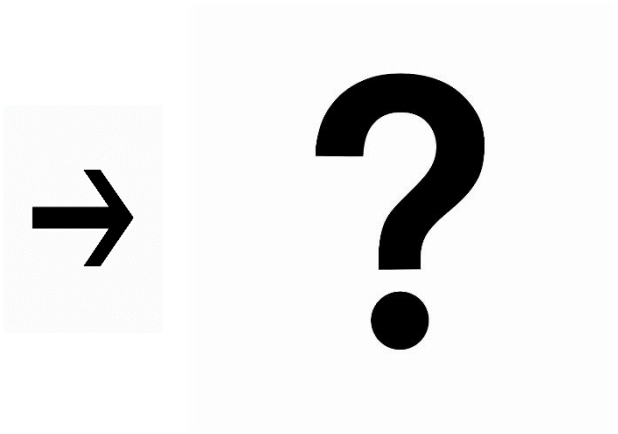
WGS 데이터로부터.

location	REF	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	A	A	T	A	T	T	A
2	T	G	G	T	G	T	T
3	C	C	C	C	C	C	G
4	G	G	G	G	G	G	G
5	A	C	A	C	A	A	A
6	A	C	C	A	A	A	A



컴퓨터가 데이터를 통해 **스스로 패턴을 학습**하고,
새로운 데이터에 대해 **예측하거나 분류**하는 알고리즘을 만들 수 있을까

변이정보로부터

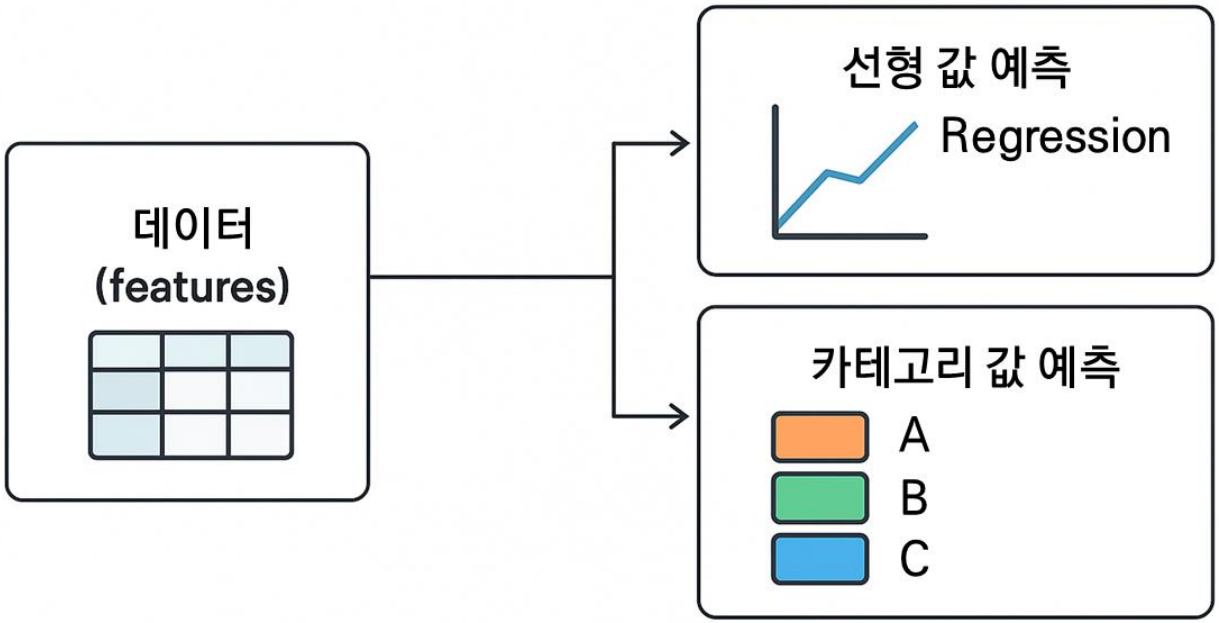
location	REF	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	A	A	T	A	T	T	A
2	T	G	G	T	G	T	T
3	C	C	C	C	C	C	G
4	G	G	G	G	G	G	G
5	A	C	A	C	A	A	A
6	A	C	C	A	A	A	A

?

