Week3. 데이터의 분류

3.1 데이터의 유사도

- 인공지능이 데이터로 수행하는 주요기능 1) 판단(decision), 2) 예측(prediction)
- 주어진 데이터의 특징에 따라 데이터가 어느 범주(class or category)에 속하는지 판단하는 것 = 분류(classification)

→ 데이터를 분류하기 위해서:

① 데이터를 계산할 수 있는 형태로 표현

② 각 범주(기준)와 얼마나 가까운지 or 유사한지 판단

↳ 데이터의 유사도 (similarity)

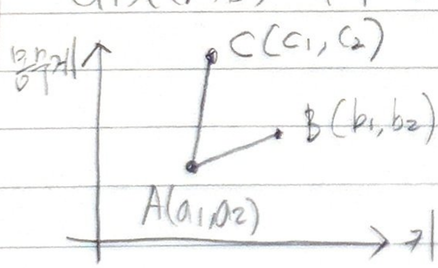
3.2 거리

'유사도'를 측정하는 척도는 매우 다양, 그 중 하나가 '거리재기'

두 점 $A(a_1, a_2)$, $B(b_1, b_2)$ 사이의 거리

$= \text{dist}(A,B) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$

dist(A,B) 작다 → 유사하다, dist(A,B) 크다 → 유사하지 않다



$\text{dist}(A,B) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$

$\text{dist}(A,C) = \sqrt{(a_1-c_1)^2+(a_2-c_2)^2}$

$\text{dist}(A,B) < \text{dist}(A,C)$

3.3 노름 (Norm)

$\mathbb{R}^n$의 벡터 a, b와 스칼라 k에 대해 다음이 성립:

① $\|a\| \geq 0$

$\|a\| \Leftrightarrow a=0$

② $\|ka\| = k\|a\|$

③ $\|a+b\| \leq \|a\| + \|b\|$

벡터 $a=(a_1, a_2)$에 대하여 a의 크기를 다음과 같이 나타내고, a의 norm이라 한다.

$\|a\| = \sqrt{a_1^2+a_2^2}$ (벡터의 크기 계산 : norm)

= 원점에서 점 $A(a_1, a_2)$에 이르는 거리

두 벡터 $a=(a_1, a_2)$, $b=(b_1, b_2)$에 대하여 $\|a-b\|$는 두 점 $A(a_1, a_2)$, $B(b_1, b_2)$ 사이의 거리가 되며, 다음이 성립한다.

$\text{dist}(A,B) = \|a-b\| = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$

$a=(a_1, a_2, a_3)$, $b=(b_1, b_2, b_3)$ >

$\|a\| = \sqrt{a_1^2+a_2^2+a_3^2}$

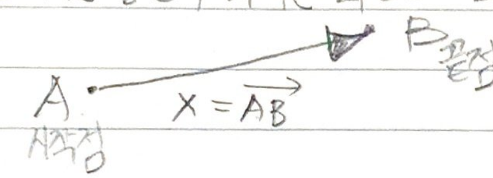
$\text{dist}(A,B) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2}$

Additional Section – vector

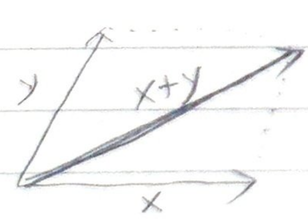
Section 1.1 공학·수학에서의 벡터 : n차원공간

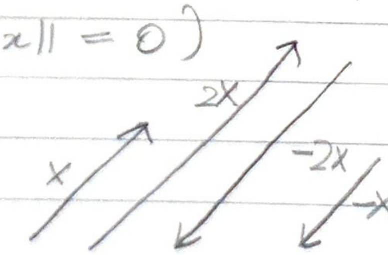
Def
스칼라(scalar) - 길이, 넓이, 질량, 온도
> 크기만 주어지면 완전히 표시되는 양

Def
벡터(vector) – 속도, 위치이동, 힘
> 크기뿐만 아니라 방향까지 지정하지 않으면 불완전하게 표시되는 양
> 벡터는 크기와 방향을 갖는 유향선분
> 2, 3차원 공간의 벡터는 화살표로 표현가능



