데이터분석을 위한 선형대수학

데이터를 벡터로 이해하기

김용담 강사

Contents

데이터분석을 위한 선형대수

1. 데이터분석과 선형대수

데이터 분석 예시

Q. 6번 고객과 가장 비슷한 고객을 어떻게 찾을것인가?

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q. 어떤 기준으로 고객의 유사성을 판단할 것인가?

괭장히 일반적인 질문이다.

절대적인 기준이 되는것이 바로 '수학'

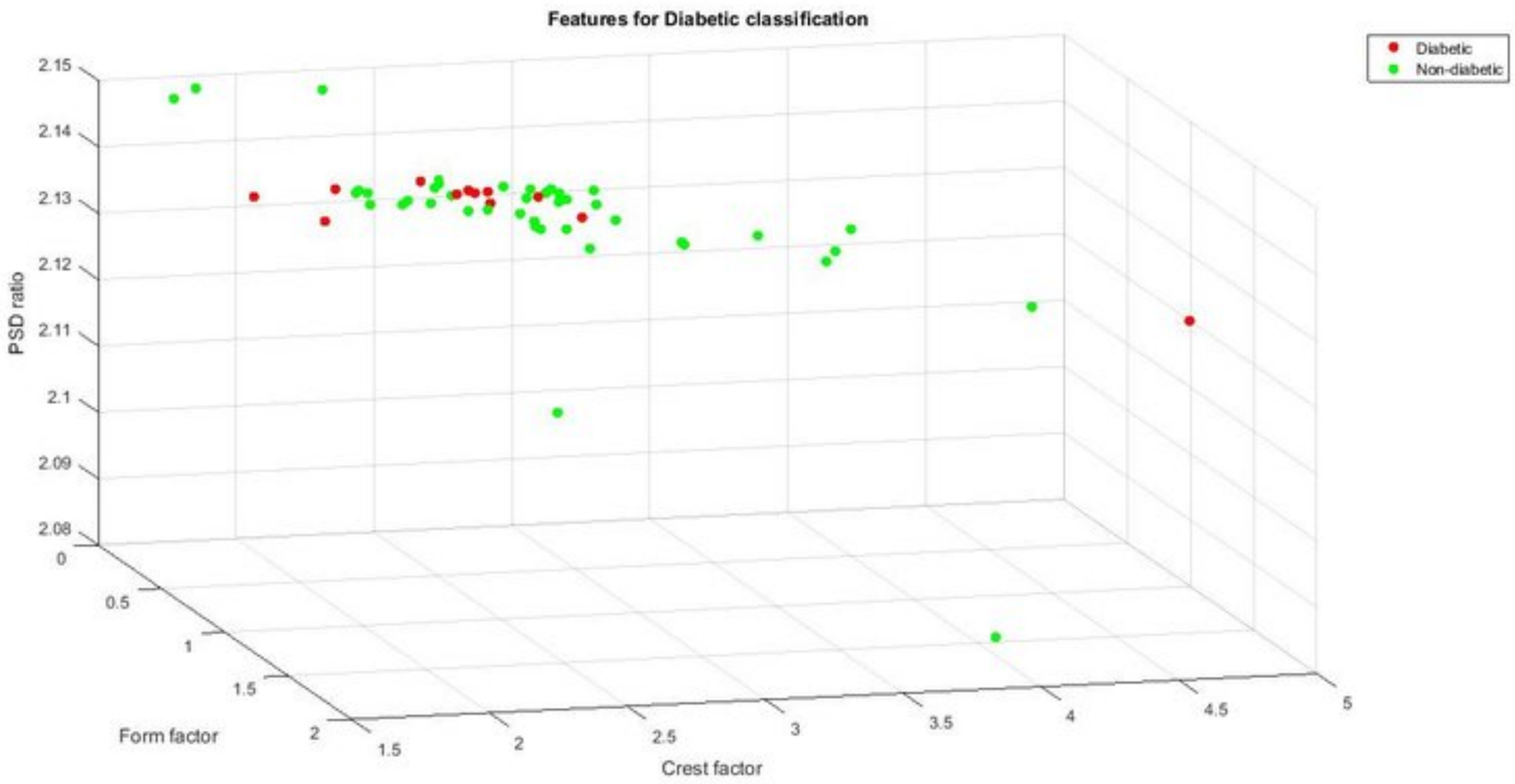
1. 행끼리 비교 (행이 하나의 고객)
2. 거리함수를 기준으로 행끼리 거리를 측정.
3. 행을 점으로 바꿀 수 있으면 두 행 사이의 거리를 측정할 수 있다. (거리 공간에서의 길이)

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터는 벡터다!!!