Supervised vs Unsupervised

Supervised Learning (지도학습)

입력 X와 레이블 Y(정답)가 쌍(pair)으로 주어질 때, 학습을 통해 새로운 입력에 대한 **NEW Y**(예측값)를 생성



기계학습의 기본 틀(Supervised)

1. 데이터 준비

학습데이터

정답
레이블

2. 전처리

전처리 된
학습데이터

정답
레이블

3. 학습모델 구축

전처리 된
학습데이터 → 정답
레이블
(y)

$$y = f(x)$$

4. 모델 평가

학습에 사용하지
않은 데이터 → 예측
레이블
(y')

$$y' = f(x')$$
$$|\tilde{y} - y'|$$

1. 데이터 준비

1. 데이터 준비(학습데이터)

location	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
1	0.815	0.278	0.957	0.792	0.679	0.706	0.695	0.766
2	0.906	0.547	0.485	0.959	0.758	0.032	0.317	0.795
3	0.127	0.958	0.800	0.656	0.743	0.277	0.950	0.187
4	0.913	0.965	0.142	0.036	0.392	0.046	0.034	0.490
5	0.632	0.158	0.422	0.849	0.655	0.097	0.439	0.446
6	0.098	0.971	0.916	0.934	0.171	0.823	0.382	0.646

Y1	A	B	C	A	A	C	C	B
Y2	3.4	2.2	5.5	6.6	8.1	9.9	7.7	6.6

2. 데이터 준비

1. 데이터 준비

학습데이터

location	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
1	0.815	0.278	0.957	0.792	0.679	0.706	0.695	0.766
2	0.906	0.547	0.485	0.959	0.758	0.032	0.317	0.795
3	0.127	0.958	0.800	0.656	0.743	0.277	0.950	0.187
4	0.913	0.965	0.142	0.036	0.392	0.046	0.034	0.490
5	0.632	0.158	0.422	0.849	0.655	0.097	0.439	0.446
6	0.098	0.971	0.916	0.934	0.171	0.823	0.382	0.646

Y1	A	B	C	A	A	C	C	B
Y2	3.4	2.2	5.5	6.6	8.1	9.9	7.7	6.6

검증데이터

location	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13
1	0.709	0.119	0.751	0.547	0.814
2	0.755	0.498	0.255	0.139	0.244
3	0.276	0.960	0.506	0.149	0.929
4	0.680	0.340	0.699	0.258	0.350
5	0.655	0.585	0.891	0.841	0.197
6	0.163	0.224	0.959	0.254	0.251

Y1	B	A	C	B	B
Y2	2.4	4.2	5.1	7.1	4.2

2. 전처리는 어떻게?

데이터가 지저분해요

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.616	0.917	NaN	0.569	0.311	0.689	NaN
0.473	NaN	NaN	0.469	0.529	0.748	0.826
0.352	0.757	0.531	NaN	NaN	0.451	0.538
0.831	0.754	0.779	0.337	0.602	NaN	0.996
0.585	0.380	0.934	NaN	NaN	NaN	NaN
0.550	0.568	NaN	0.794	0.654	0.913	0.443

2. 전처리는 어떻게?

데이터가 지저분해요

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.616	0.917	NaN	0.569	0.311	0.689	NaN
0.473	NaN	NaN	0.469	0.529	0.748	0.826
0.352	0.757	0.531	NaN	NaN	0.451	0.538
0.831	0.754	0.779	0.337	0.602	NaN	0.996
0.585	0.380	0.934	NaN	NaN	NaN	NaN
0.550	0.568	NaN	0.794	0.654	0.913	0.443

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.616	0.917	0.620	0.569	0.311	0.689	NaN
0.473	0.609	0.609	0.469	0.529	0.748	0.826
0.352	0.757	0.531	0.525	0.525	0.451	0.538
0.831	0.754	0.779	0.337	0.602	NaN	0.996
삭제						
0.550	0.568	0.653	0.794	0.654	0.913	0.443

다른 샘플들의 평균 등으로 채우거나,해당 Feature 를 삭제

2. 전처리는 어떻게?

샘플 마다 실험 환경이 달랐어요

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
2	…	4	…	12	…	…
2	…	4	…	12	…	…
4	…	8	…	16	…	…
4	…	8	…	16	…	…
5	…	10	…	20	…	…
1	…	2	…	4	…	…

2. 전처리는 어떻게?

샘플 마다 실험 환경이 달랐어요

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
2	…	4	…	12	…	…
2	…	4	…	12	…	…
4	…	8	…	16	…	…
4	…	8	…	16	…	…
5	…	10	…	20	…	…
1	…	2	…	4	…	…

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
2	…	2	…	2	…	…
2	…	2	…	2	…	…
4	…	4	…	4	…	…
4	…	4	…	4	…	…
5	…	5	…	5	…	…
1	…	1	…	1	…	…

실험이 잘못된 게 아니라면 정규화 작업을 진행!!
예) Z-score 변환

$$z_i = \frac{x_i - \mu}{\sigma}$$

2. 전처리는 어떻게?

