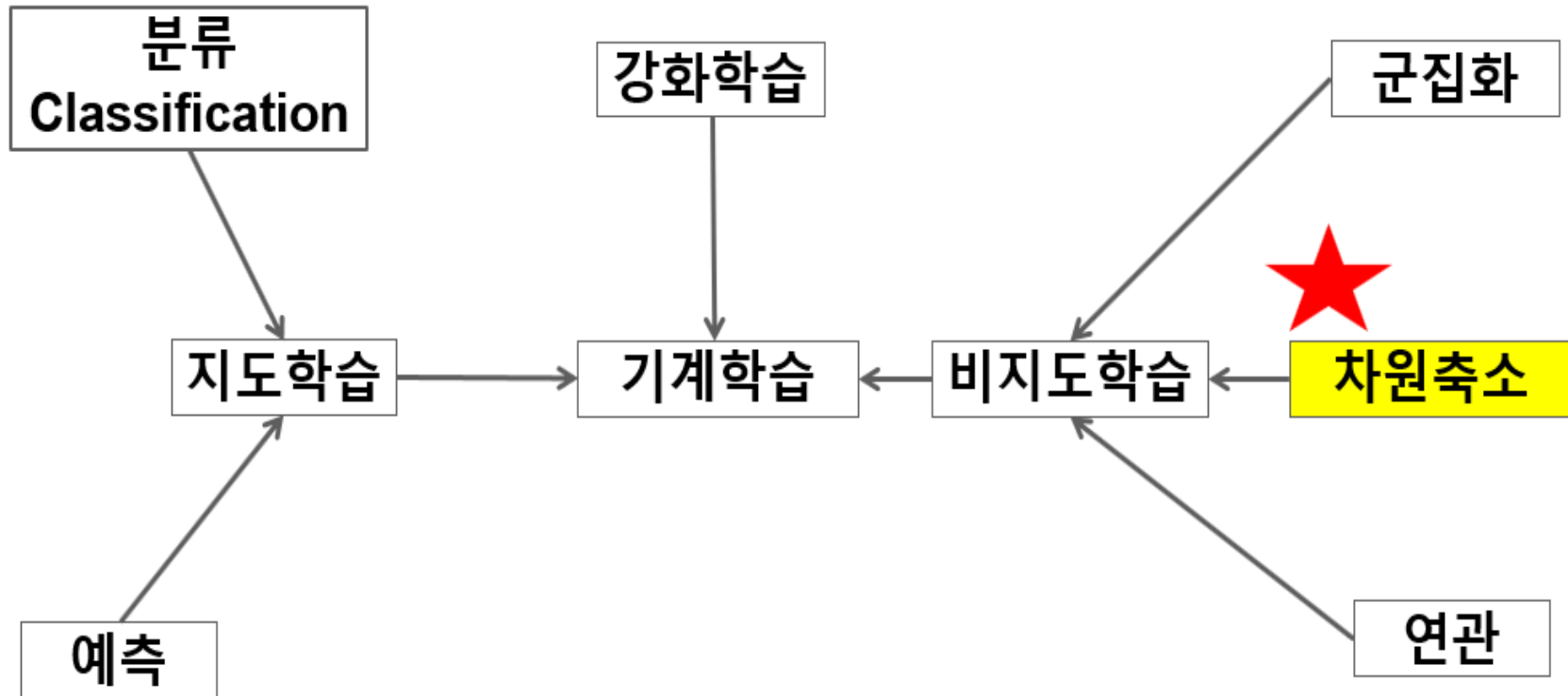


비지도학습(차원 축소)

홍익 대학교
Hyun-Sun Ryu

머신러닝의 종류



차원 축소(Dimension Reduction)

차원 축소

- 우리가 사용하는 데이터의 대부분은 가로 방향으로 개별 관찰값인 instance, 세로 방향으로 instance의 특징인 변수를 나타냄.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
6	0.06905	0	2.18	0	0.458	7.147	94.2	6.0622	3	222	18.7	396.9	5.35	36.2
7	0.02985	0	2.18	0	0.458	6.43	18	6.0622	22	187	14.1	395.6	5.2	28.7
8	0.08829	12.5	7.87	0	0.524	6.012	66	5.505	31	15.2	395.6	2.43	22.9	22.9
9	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
10	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.3
11	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
12	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
13	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
14	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
15	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
16	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
17	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
18	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
19	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
20	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2
21	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2
22	1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6
23	0.85204	0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21	392.53	13.83	19.6
24	1.23247	0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21	396.9	18.72	15.2
25	0.98843	0	8.14	0	0.538	5.813	100	4.0952	4	307	21	394.54	19.88	14.5
26	0.75026	0	8.14	0	0.538	5.924	94.1	4.3996	4	307	21	394.33	16.3	15.6
27	0.84054	0	8.14	0	0.538	5.599	85.7	4.4546	4	307	21	303.42	16.51	13.9
28	0.67191	0	8.14	0	0.538	5.813	90.3	4.682	4	307	21	376.88	14.81	16.6
29	0.95577	0	8.14	0	0.538	6.047	88.8	4.4534	4	307	21	306.38	17.28	14.8
30	0.77299	0	8.14	0	0.538	6.495	94.4	4.4547	4	307	21	387.94	12.8	18.4
31	1.00245	0	8.14	0	0.538	6.674	87.3	4.239	4	307	21	380.23	11.98	21
32	1.13081	0	8.14	0	0.538	5.713	94.1	4.233	4	307	21	360.17	22.6	12.7
33	1.35472	0	8.14	0	0.538	6.072	100	4.175	4	307	21	376.73	13.04	14.5
34	1.38799	0	8.14	0	0.538	5.95	82	3.99	4	307	21	232.6	27.71	13.2
35	1.15172	0	8.14	0	0.538	5.701	95	3.7872	4	307	21	358.77	18.35	13.1
36	1.61282	0	8.14	0	0.538	6.096	96.9	3.7598	4	307	21	248.31	20.34	13.5
37	0.06417	0	5.96	0	0.499	5.933	68.2	3.3603	5	279	19.2	396.9	9.68	18.9
38	0.09744	0	5.96	0	0.499	5.841	61.4	3.3779	5	279	19.2	377.56	11.41	20
39	0.08014	0	5.96	0	0.499	5.85	41.5	3.9342	5	279	19.2	396.9	8.77	21
40	0.17505	0	5.96	0	0.499	5.966	30.2	3.8473	5	279	19.2	393.43	10.13	24.7
41	0.02763	75	2.95	0	0.428	6.595	21.8	5.4011	3	252	18.3	395.63	4.32	30.8
42	0.03359	75	2.95	0	0.428	7.024	15.8	5.4011	3	252	18.3	395.62	1.98	34.9
43	0.12744	0	6.91	0	0.448	6.77	2.9	5.7209	3	233	17.9	385.41	4.84	26.6
44	0.1415	0	6.91	0	0.448	6.169	6.6	5.7209	3	233	17.9	383.37	5.81	25.3
45	0.15936	0	6.91	0	0.448	6.211	6.5	5.7209	3	233	17.9	394.46	7.44	24.7