데이터 분석 예시

A. 데이터는 벡터다



2. 벡터란 무엇인가?

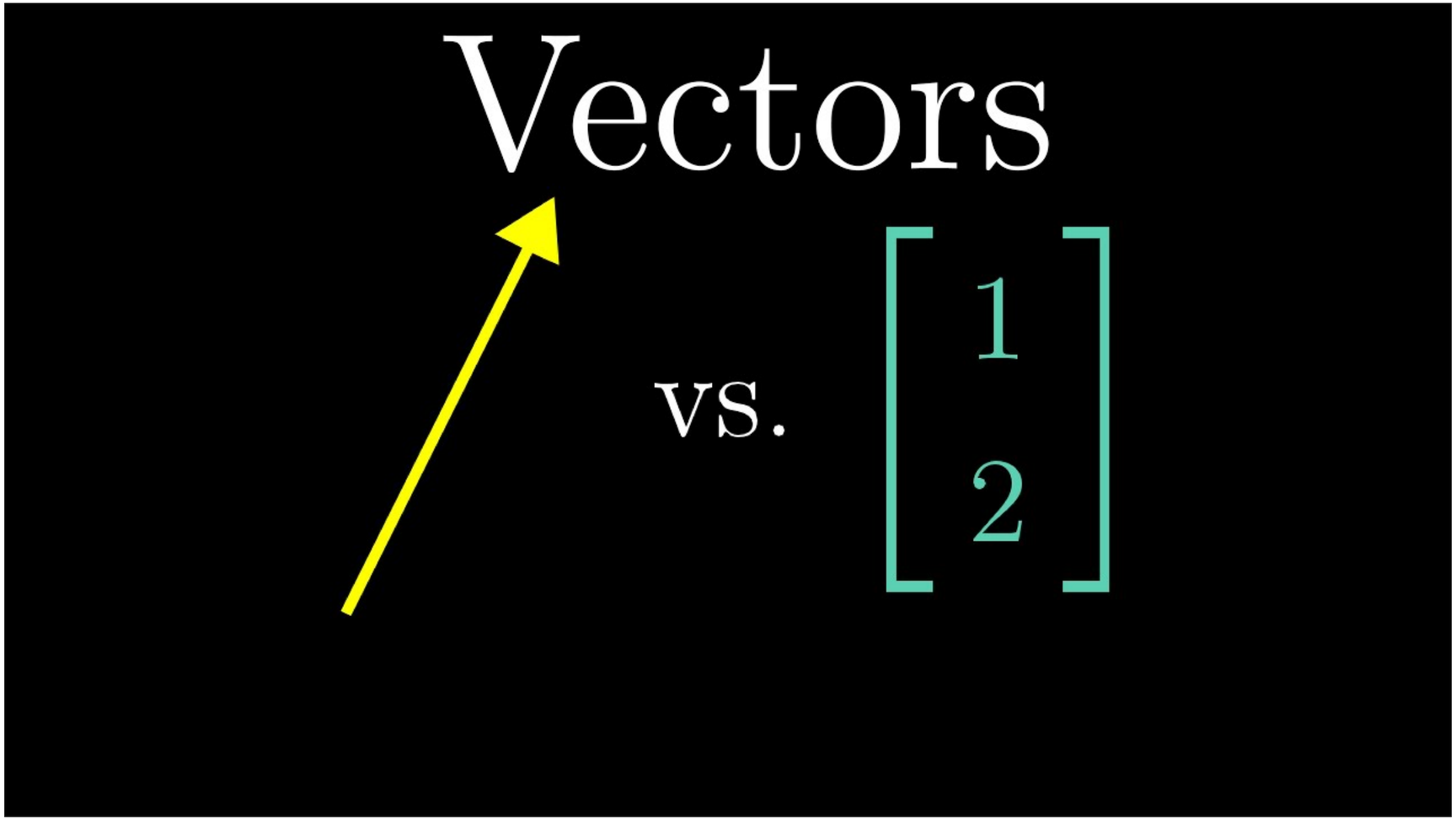
벡터의 정의

벡터 공간의 원소

Definition(약속) 주장하는 논리를 설명하기 위한 약속.

