

Dimensionality Reduction

12 dimension	name
	ticket
(+) family size	(8)

100 - 200 feature

→ feature extraction

→ " selection

$$f = S + P + 1 \quad \text{max}$$

$$a + b + c + d + e = x$$

feature engineering

$$\Rightarrow \sqrt{ab} + 0 + \sqrt{d} + \sqrt{e} = x$$

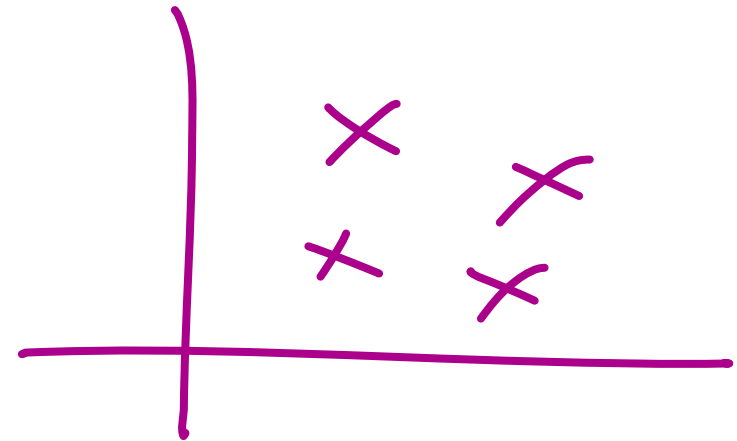
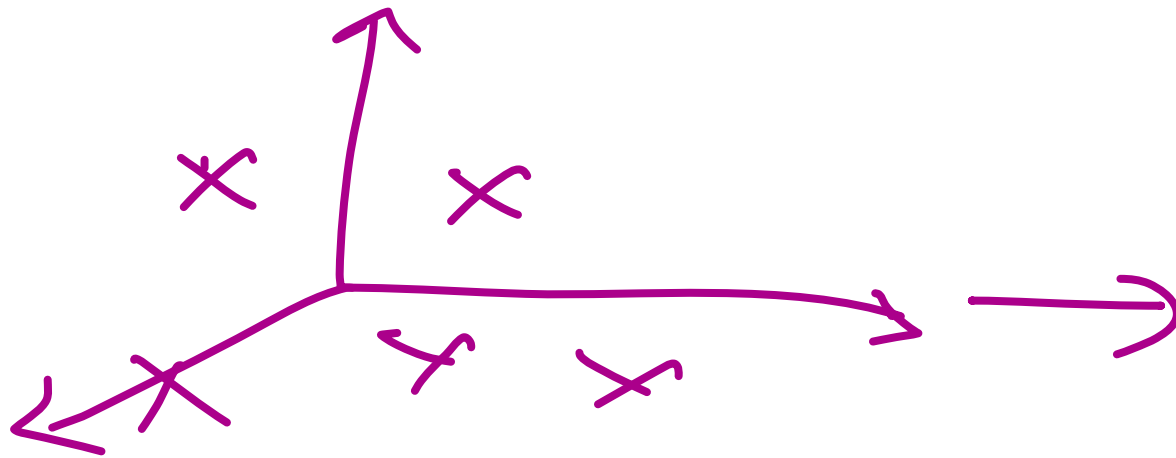
$$\therefore ab + d + e = x$$

feature selection

Feature Selection

- Filter Methods
- Wrapper
 - Forward
 - Backward
- Embedded
 - RFE

Feature Engineering



3D

100 → 6




/ x x

FE/

- PCA → Principal component analysis

- LDA → Linear discriminant

- t-SNE → t-distributed stochastic neighbour embedding

- ICA \rightarrow Independent component
- LLE \rightarrow Locally linear embedding
- Autoencoder
 Neural Network