instance

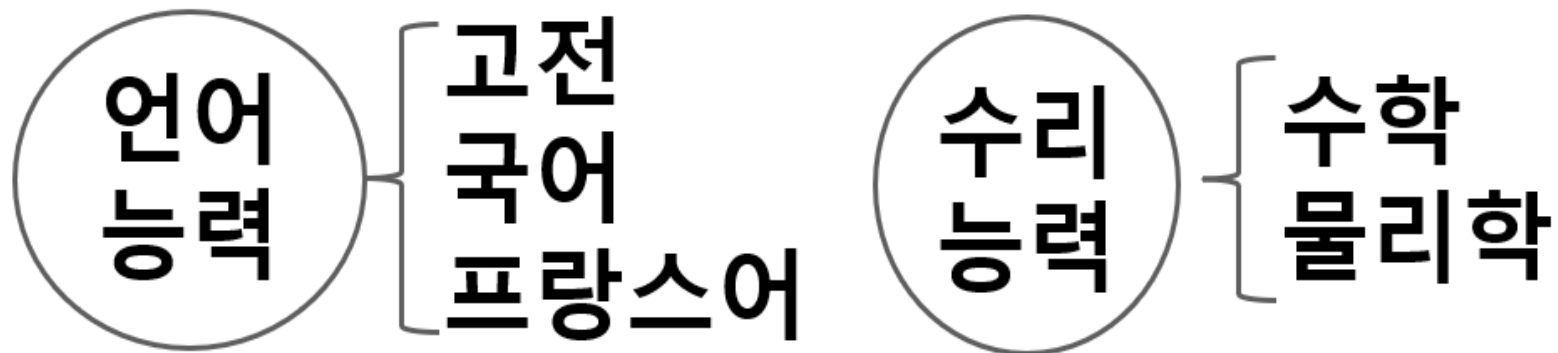
관찰값

변수

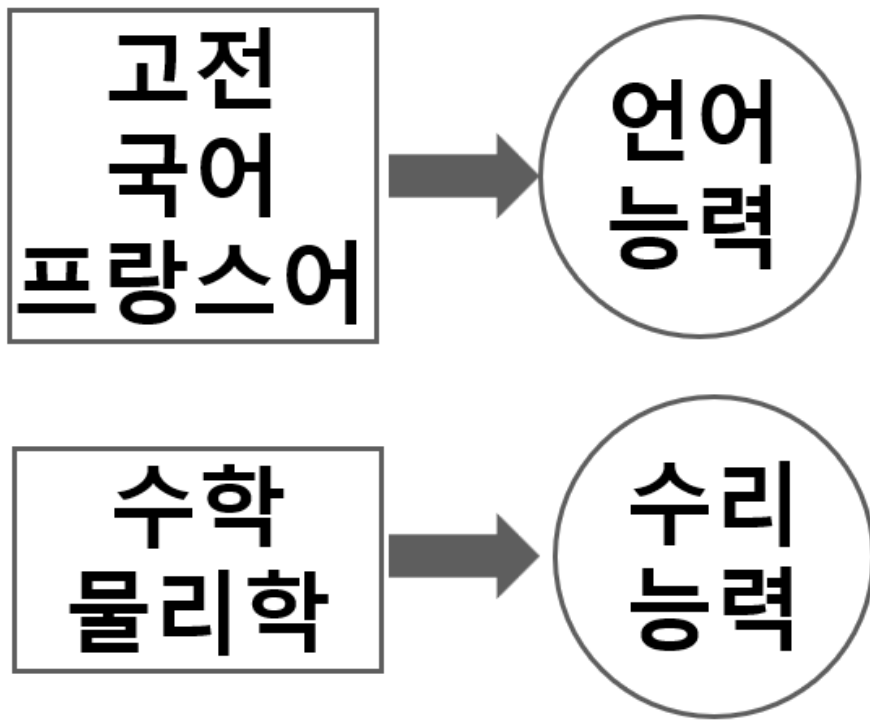
차원 축소

	고전	국어	프랑스어	수학	음악	소리·빛 반응
Student1	80	90	88	88	78	90
Student2	85	91	85	80	80	82
Student3	86	87	87	82	84	84
Student4