받아들이고 모르는것을 찾아보자

그럼 벡터공간이 뭐야?



벡터의 정의

여러 개의 숫자 모음 (list of numbers)

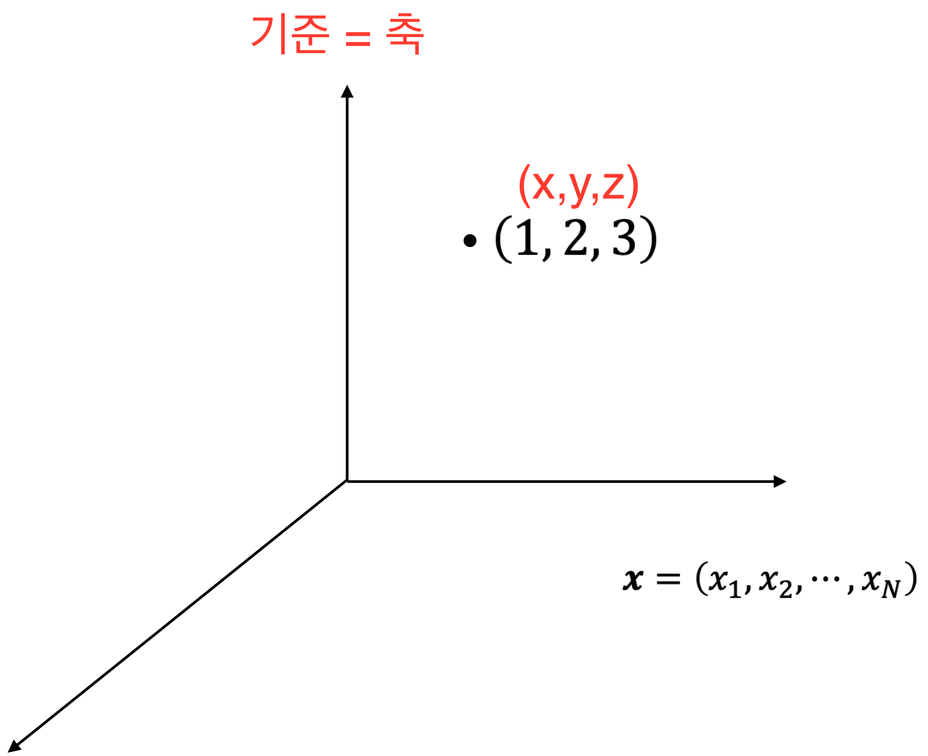
- 행 벡터 $(1, 2, 3)$

- 열 벡터 $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$

- 여러 개의 숫자를 묶어서 표현한 것

문자같은경우 숫자로 변환해줘야 한다. (주소 —> 숫자.)