Def
Vector Sum / Scalar multiplication
벡터 x, y 와 스칼라 k에 대하여, $x+y$ 와 $k \cdot x$를 다음과 같이 정의한다.
(1) $x+y$는 x, y에 의하여 결정되는 평행사변형의 대각선으로 표시되는 벡터
(2) kx는 $k>0$이면 x와 방향이 같으면서 길이는 k배 한 벡터
$k<0$이면 x와 방향이 반대이면서 길이는 $|k|$배 한 벡터
($k=0$이면 $\|kx\| = 0$)





Def
평면벡터 vector in the plane
- 두 실수를 성분(component)으로 하는 벡터 $x = (x_1, x_2)$

Def
공간벡터 vector in space
- 세 실수를 성분으로 하는 벡터 $x = (x_1, x_2, x_3)$

Def **일차결합 (linear combination)**
$v_1, v_2, \cdots v_k$가 R^n의 벡터이고, 계수 $c_1, \cdots c_k$가 실수일때
$$\boxed{x = c_1v_1 + c_2v_2 + \cdots + c_kv_k}$$
인 형태를 $v_1, v_2, \cdots v_k$의 일차결합(linear combination)이라 한다

* n개 실수성분 → n차원벡터 n-dimensional vector

3.4 노름(norm - 크기, 거리)을 활용한 데이터 유사도 비교

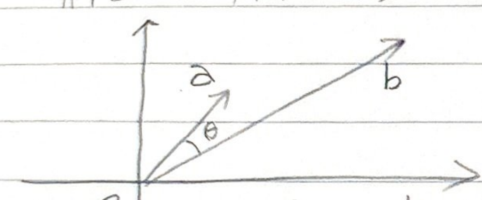
(예제) $A(0, 1, -7, 1)$, $B(5, 2, -1, 3)$, $C(-2, 0, -4, 6)$
이 대하여 B는 A와 C 중 어느 데이터랑 더 가까운가?

$$dist(A,B) = \sqrt{(0-5)^2+(1-2)^2+(-7+1)^2+(1-3)^2} = \sqrt{66}$$
$$dist(B,C) = \sqrt{(5+2)^2+(2-0)^2+(-1+4)^2+(3-6)^2} = \sqrt{71}$$

∴ B는 C보다 A에 더 가깝다.

3.5 사잇각을 활용한 데이터의 비교

'유사도 = 거리 가까움' 아닌 '유사도 = 데이터 패턴(방향)' 이라면?



$\theta = \angle(a, b)$

$\boxed{a=(a_1, a_2) \text{ 와 } b(b_1, b_2)}$

→ 거리는 멀지만 방향은 유사하다

→ 척도에 따라 매우 상반된 평가

3.6 코사인 유사도의 개념

사잇각 θ은 벡터의 내적(inner product)으로부터 정의된다

→ 벡터 내적을 이용, θ의 코사인 값으로 유사도 측정 = 코사인 유사도.

3.7 내적

두 벡터 $a(a_1, a_2)$, $b(b_1, b_2)$의 내적

$$a \cdot b = a_1 b_1 + a_2 b_2$$

두 벡터의 내적은 다음의 성질을 만족한다.

① $a \cdot a = \|a\|^2 \geq 0$, $a \cdot a = 0 \iff a = 0$

② $a \cdot b = b \cdot a$ (교환법칙)

③ $(a+b) \cdot c = a \cdot c + b \cdot c$

④ $(ka) \cdot b = a \cdot (kb) = k(a \cdot b)$

Section 1.2 내적과 직교

for $x = (x_1, x_2, \cdots, x_n)$ in $\mathbb{R}^n$

$$\|x\| = \sqrt{x_1^2 + x_2^2 \cdots + x_n^2}$$

↳ norm, length, magnitude

def
내적 (Euclidean Inner Product, dot product)

$\mathbb{R}^n$의 벡터 $x = (x_1, x_2, \cdots x_n)$, $y = (y_1, y_2, \cdots y_n)$에 대하여
실수 $x_1y_1 + x_2y_2 + \cdots x_ny_n$을 x와 y의 내적이라 하고 $x \cdot y$로 나타낸다.