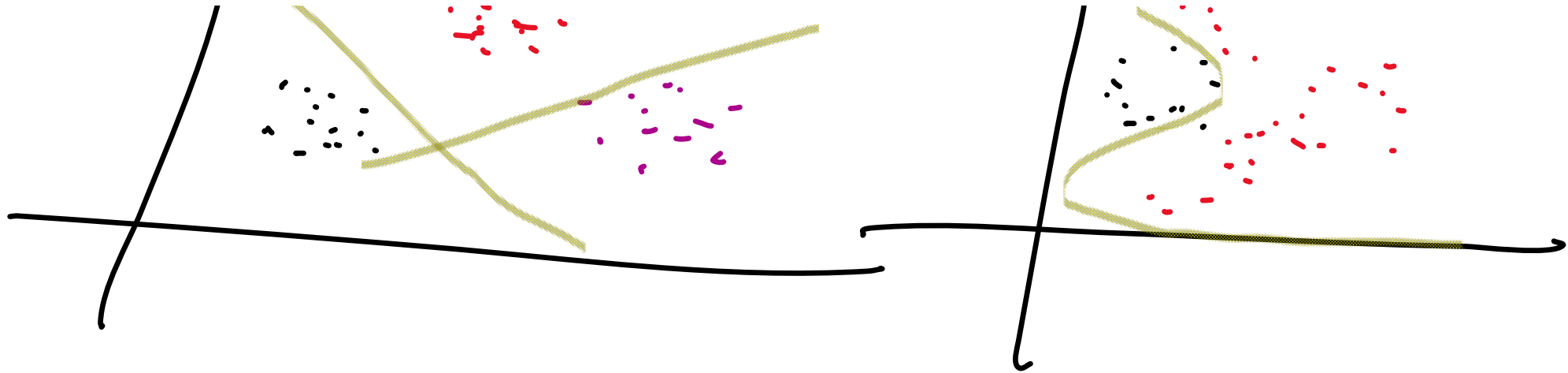
t-SNE

feature	target
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0
27	0
28	0
29	0
30	0
31	0
32	0
33	0
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	0
48	0
49	0
50	0
51	0
52	0
53	0
54	0
55	0
56	0
57	0
58	0
59	0
60	0
61	0
62	0
63	0
64	0
65	0
66	0
67	0
68	0
69	0
70	0
71	0
72	0
73	0
74	0
75	0
76	0
77	0
78	0
79	0
80	0
81	0
82	0
83	0
84	0
85	0
86	0
87	0
88	0
89	0
90	0
91	0
92	0
93	0
94	0
95	0
96	0
97	0
98	0
99	0

— unsupervised non-linear

DR technique





t-SNE

PCA

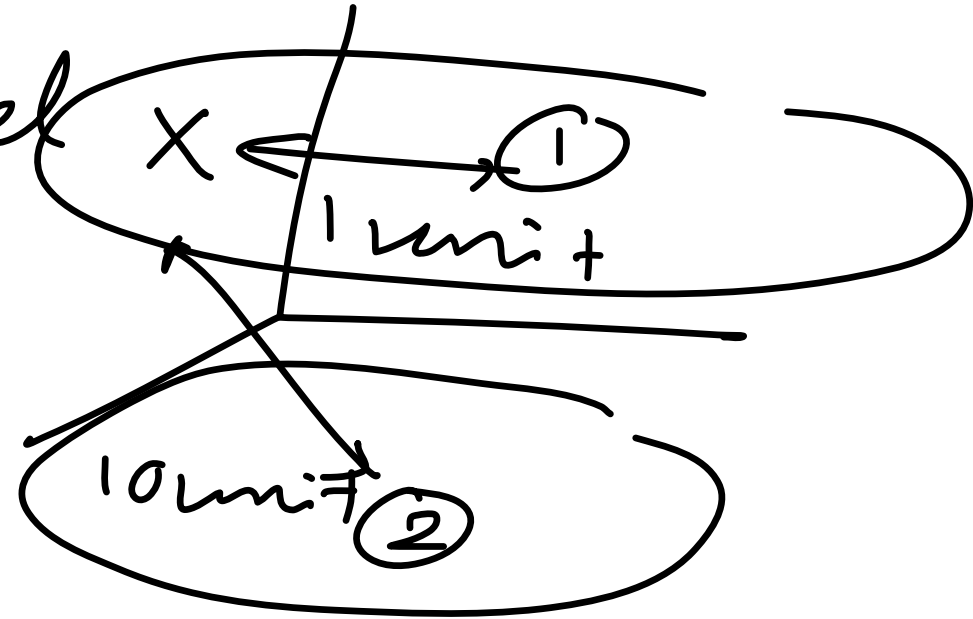
- non-linear

- linear

- relationship
betⁿ data points

- variance

- ① gaussian kernel
- ② Probability
of relation betⁿ data points
- ③ divergence betⁿ probability
distribution \rightarrow minimize



0-1 → Scoring

Standardization

ID		W	H	T
1	0.02	60	120	0.2
2	0.01	58		0.3
3	0.13	90		0.07
		40	2X	

Applications of t-SNE

- Clustering \rightarrow Classify
- Anomaly \rightarrow outliers
- NLP, Computer Security
- Cancer Research