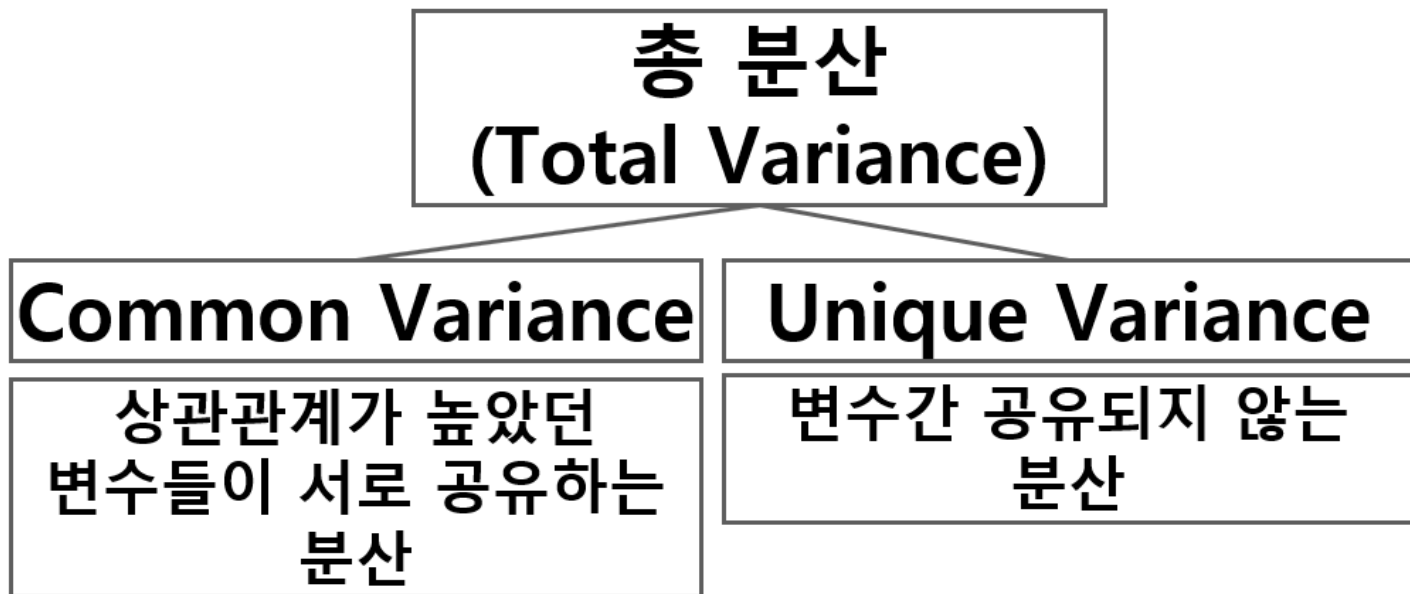
차원 축소



차원 축소



PCA(주성분분석)과 FA(요인분석)



PCA(주성분분석)과 FA(요인분석)

PCA(Principal Component Analysis)

- Total variance = Common variance

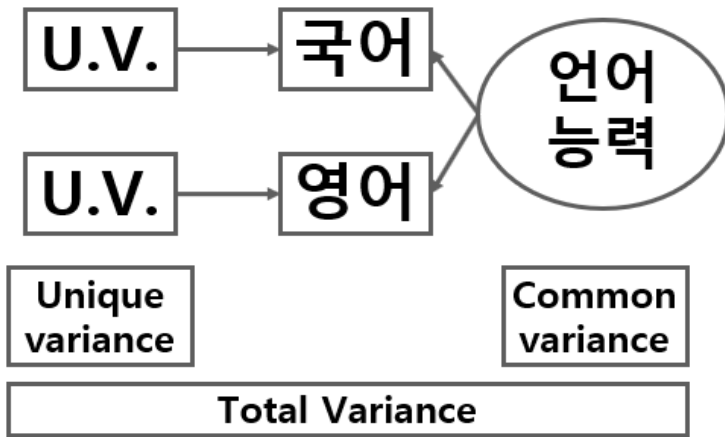
Common variance

Total variance

FA(Factor Analysis)

- Total variance =

Common variance + Unique variance

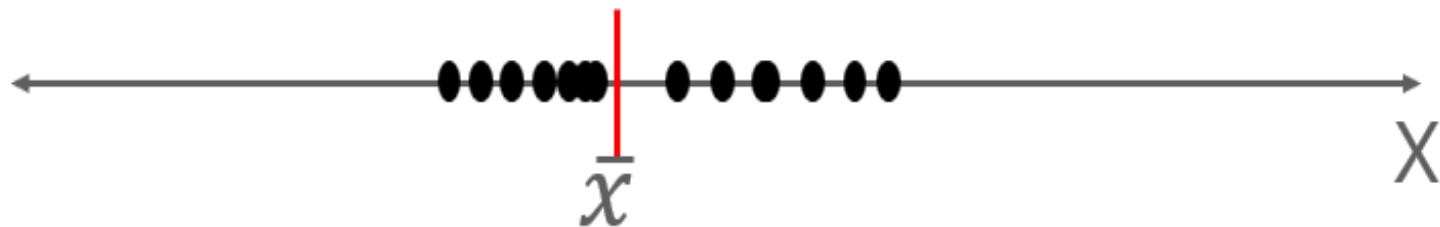


차원 축소

- 차원축소의 장점
 - 관측이 불가능한 보이지 않는 대상을 측정하기 위해 관측이 가능한 것을 이용하여 연구가 가능.
 - 고전, 국어, 프랑스어의 정량적으로 보이는 점수를 활용하여 보이지 않는 언어 능력을 연구
 - 너무 많은 변수를 줄여주어 보다 적은 변수로도 원하는 대상을 측정
- 차원축소의 단점
 - 차원축소가 데이터 과정의 만능이 될 수는 없음
 - 설명이 불가능한 경우가 많고 연구자의 의도대로 결과가 흘러갈 가능성
 - 정보의 손실이 발생하며 직관적인 이해가 어려움

상관관계와 분산, 공분산

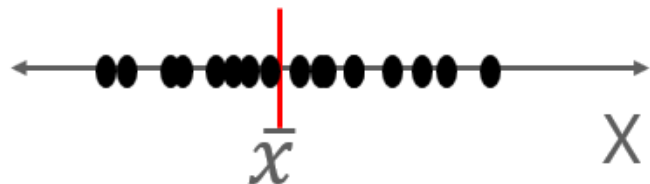
- 분산 = $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$