* 코시-슈바르츠 부등식 (Cauchy-Schwarz inequality)

$\mathbb{R}^n$ 임의의 벡터 x, y에 대하여 다음 부등식이 성립한다.

$$\boxed{|x \cdot y| \leq \|x\| \, \|y\|}$$

(단, 등호는 x, y중 하나가 다른 하나의 실수배일 때만 성립)

def
두 벡터 사이의 각

$\mathbb{R}^n$의 벡터 x, y에 대하여

$$\boxed{x \cdot y = \|x\| \, \|y\| \cos\theta \quad (0 \leq \theta \leq \pi)}$$

인 θ가 x와 y가 이루는 각 (angle, 사잇각)이라 한다.

* 직교와 평행

- $x \cdot y = 0$일 때 x와 y는 서로 직교 (orthogonal)한다.
- 실수 k에 대하여 $x = ky$인 경우 x는 y와 평행하다.

def
단위벡터, 직교벡터, 정규직교 벡터
unit vector, orthogonal vector, orthonormal vector

$\mathbb{R}^n$의 벡터 x에 대하여 노름이 1인 벡터, 즉 $\|x\| = 1$인 벡터 = 단위벡터
$\mathbb{R}^n$의 벡터 x, y가 직교 → x, y는 직교 (orthogonal)
$\mathbb{R}^n$의 벡터 x, y가 직교하면서 각각 단위벡터 → x, y는 정규직교 (orthonormal)

def
벡터에 대한 삼각부등식

$\mathbb{R}^n$의 벡터 x, y에 대하여 $\|x+y\| \leq \|x\| + \|y\|$ (단, 등호는 $y = kx$, $k \geq 0$일 때만)

Additional Section

def
기본단위벡터
Standard unit vector

임의의 벡터 x에 대하여,

$u = \frac{1}{\|x\|} x$

→ 단위벡터

$\mathbb{R}^n$의 단위벡터 중 다음 n개의 벡터

$e_1 = (1, 0, 0, \cdots 0)$
$e_2 = (0, 1, 0, \cdots 0)$
$\vdots$
$e_n = (0, 0, 0, \cdots n)$

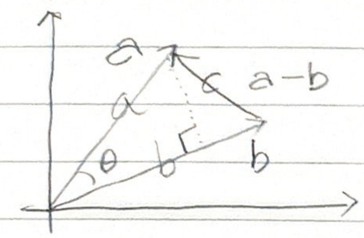
→ $\boxed{\text{기본단위벡터}}$

$x = (x_1, x_2, \cdots x_n) = x_1e_1 + x_2e_2 + \cdots x_ne_n$

3.8 사잇각

벡터의 내적은 두 벡터가 이루는 사잇각과 관련이 있다.

<피타고라스 정리 이용>



$$c^2 = (a\sin\theta)^2 + (b - a\cos\theta)^2$$
$$= a^2 + b^2 - 2ab\cos\theta$$

↓ 벡터를 이용하여 표현

1 $\|a-b\|^2 = \|a\|^2 + \|b\|^2 - 2\|a\|\|b\|\cos\theta$

↓ 내적의 정의와 성질에 의해

$$\|a-b\|^2 = (a-b)\cdot(a-b)$$
$$= a\cdot a - a\cdot b - b\cdot a + b\cdot b$$

2 $= \|a\|^2 + \|b\|^2 - 2(a\cdot b)$

두 식을 비교하면

$a\cdot b = \|a\|\|b\|\cos\theta$

$$\boxed{\cos\theta = \frac{a\cdot b}{\|a\|\|b\|}}$$

$(0 \le \theta \le \pi)$

3.9 코사인 유사도의 계산

cosine similarity

→ 코사인 값이 크면, 사잇각은 작아지고, 유사도는 높아진다.

$$\cos\theta = \frac{a\cdot b}{\|a\|\|b\|} = \boxed{\left(\frac{a}{\|a\|}\right)\cdot\left(\frac{b}{\|b\|}\right)}$$

for $a = (a_1, a_2, \dots$

$b = (b_1, b_2, \dots b_n$

$a, b \in \mathbb{R}^n$,

$0 \le \theta \le \pi$.

"$\frac{a}{\|a\|}$ 와 $\frac{b}{\|b\|}$ 는 크기(norm)가 항상 1인 단위벡터(unit vector)"

→ 코사인 유사도는 크기 / 거리는 무시하고 데이터의 패턴(방향)만 고려

→ 코사인 값 ∝ 코사인 유사도