<Feature engineering>



벡터와 관련된 용어들

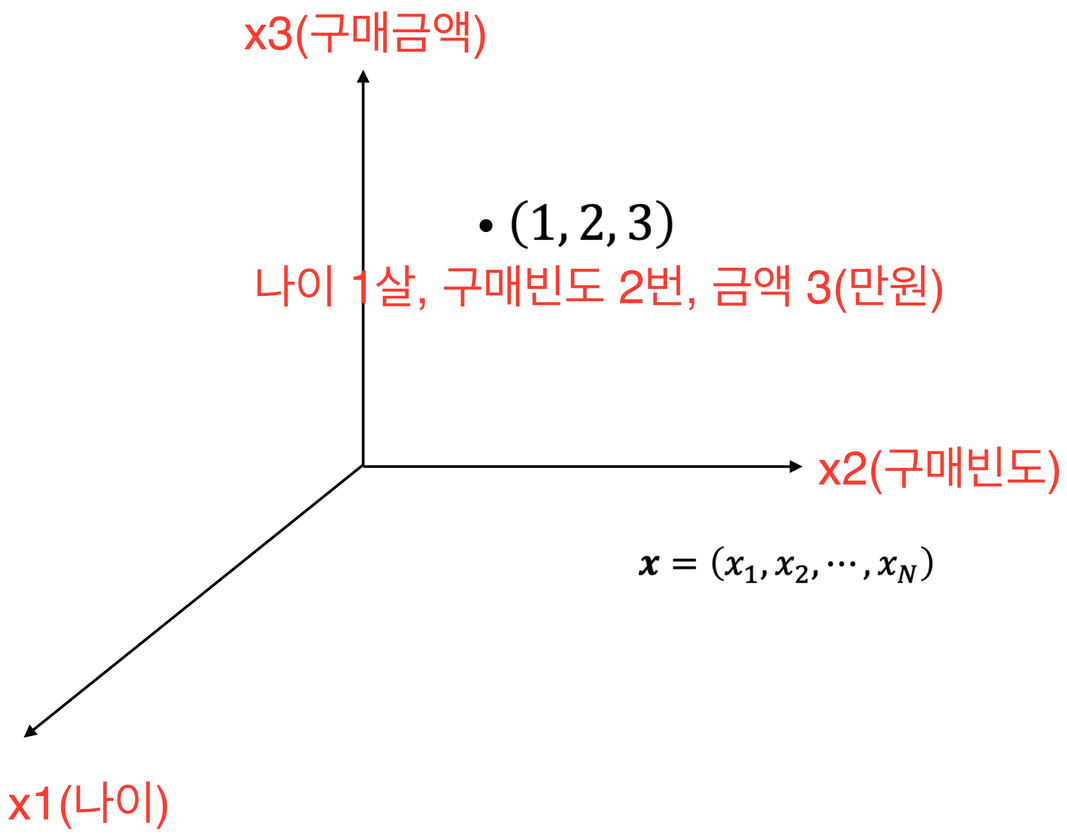
벡터 공간에 대한 이해

- 기저(basis) = 축(axis) x축, y축, z축 ….
- 차원(dimension) = 벡터의 원소 개수
  = 벡터공간의 기저 개수
- 원소(element) = 벡터

Vector Space

여기서는 사람이 나이, 구매빈도, 구매금액으로만 표시된다.

정의된 기저를 바탕으로 정의되는 공간 - vector space



벡터와 관련된 용어들

벡터 공간에 대한 이해

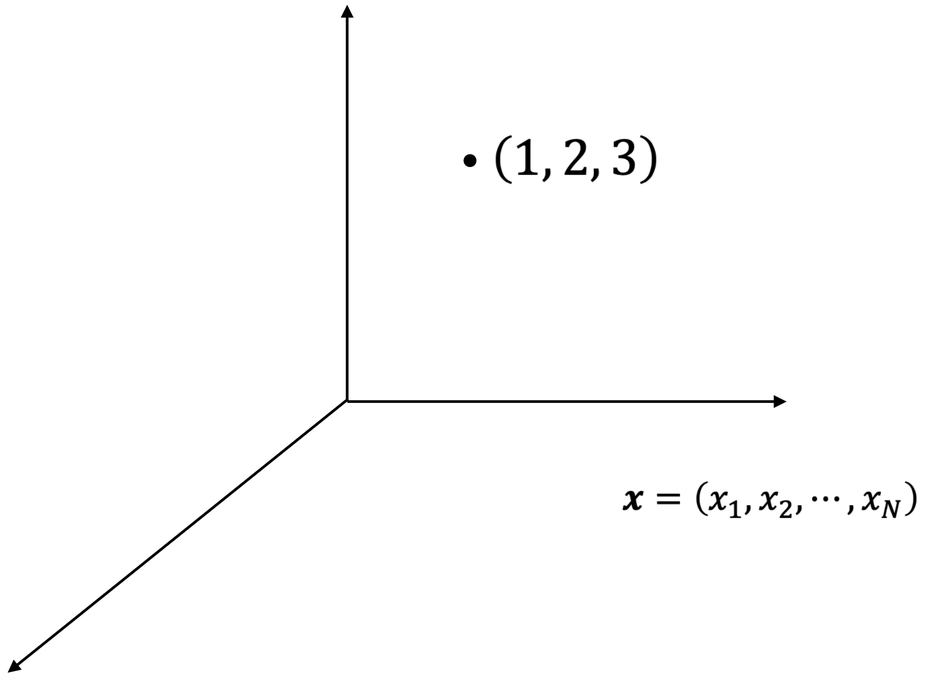
- 벡터 공간(Vector Space) = 집합 모집합?

