합법칙(0), 분배법칙(0))

[핵심07] **행렬의 곱셈 - 행렬 A의 열과 B의 행의 개수가 같아야 합니다.**

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

2*3 행렬 $\mathbf{A} = \begin{pmatrix} 3 & 4 & 1 \\ 0 & 5 & 2 \end{pmatrix}$와 3*3행렬 $\mathbf{B} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 0 & 5 \\ 4 & 5 & 1 \end{pmatrix}$을 곱하면 2*3 행렬 $\mathbf{C} = \mathbf{AB} = \begin{pmatrix} 14 & 5 & 24 \\ 13 & 10 & 27 \end{pmatrix}$

3차원 이상의 구조를 갖는 배열 --> **텐서**(tensor)

$\mathbf{A} =$ 

1차원 행렬(tensor): vector
2차원 행렬(tensor): matrix
3차원 행렬(tensor): cube = tensor
n차원 행렬(tensor): tensor

=> tensor: 행렬과 같은 배열 구조

행렬 A와 **곱하면** <u>**단위행렬I**가 나오는 행렬</u> --> **역행렬**(inverse matrix)
정방행렬에 대해서만 정의되며, 역행렬이 없다면 특이행렬로 분류됩니다. # 교수님이 어렵습니다. 고 하셨습니다.

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \mathbf{A}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$ --> 잘 알고 있는 공식이라는데

(예)

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 6 & 4 \end{pmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{2*4-1*6}\begin{pmatrix} 4 & -1 \\ -6 & 2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 4 & -1 \\ -6 & 2 \end{pmatrix} = \begin{pmatrix} 2 & -0.5 \\ -3 & 1 \end{pmatrix}$$

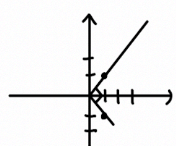
행렬식 - 어떤 행렬의 역행렬 존재여부에 대한 판별값 --> det의 값이 0이면 역행렬이 없습니다.

[핵심08] 고유벡터(v)와 고유값(람다) $\mathbf{Av} = \lambda\mathbf{v}$

어떤 행렬 A에 대하여 벡터(v)를 곱했더니, 람다와 같아졌습니다. v는 고유벡터입니다.
고유벡터(v)는 <u>벡터의 방향</u>을 나타내고, 고유값(람다)은 <u>벡터의 길이</u>를 나타냅니다.

(예) $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

- $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\qquad \lambda = 3, \mathbf{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
- $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ $\qquad \lambda = 1, \mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$



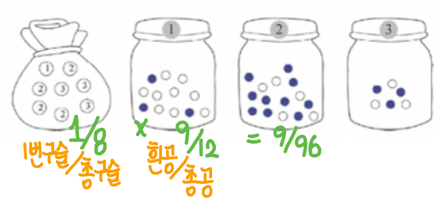
m*m행렬은 최대 m개의 고유벡터와 고유값을 가질 수 있습니다.

확률과 통계 #기계학습과 관련된 부분만 다룰 겁니다.
확률분포 - 정의역 전체의 확률을 표현
정의역 - 확률을 표현하는 변수가 가질 수 있는 값의 범위 #확률변수: 확률을 수식으로 표현하기 위한 변수
결합확률 - 두 사건이 결합된 상태의 확률 $P(y,x)$

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

- $P(x|y)$: 조건부확률(conditional probability)

(문제) 주머니에서 1, 1번 병에서 흰 공을 꺼낼 확률



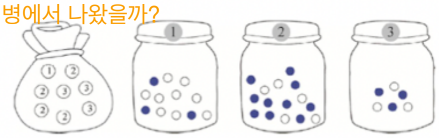
[핵심09] 베이즈정리

일반적으로 **x와 y가 동시에 일어날 결합확률**과 **y와 x가 동시에 일어날 결합확률**이 **같습니다.**

$$\boldsymbol{P(y,x) = P(x|y)P(y) = P(x,y) = P(y|x)P(x)}$$

$$\boxed{\underbrace{\boldsymbol{P(y|x)}}_{\text{사후확률}} = \frac{\overbrace{\boldsymbol{P(x|y)}}^{\text{우도(likehood)}}\underbrace{\boldsymbol{P(y)}}_{\text{사전확률}}}{\boldsymbol{P(x)}}}$$

(문제) 주머니에서 숫자를 뽑은 번호 병에서, 흰공이 나왔다. 어느 병에서 나왔을까?