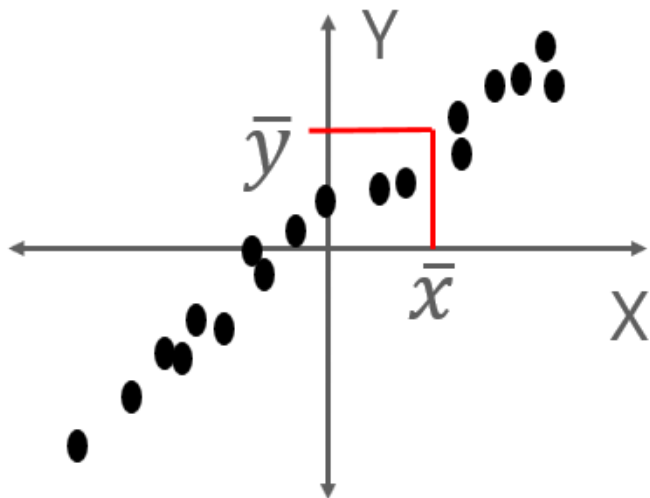
- 공분산(S_{xy}) = $\frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

상관관계와 분산, 공분산

1차원 = 변수 1개

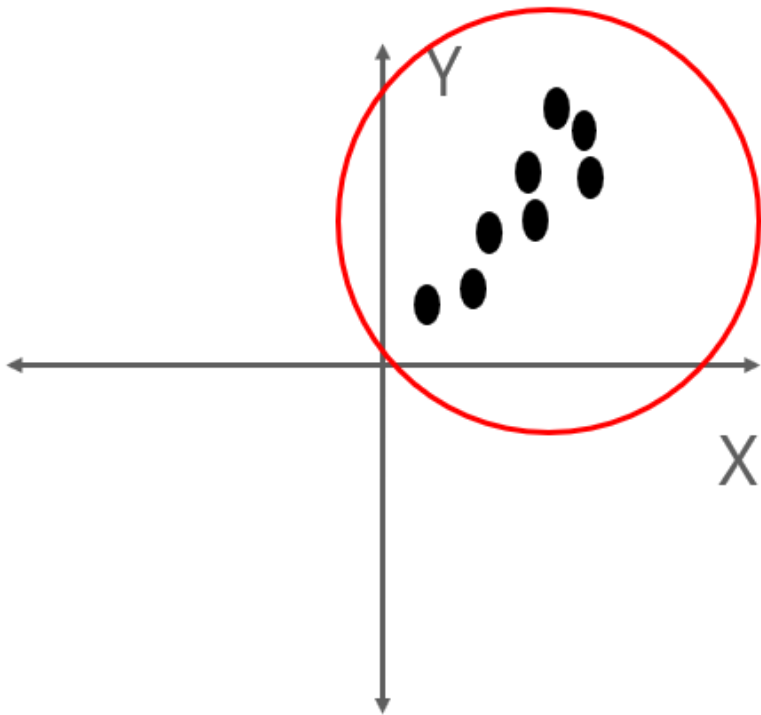


2차원 = 변수 2개
→ 공변량(같이 변하는 양)