벡터의 크기

- 크기(norm) = 벡터의 길이

  원점에서부터 얼마나 떨어져 있는가?

- 방향(direction) = 단위 벡터(unit vector)



데이터 분석 예시

column을 기준(축, domain)으로 행의 숫자값이 입력되어 있다.

Q1. 다음 주어진 데이터를 벡터로 표현한다면 기준은 무엇으로 잡을까?

축

컬럼을 기준으로 숫자로 표현 - > 데이터를 벡터로 표현 할 수 잇다.

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터를 벡터로 표현하게되면 그걸 이용해서 할 수 잇는게 많아진다.(데이터간의 거리계산, ...)

데이터 분석 예시

기준을 몇개 잡냐에 따라 정의되는 벡터의 차원이 바뀐다.

Q2. 다음 주어진 데이터는 <u>몇 차원 벡터</u>일까?

=원소 개수

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q3. 다음 데이터의 Product_Info_2 열의 경우엔 공간에 어떻게 표현되는가?

숫자로 바꾸어야 함. 여러 방법이 있음 (onehot- encoding)

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

3. 데이터 분석을 위한 벡터 연산

벡터 기본 연산 같은 차원의 벡터들 사이에서만 가능!

데이터의 특징을 파악할 수 있는 기본 연산

- N차원의 벡터 $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)$와 $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)$에 대해,

정의

- 벡터의 크기 : $|\boldsymbol{x}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2}$ (L2 norm)
- 벡터의 덧셈 : $\boldsymbol{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_N + y_N)$
- 벡터의 뺄셈 : $\boldsymbol{x} - \mathbf{y} = (x_1 - y_1, x_2 - y_2, \ldots, x_N - y_N)$
- 스칼라 배 : $a\boldsymbol{x} = (ax_1, ax_2, \cdots, ax_N)$
- 백터의 내적 : $\boldsymbol{x} \cdot \mathbf{y} = (x_1 \times y_1, x_2 \times y_2, \ldots, x_N \times y_N) = |\boldsymbol{x}||\boldsymbol{y}| \cos\theta$ (단, θ는 $\boldsymbol{x}$와 $\boldsymbol{y}$의 사이각)

데이터 분석 예시

Q1. 6번 고객 벡터의 크기를 계산하여라. (단, 두번째 column 제외)

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q2. 6번 고객의 데이터와 17번 고객 데이터의 차이를 구하여라. (단, 두번째 column 제외)

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q3. 주어진 6, 17번 데이터의 일부를 벡터로 표현할 때, 두 벡터는 같은가?

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q4. 주어진 5, 20번 데이터의 일부를 벡터로 표현할 때, 두 벡터의 내적값을 구하여라.

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

벡터 간 거리

데이터 사이의 유사성 측정

거리함수

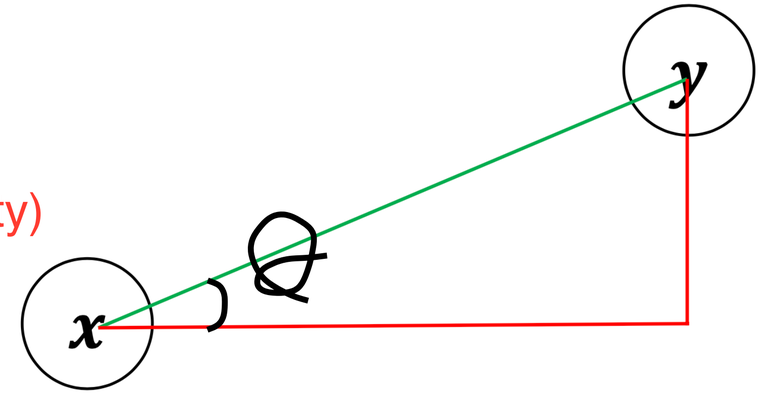
- N차원의 벡터 $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)$와 $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)$에 대해,