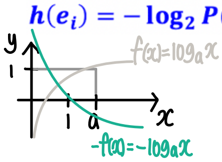
<u>**(1) 사전확률과 (2) 우도를 구할 수 있다면, 사후확률을 간접적으로 계산할 수 있습니다.**</u>

평균과 분산
특징벡터는 평균과 공분산으로 계산 합니다.

정보이론과 자기정보
여러가지 형태로 들어오는 정보의 **정보량**을 **수치**로 나타내봅시다.
기본원리 - 낮은 확률의 사건일수록 더 많은 정보를 전달합니다.
자기정보 - 어떤 사건 e에 대한 확률 변수가 x = {e1, e2, ... ek}일 때, 사건이 일어날 확률을 추정할 수 있다면, 그 사건에 대한 정보량을 측정할 수 있습니다. --> 특정사건 ei의 정보량

$$\boldsymbol{h(e_i) = -\log_2 P(e_i)} \quad \textbf{또는} \quad \boldsymbol{h(e_i) = -\log_e P(e_i)}$$

자연로그



x→0 y→∞
x→1 y→0

엔트로피
불확실성을 수치화 하는 방법: (P의 확률분포 * P의 자기정보량)의 총합

$$H(x) = -\sum_{i-1,k} P(e_i)\log_2 P(e_i) \qquad H(x) = -\sum_{i-1,k} P(e_i)\log_e P(e_i)$$

[핵심10] **교차 엔트로피**
서로 다른 확률분포 P, Q의 차이를 수치화 하는 방법: (P의 확률분포 * Q의 자기정보량)의 총합

$$E(P,Q) = -\sum_{x} P(x)\log_2 Q(x) = -\sum_{i=1,k} P(e_i)\log_2 Q(e_i)$$

차이가 없으면 -> 0
차이가 많으면 -> 무한대

기계학습의 최적화
학습데이터에 따라 정해지는 **목적함수**의 **최저점**을 탐색합니다. --> **모든 지점**에서의 **순간변화율**을 알아야 합니다.
목적함수: 미분가능한 함수, 순간변화율: 미분 값
최적화 이론의 알고리즘 - 경사하강 알고리즘
(1) 배치 경사하강 알고리즘 - [0] 샘플의 gradient를 평균하고 [1] 한꺼번에 갱신합니다.
(2) 스토캐스틱 경사하강 알고리즘 - [0] 한 샘플의 gradient를 계산하고 [1] 즉시 갱신합니다.
<-------------- 스토캐스틱 경사하강법, 오차역전파(미분하는 과정)을 이용 ---------->
[0] 난수를 생성하여 초기해 θ 를 구합니다.
[1] repeat

목적함수 $J(\theta)$ 가 작아지는 방향 $d\theta$ 을 구합니다. $\theta = \theta + d\theta$	샘플의 순서를 섞는다. for(i=1 to n)

i번째 샘플에 대한 gradient Δ 계산
$\theta = \theta - \rho\Delta_i$
학습률

[2] until (멈춤조건)
[3] $\hat{\theta} = \theta$

그냥 보세요

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

미분은 **함수의 기울기**를 나타냅니다.
점점 커지는 방향을 나타내기 때문에, **-미분**을 해야 작아지는 방향을 알려주고, **목적함수**의 **최저점**을 찾을 수 있습니다.