데이터가 너무 많아요

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.107	0.084	0.182	0.550	0.402	0.417	0.338
0.962	0.400	0.264	0.145	0.076	0.050	0.900
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.817	0.431	0.869	0.351	0.184	0.491	0.780
0.869	0.911	0.580	0.513	0.240	0.489	0.390

2. 전처리는 어떻게?

데이터가 너무 많아요

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.107	0.084	0.182	0.550	0.402	0.417	0.338
0.962	0.400	0.264	0.145	0.076	0.050	0.900
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.817	0.431	0.869	0.351	0.184	0.491	0.780
0.869	0.911	0.580	0.513	0.240	0.489	0.390

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.107	0.084	0.182	0.550	0.402	0.417	0.338
0.962	0.400	0.264	0.145	0.076	0.050	0.900
0.869	0.911	0.580	0.513	0.240	0.489	0.390

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.762	0.598	0.298	0.279	0.628	0.165	0.932
0.176	0.766	0.920	0.555	0.612	0.768	0.126
0.996	0.597	0.625	0.521	0.352	0.186	0.544

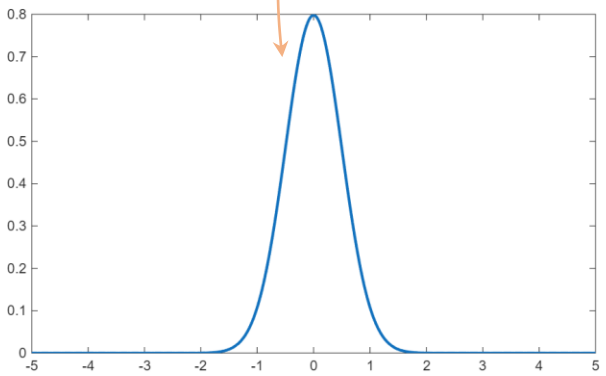
의미 있는 Feature 들만 선택하거나,
차원 축소의 방법을 사용!

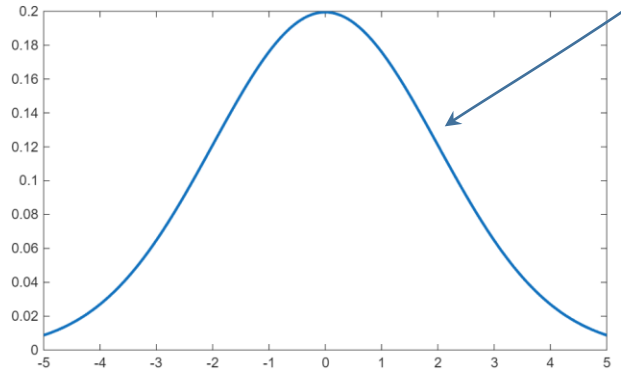
2. 전처리는 어떻게?

의미 있는 Feature 데이터란?

① 분포가 고른 데이터

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
0.107	0.084	...			0.417	-0.338
0.962	2.400	...			0.050	-3.900