- Manhattan Distance (L1 distance) : $\sum_{i=1}^{N} |x_i - y_i|$ 축을 따라감.(축별 차이를 더 중요시하는 경우)
- Euclidean Distance (L2 distance) : $\sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$

- 기하학적 표현

코사인 유사도 (cosine similarity)

작을수록 유사도가 작다.



맨하탄 거리 (1-놈)

유클리디안 거리 (2-놈)

데이터 분석 예시

Q5. 6, 17번 고객 데이터 사이의 L1 distance를 구하여라. (단, 두번째 column 제외)

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q6. 주어진 6, 16번 데이터의 일부를 벡터로 표현할 때, 두 벡터 사이의 L2 distance를 구하여라.

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

데이터 분석 예시

Q7. 주어진 6번 데이터의 일부를 기준으로, 가장 유사하지 않은 고객은 누구인가? (단, L2 distance를 기준으로 한다.)

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

4. Feature Space

Feature Space의 정의

주어진 데이터의 특징을 정의한 벡터 공간

feature engineering

raw data → feature vector

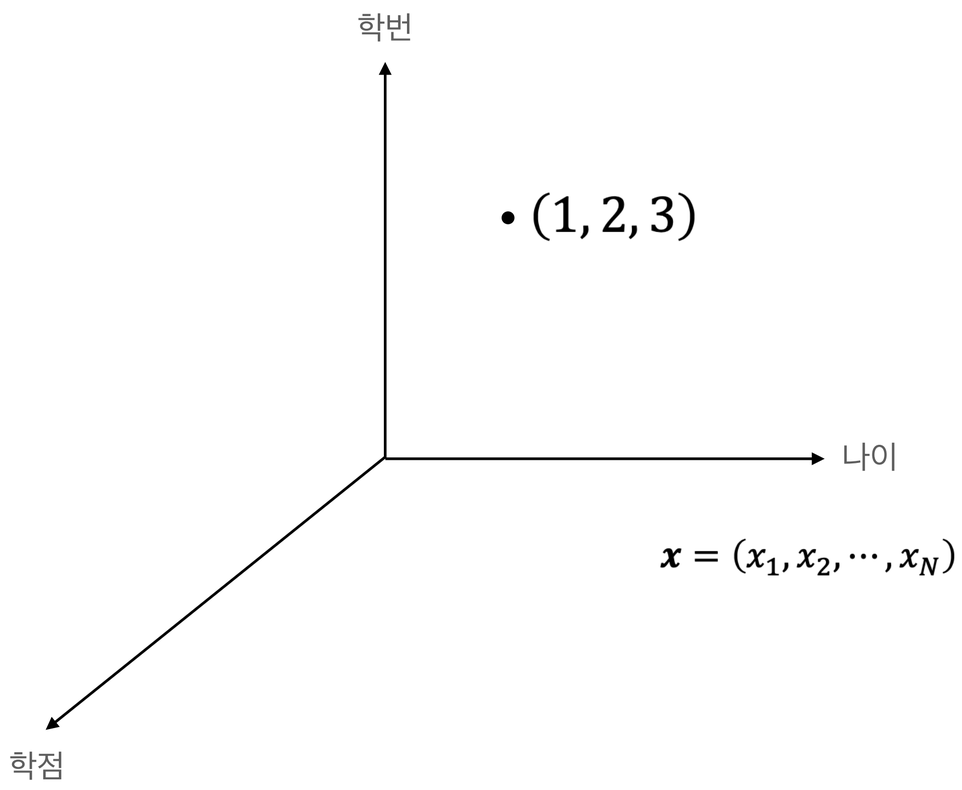
- 주어진 데이터를 input vector라고 할 때, input vector 중에 필요한 특징만을 추출(또는 선별)하여 벡터로 표현한 것을 "feature vector" 라고 한다.

feature selection

- 필요한 특징을 선별하는 방법은 데이터를 잘 이해하고 있는 분석가가 담당한다.