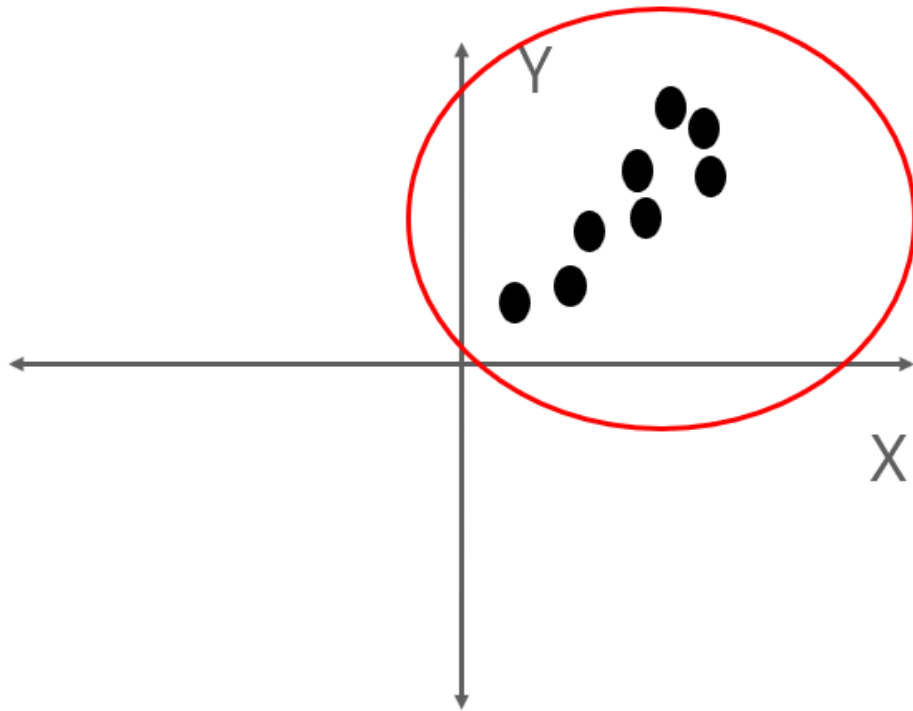


상관관계와 분산, 공분산

	고전(X)	국어(Y)
Student1	80	78
Student2	89	91
Student3	86	87
Student4	68	72

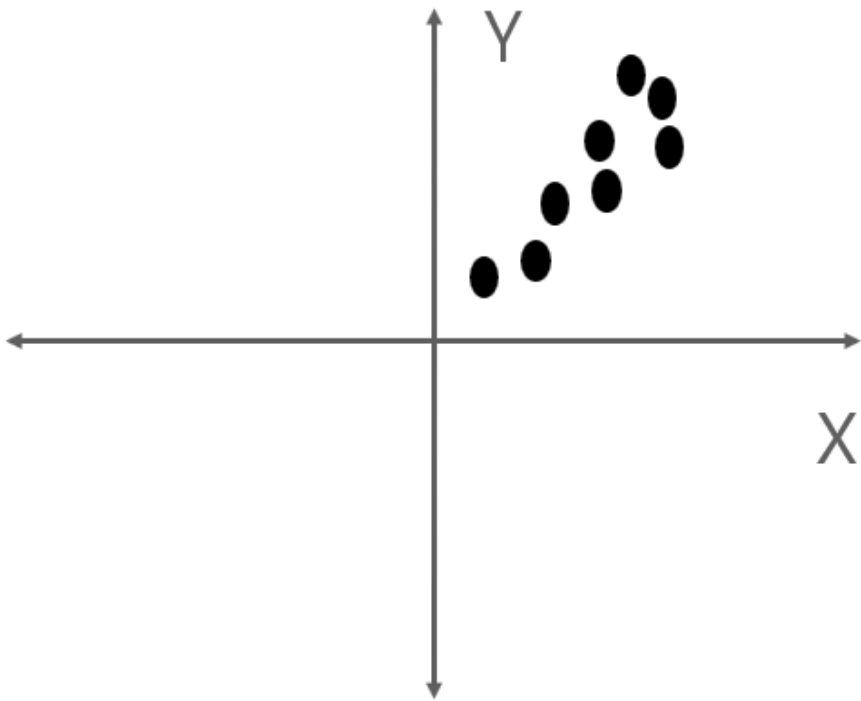


상관관계와 분산, 공분산

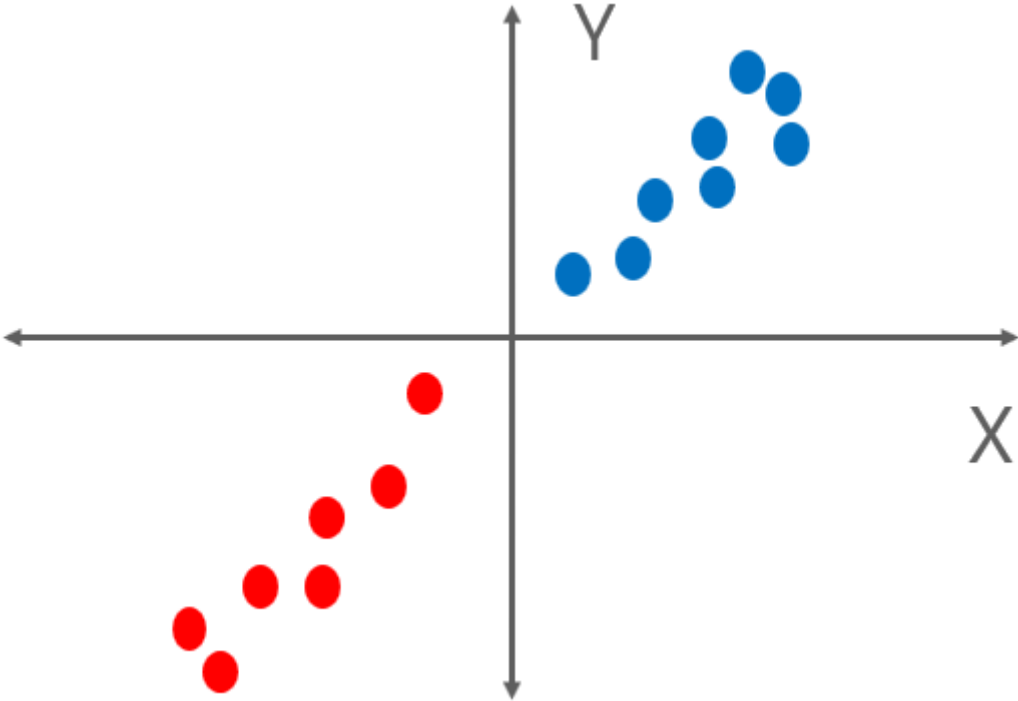


상관관계와 분산, 공분산

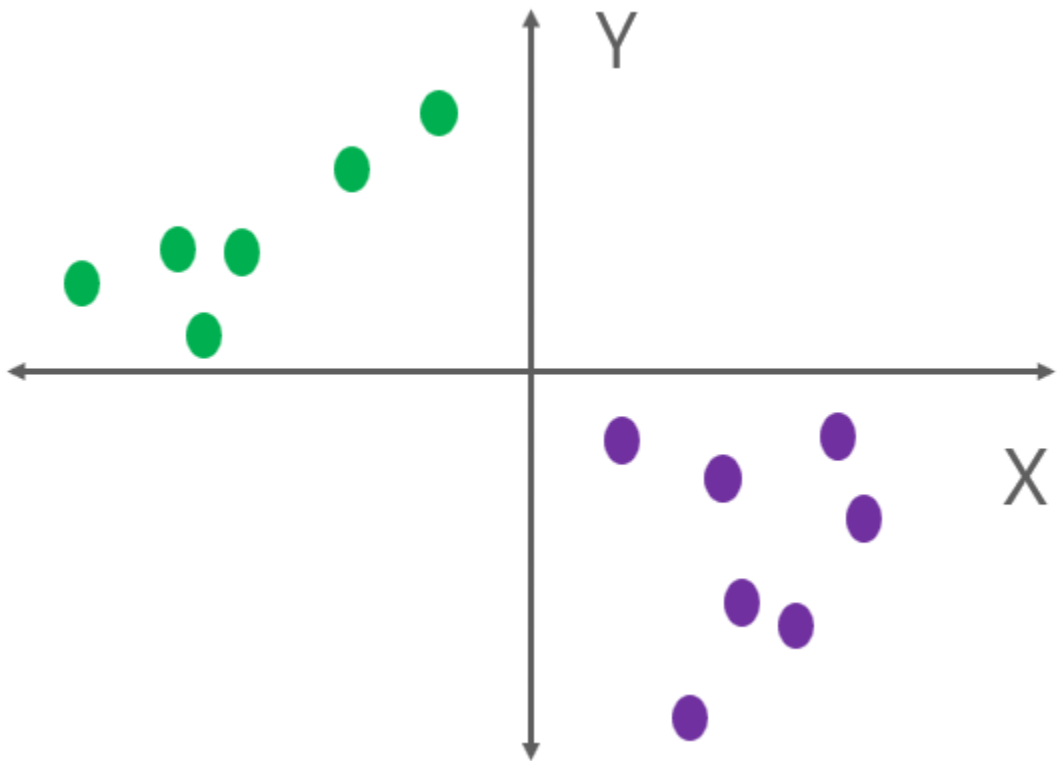