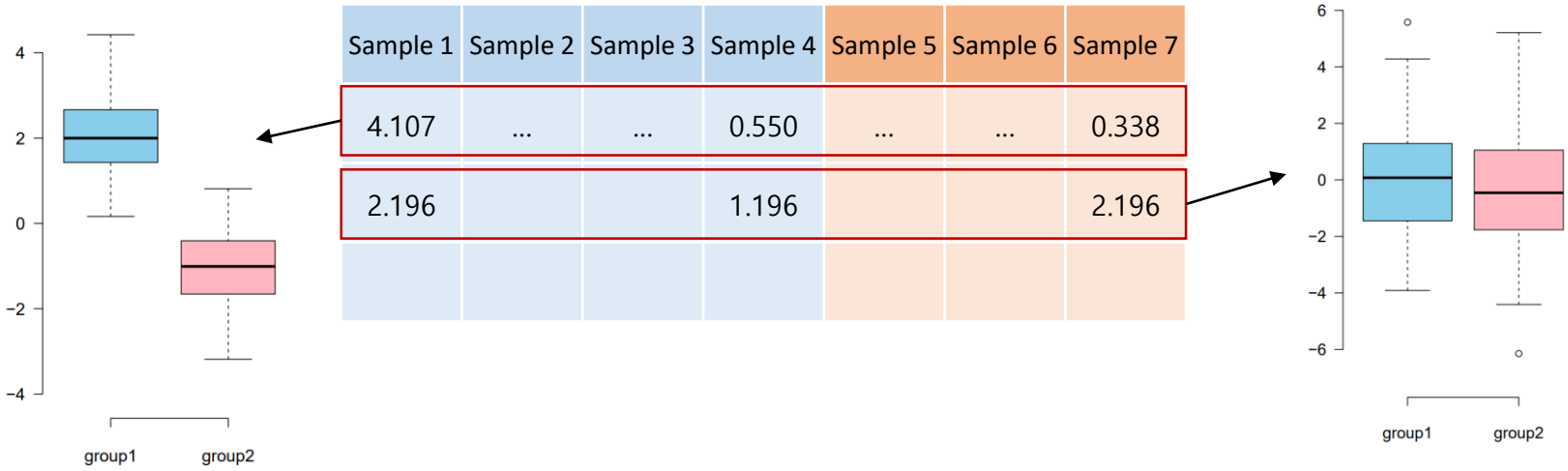


모든 샘플에서 비슷한 값이면 Filtering:
예) 표준편차가 작은 Feature 는 학습에서 제외

2. 전처리는 어떻게?

의미 있는 Feature 데이터란?

② 목적에 맞게 데이터값의 분포가 다른 Feature

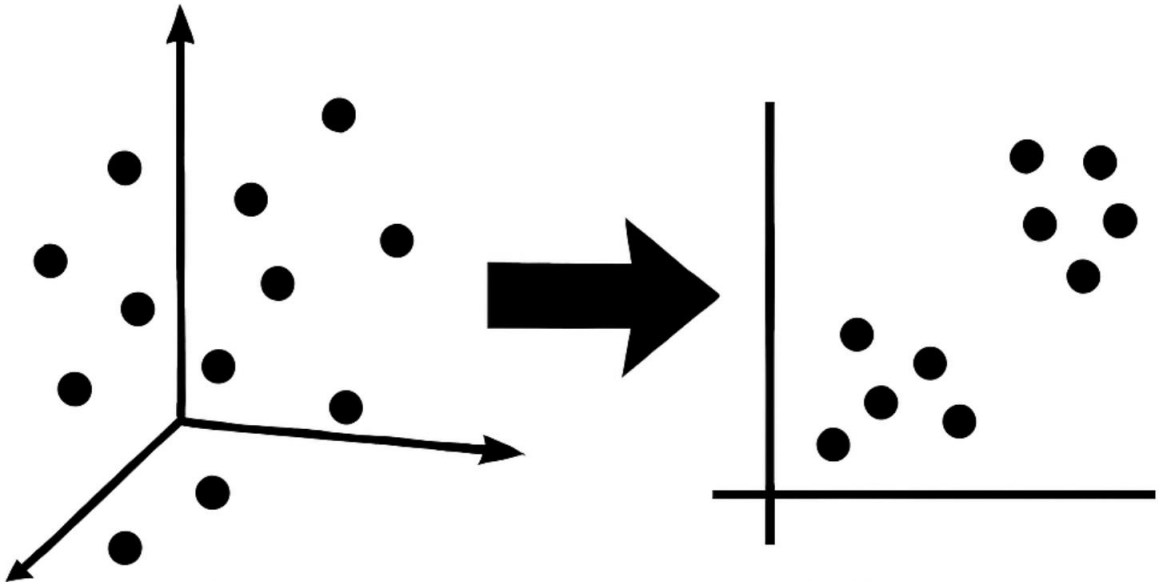


차이가 확연한 Feature 를 선택
예) T-test 유의성 검정 등 (Y label 을 선택하는데 효력이 없지 않을까?)

2. 전처리는 어떻게?

차원 축소 데이터란?

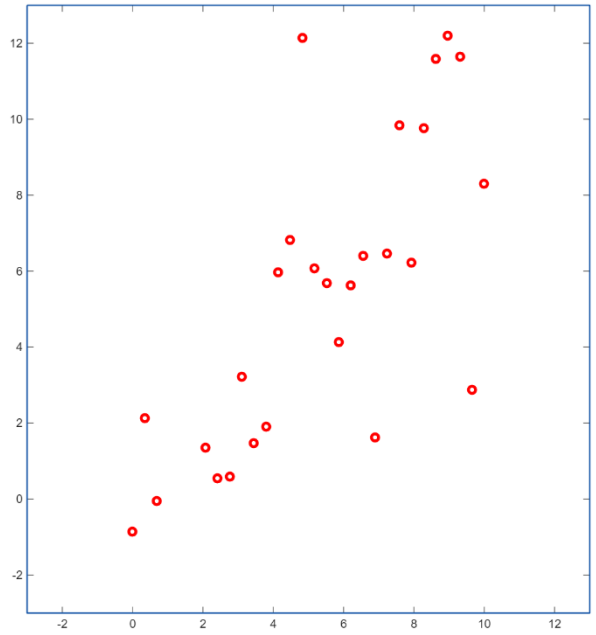
N 차원 데이터를 N 보다 작은 데이터로 표현(정보는 최대한 간직하면서!)