- 필요한 특징을 추출하는 방법은 차원 축소 모델을 사용한다.

feature extraction



$$\boldsymbol{x} = (x_1, x_2, \cdots, x_N)$$

데이터 분석 예시

column의 의미를 기준으로 선별하는 경우

Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_2	Employment_Info_3
2	1	D3	10	0.076923077	2	1	1	0.641791045	0.581818182	0.148535565	0.323007976	0.028	12	1
5	1	A1	26	0.076923077	2	3	1	0.059701493	0.6	0.131799163	0.272287744	0	1	3
6	1	E1	26	0.076923077	2	3	1	0.029850746	0.745454545	0.288702929	0.428780429	0.03	9	1
7	1	D4	10	0.487179487	2	3	1	0.164179104	0.672727273	0.205020921	0.352437744	0.042	9	1
8	1	D2	26	0.230769231	2	3	1	0.417910448	0.654545455	0.234309623	0.424045645	0.027	9	1
10	1	D2	26	0.230769231	3	1	1	0.507462687	0.836363636	0.29916318	0.364886708	0.325	15	1
11	1	A8	10	0.166193846	2	3	1	0.373134328	0.581818182	0.173640167	0.376586717	0.11	1	3
14	1	D2	26	0.076923077	2	3	1	0.611940299	0.781818182	0.40376569	0.571611506	0.12	12	1
15	1	D3	26	0.230769231	2	3	1	0.52238806	0.618181818	0.184100418	0.36264306	0.165	9	1
16	1	E1	21	0.076923077	2	3	1	0.552238806	0.6	0.284518828	0.587795766	0.025	1	3
17	1	D3	26	0.128205128	2	3	1	0.537313433	0.690909091	0.309623431	0.521668453	0.05	9	1
18	1	D4	26	0.230769231	2	3	1	0.298507463	0.690909091	0.271966527	0.455050111	0.09	3	1
19	1	A2	26	0.102564103	2	3	1	0.567164179	0.618181818	0.163179916	0.320783966	0.075	9	1
20	2	D1	26	0.487179487	2	3	1	0.223880597	0.781818182	0.361924686	0.507514769	0.1	9	1
22	1	D4	26	0.487179487	2	3	1	0.328358209	0.636363636	0.142259414	0.264648223	0.16	3	1
23	1	A7	26	0	2	3	1	0.626865672	0.672727273	0.330543933	0.58127899	0.075	9	1
24	2	D4	26	0.487179487	2	3	1	0.208955224	0.745454545	0.246861925	0.360968696	0.1	14	1
25	1	D3	26	0.384615385	2	3	1	0.268656716	0.636363636	0.228033473	0.430949212	0.0378	9	1
26	1	D3	26	0.076923077	2	3	1	0.388059701	0.781818182	0.309623431	0.427393846	0.08	9	1
27	1	D4	26	0.487179487	2	3	1	0.223880597	0.6	0.138075314	0.285253828	0.055	9	1
29	1	D2	26	0.435897436	2	3	1	0.388059701	0.745454545	0.246861925	0.360968696	0.083	9	1