	고전	국어
고전	65	58
국어	58	56



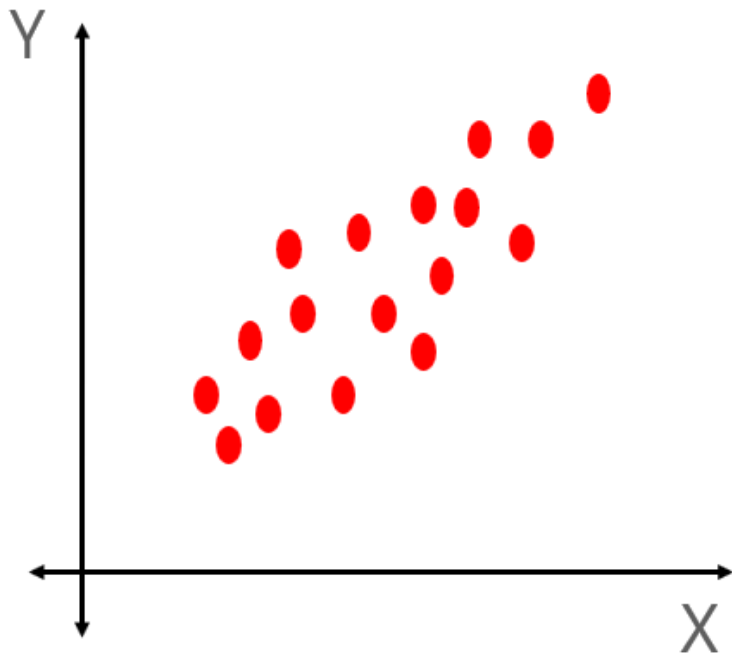
상관관계와 분산, 공분산



상관관계와 분산, 공분산

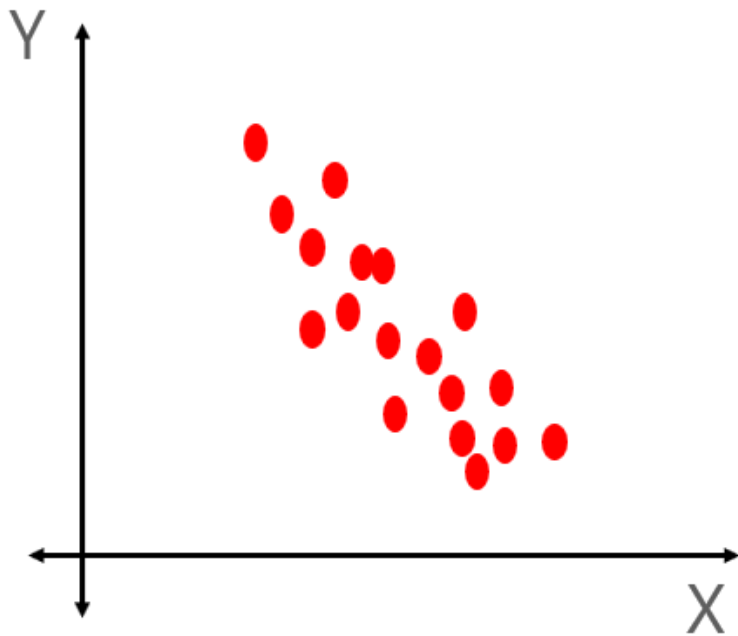


상관관계와 분산, 공분산



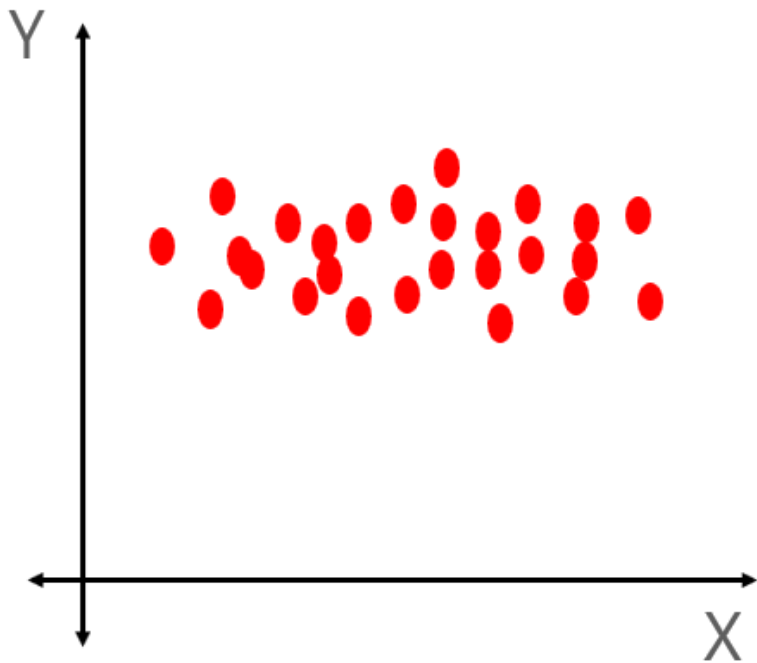
$$\begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix}$$