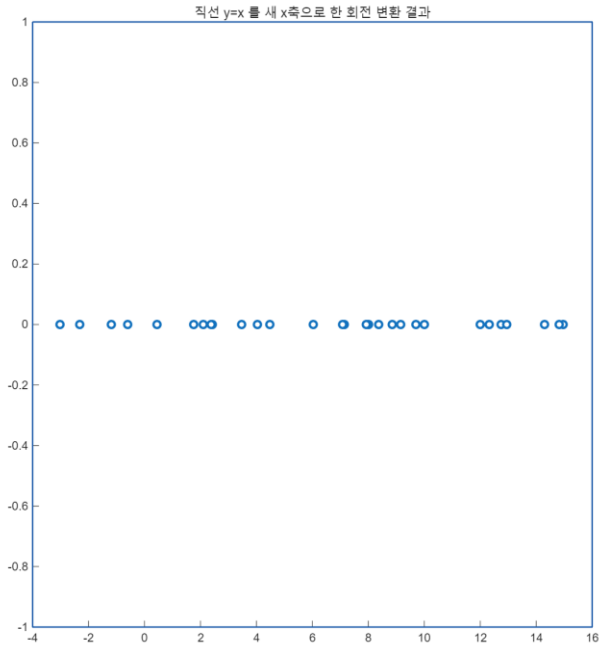


2. 전처리는 어떻게?

차원 축소 데이터란?

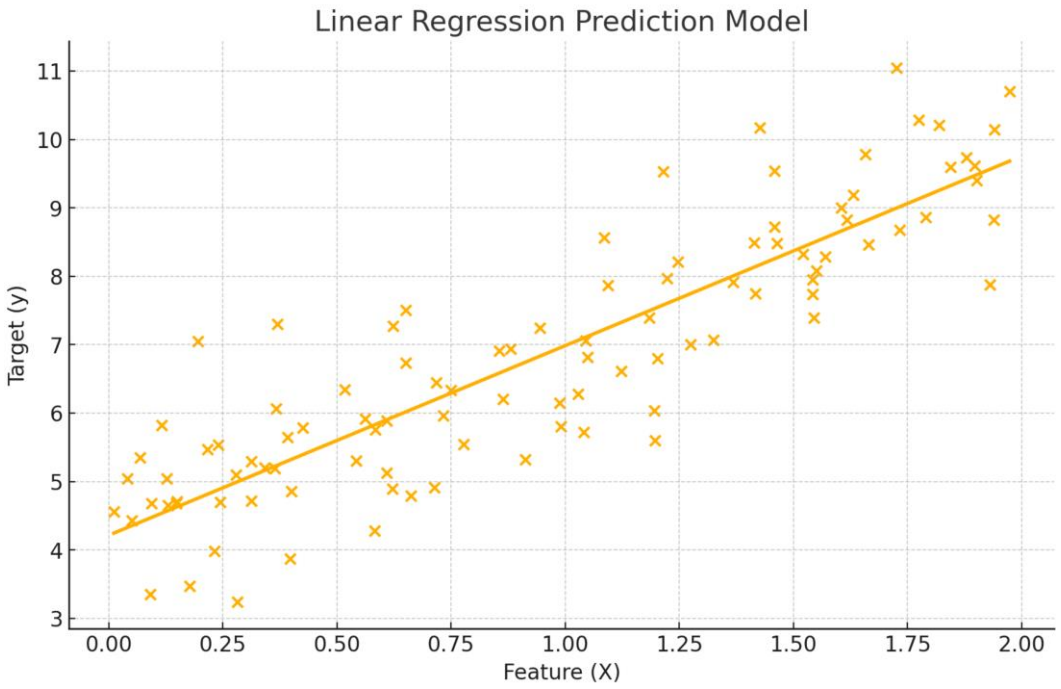
Principle component analysis (PCA)





3. 학습 모델이란

선형 모델 (Linear regression)