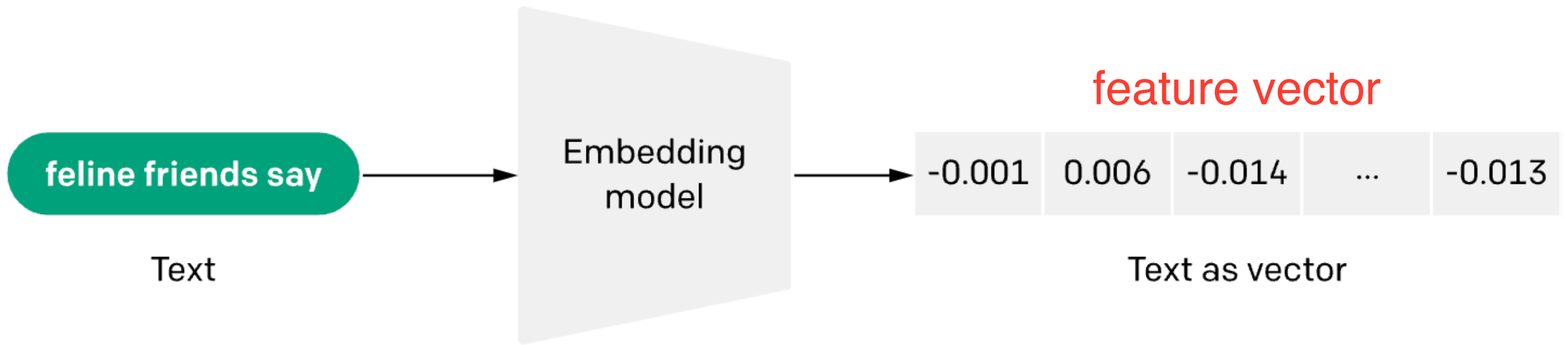
데이터 분석 예시

추출 기법을 통하여 새로운 Feature vector를 생성한 경우 (e.g. PCA)

V1	V2	V3	V4	V5	V6	V7	V8
-1.3598071336738	-0.0727811733098497	2.53634673796914	1.37815522427443	-0.338320769942518	0.462387777762292	0.239598554061257	0.0986979012610507
1.19185711131486	0.26615071205963	0.16648011335321	0.448154078460911	0.0600176492822243	-0.0823608088155687	-0.0788029833323113	0.0851016549148104
-1.35835406159823	-1.34016307473609	1.77320934263119	0.379779593034328	-0.503198133318193	1.80049938079263	0.791460956450422	0.247675786588991
-0.966271711572087	-0.185226008082898	1.79299333957872	-0.863291275036453	-0.0103088796030823	1.24720316752486	0.23760893977178	0.377435874652262
-1.15823309349523	0.877736754848451	1.548717846511	0.403033933955121	-0.407193377311653	0.0959214624684256	0.592940745385545	-0.270532677192282
-0.425965884412454	0.960523044882985	1.14110934232219	-0.168252079760302	0.42098688077219	-0.0297275516639742	0.476200948720027	0.260314333074874
1.22965763450793	0.141003507049326	0.0453707735899449	1.20261273673594	0.191880988597645	0.272708122899098	-0.00515900288250983	0.0812129398830894
-0.644269442348146	1.41796354547385	1.0743803763556	-0.492199018495015	0.948934094764157	0.428118462833089	1.12063135838353	-3.80786423873589
-0.89428608220282	0.286157196276544	-0.113192212729871	-0.271526130088604	2.6695986595986	3.72181806112751	0.370145127676916	0.851084443200905
-0.33826175242575	1.11959337641566	1.04436655157316	-0.222187276738296	0.49936080649727	-0.24676110061991	0.651583206489972	0.0695385865186387
1.44904378114715	-1.17633882535966	0.913859832832795	-1.37566665499943	-1.97138316545323	-0.62915213889734	-1.4232356010359	0.0484558879088564
0.38497821518095	0.616109459176472	-0.874299702595052	-0.0940186259679115	2.92458437838817	3.31702716826156	0.470454671805879	0.53824722837695
1.249998742053	-1.22163680921816	0.383930151282291	-1.23489868766892	-1.48541947377961	-0.753230164566149	-0.689404975426345	-0.227487227519552
1.0693735878819	0.287722129331455	0.828612726634281	2.71252042961718	-0.178398016248009	0.337543730282968	-0.0967168617395962	0.115981735546597
-2.7918547659339	-0.327770756658658	1.64175016056605	1.76747274389883	-0.136588446465306	0.80759646826532	-0.422911389711497	-1.90710747624096
-0.752417042956605	0.345485415344747	2.05732291276727	-1.46864329840046	-1.1583936804082	-0.0778498291166733	-0.608581418236123	0.00360348436201849
1.10321543528383	-0.0402962145973447	1.2673320885949	1.28909146962552	-0.735997163604068	0.288069162976262	-0.586056786337461	0.189379713679593

데이터 분석 예시

학습을 통하여 새로운 Feature vector를 생성한 경우 (e.g. embedding)



Questions?