상관관계와 분산, 공분산



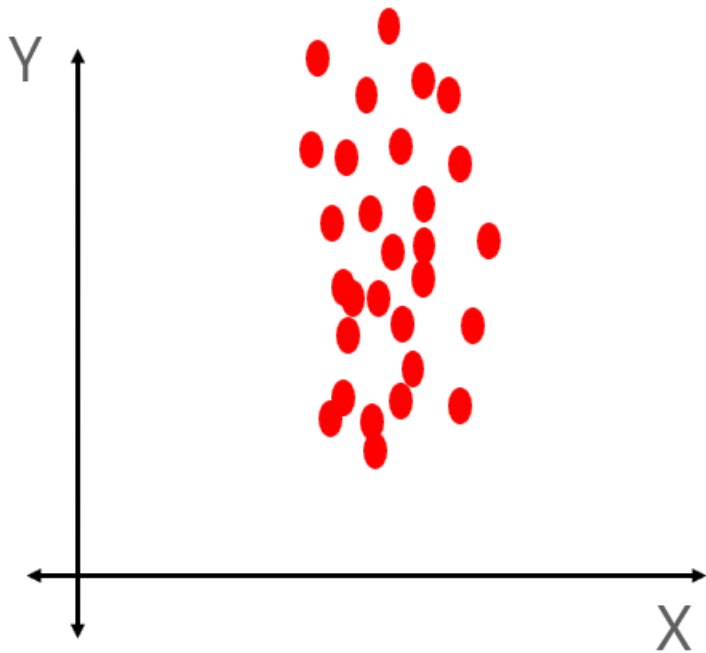
$$\begin{bmatrix} 5 & -4 \\ -4 & 6 \end{bmatrix}$$

상관관계와 분산, 공분산




$$\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$

상관관계와 분산, 공분산



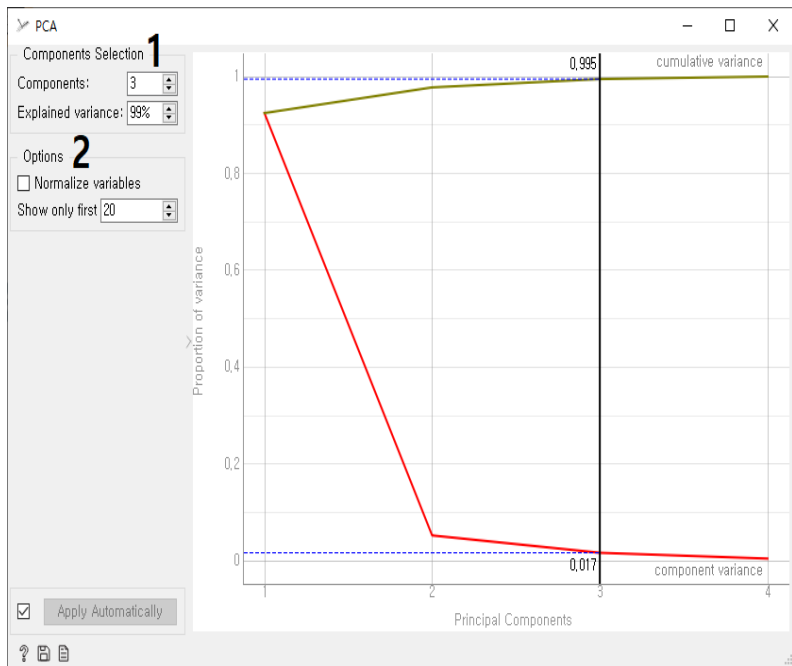
$$\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

Iris 데이터 활용하기(PCA)

위젯	설명	입력	출력
 PCA	입력 데이터의 선형 변환을 수행한다.	Data	Transformed data, Data, Components, PCA

- PCA 위젯은 입력 데이터의 PCA 선형 변환을 계산
- 개별 인스턴스의 가중치와 함께 변환된 데이터 세트 또는 주요 구성 요소의 가중치를 출력

Iris 데이터 활용하기(PCA)



① Components Selection	출력에서 원하는 주성분 수를 선택한다. 분산이 가능한 한 높게 적용된 상태에서 가능한 적게 선택하는 것이 좋다. 주성분으로 포함할 분산의 양을 설정할 수도 있다.
② Options	데이터를 정규화하여 값을 공통 척도로 조정할 수 있다.

아이리스 데이터

- 아이리스는 꽃잎의 모양과 길이에 따라 여러 가지 품종으로 나뉨
- 사진을 보면 품종마다 비슷해 보이는데 과연 딥러닝 모델을 사용하여 이들을 구별해 낼 수 있을까?



Iris-virginica



Iris-setosa



Iris-versicolor

아이리스 데이터

		속성				클래스
		정보 1	정보 2	정보 3	정보 4	품종
샘플	1번째 아이리스	5.1	3.5	4.0	0.2	Iris-setosa
	2번째 아이리스	4.9	3.0	1.4	0.2	Iris-setosa
	3번째 아이리스	4.7	3.2	1.3	0.3	Iris-setosa