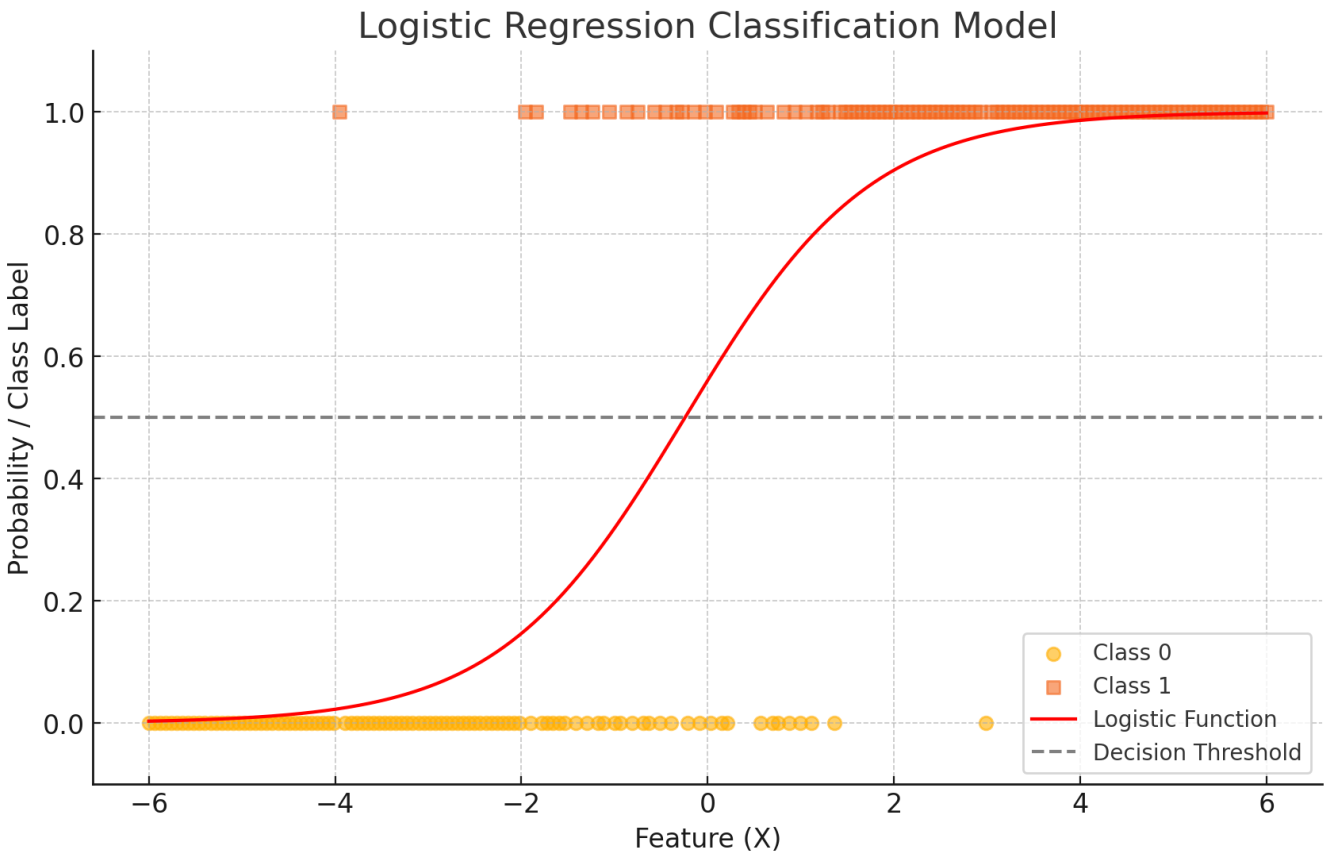


$$Y' \sim a'x + b'$$

선형모델에서 최소화해야할 값: $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

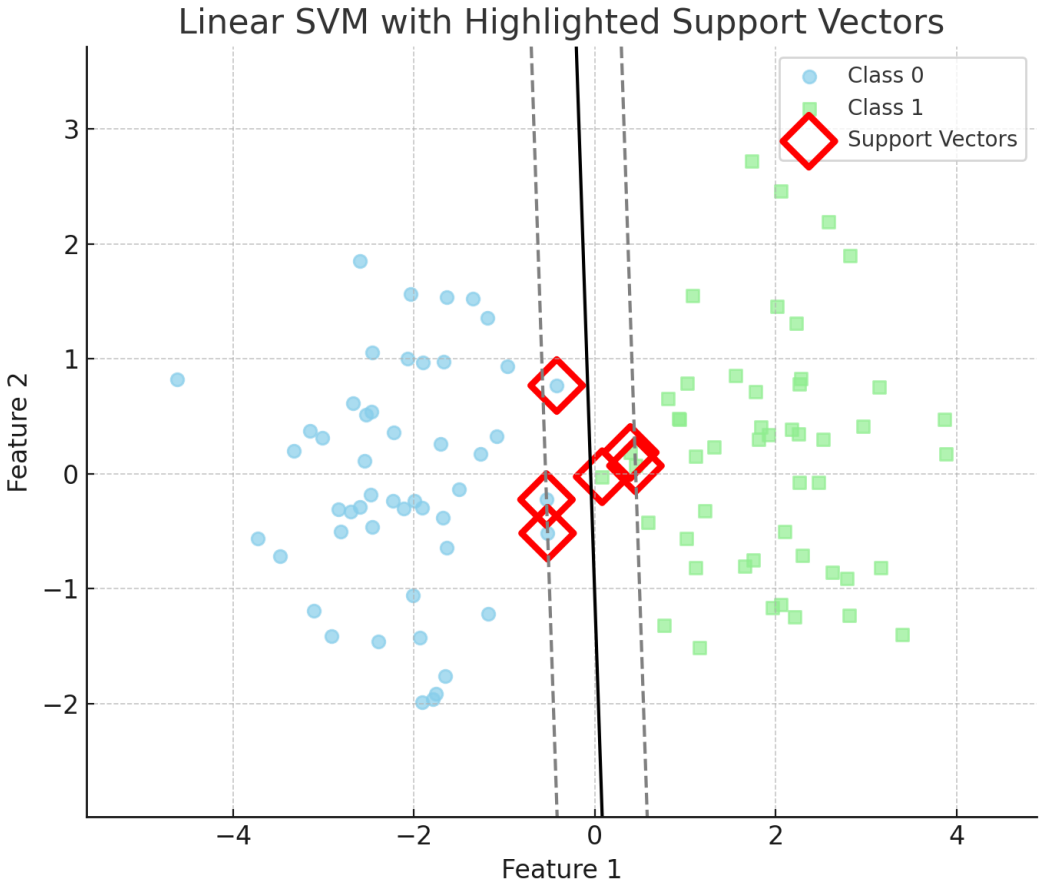
3. 학습 모델이란

분류 모델 (Classification)



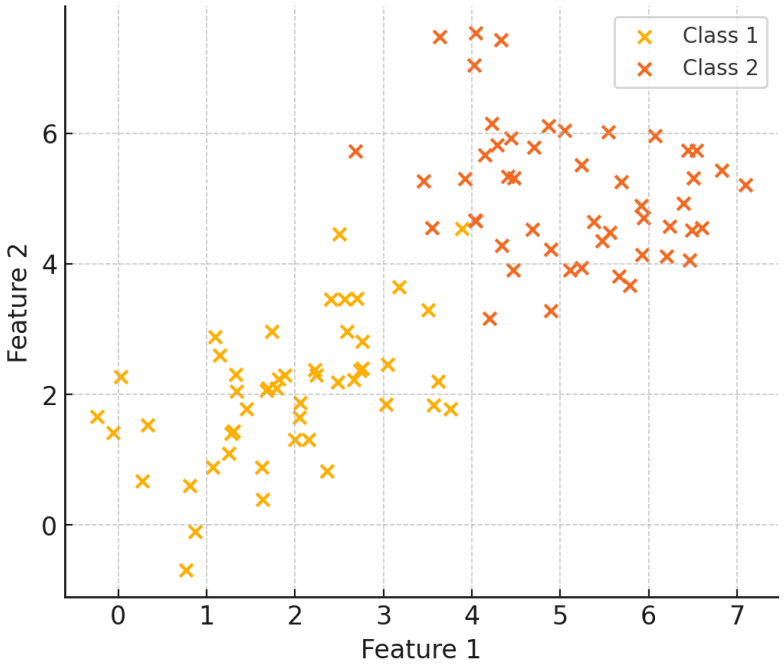
3. 학습 모델이란

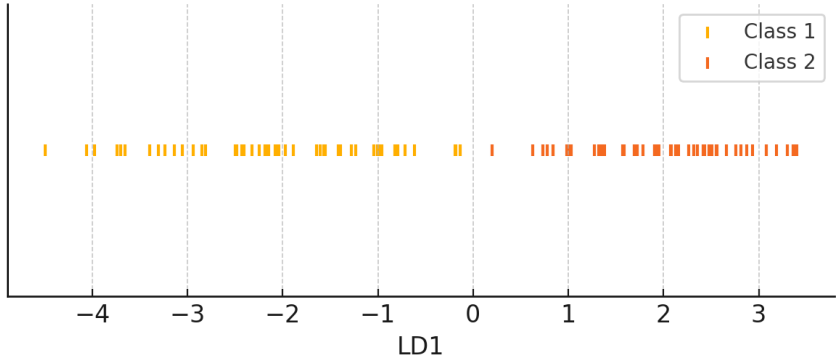
분류 모델 (SVM Classification)



3. 학습 모델이란

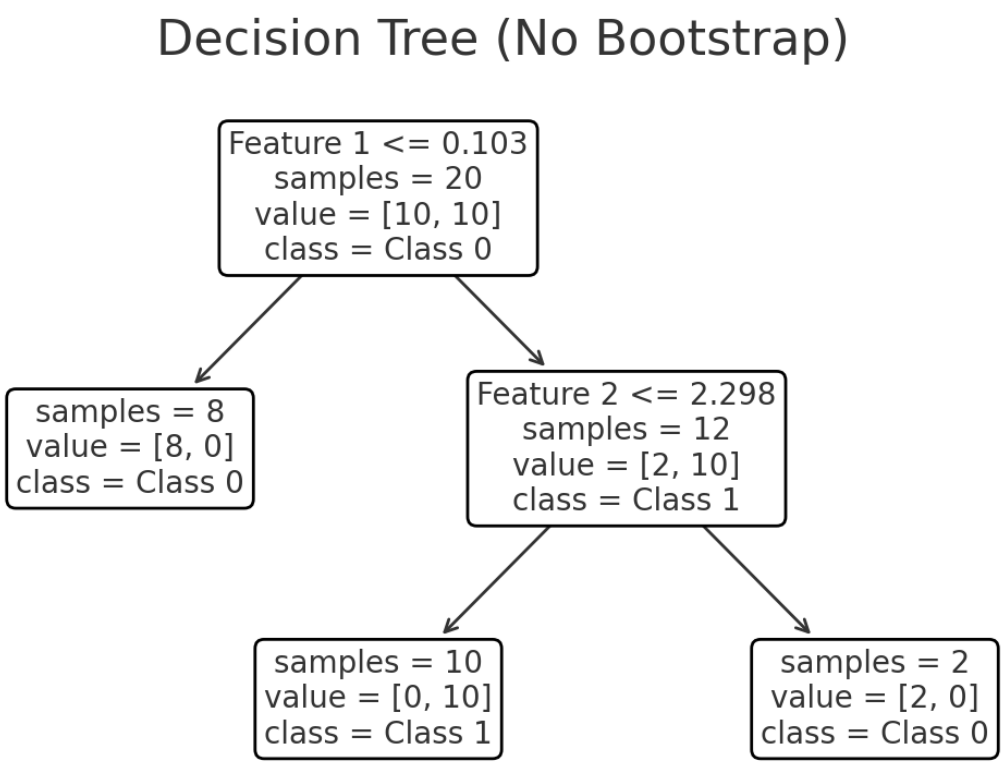
분류 모델 (Linear Discriminant Analysis)





3. 학습 모델이란

분류 모델 (Tree Classification)



4. 모델평가는 어떻게

Y	예측	Y-예측
0	1.9016	-1.9016
0.3448	2.3487	-2.0039
0.6897	1.7574	-1.0678
1.0345	-0.0143	1.0488
1.3793	-1.0298	2.4091
1.7241	5.9303	-4.2062
2.069	6.1714	-4.1024
2.4138	3.9971	-1.5833
2.7586	3.4109	-0.6523
3.1034	-4.6207	7.7241
3.4483	2.0249	1.4233
3.7931	7.3159	-3.5228

		실제				
		AFR	EUR	EAS	SAS	AMR
예측	AFR	178	0	0	0	1
	EUR	0	127	0	2	22
	EAS	0	0	117	0	0
	SAS	0	0	0	118	0
	AMR	1	0	0	0	75

선형 모델은 RMSE, 분류 모델은 Accuracy

SNV 를 수치 값으로

SNV

location	REF	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	A	A	T	A	T	T	A
2	T	G	G	T	G	T	T
3	C	C	C	C	C	C	G
4	G	G	G	G	G	G	G
5	A	C	A	C	A	A	A
6	A	C	C	A	A	A	A

Data matrix

location	REF	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	0	0	1	0	1	2	0
2	0	1	2	0	1	0	0
3	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0
5	0	1	0	1	0	0	0
6	0	1	2	0	0	0	0

참조 유전체서열(REF)과 다르면 1 같으면 0

그렇다면 2가 의미하는 것은?

변이정보로부터

location	REF	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	0	0	1	0	1	2	0
2	0	1	2	0	1	0	0
3	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0
5	0	1	0	1	0	0	0
6	0	1	2	0	0	0	0

?

Classification

실습