	150번째 아이리스	5.9	3.0	5.1	1.8	Iris-virginica

<표> 아이리스 데이터의 샘플, 속성, 클래스 구분

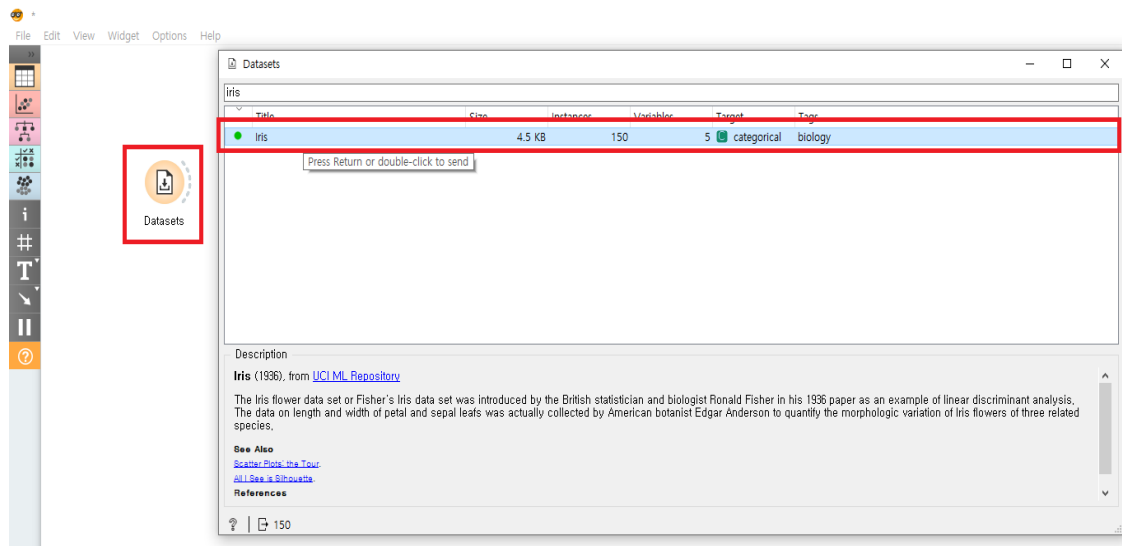
아|오|리|스|터|이|터|

- 샘플 수: 150
- 속성 수: 4
 - 정보 1: 꽃받침 길이 (sepal length, 단위: cm)
 - 정보 2: 꽃받침 너비 (sepal width, 단위: cm)
 - 정보 3: 꽃잎 길이 (petal length, 단위: cm)
 - 정보 4: 꽃잎 너비 (petal width, 단위: cm)
- 클래스: Iris-setosa, Iris-versicolor, Iris-virginica

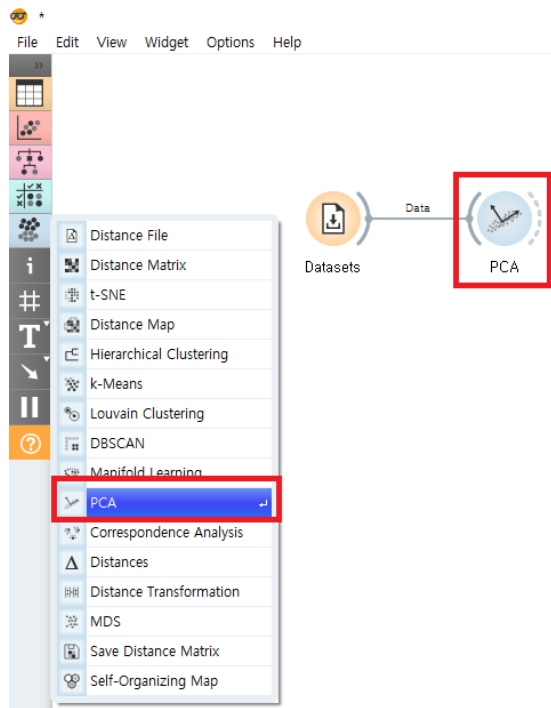
[illegible]

Iris 데이터 활용하기(PCA)

- Iris데이터를 기반으로 PCA를 통해 대규모 데이터 세트의 시각화를 단순화
- PCA위젯을 활용하면 개별 instance의 가중치와 함께 변환된 데이터 세트 또는 주요 구성 요소의 가중치를 출력
- Data메뉴에서 datasets위젯을 클릭하여 Iris데이터를 불러옴



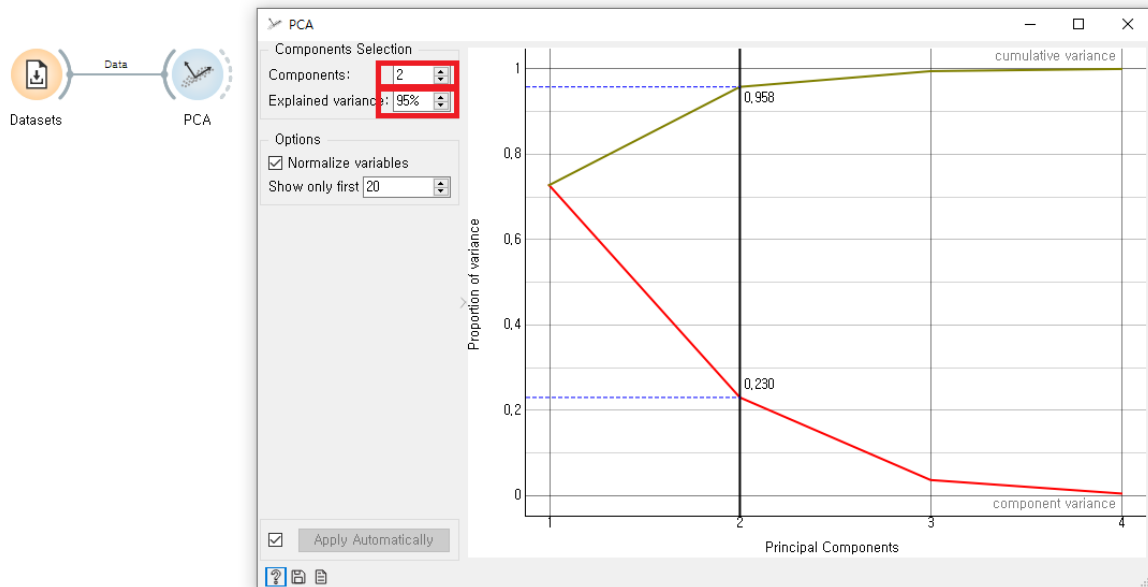
Iris 데이터 활용하기(PCA)



- Iris 데이터 세트에 PCA위젯을 연결
- PCA위젯은 왼쪽 Unsupervised learning 메뉴에서 PCA를 클릭하거나 드래그&드랍
- 데이터 세트에 연결된 PCA위젯을 더블클릭하여 그 값을 조정

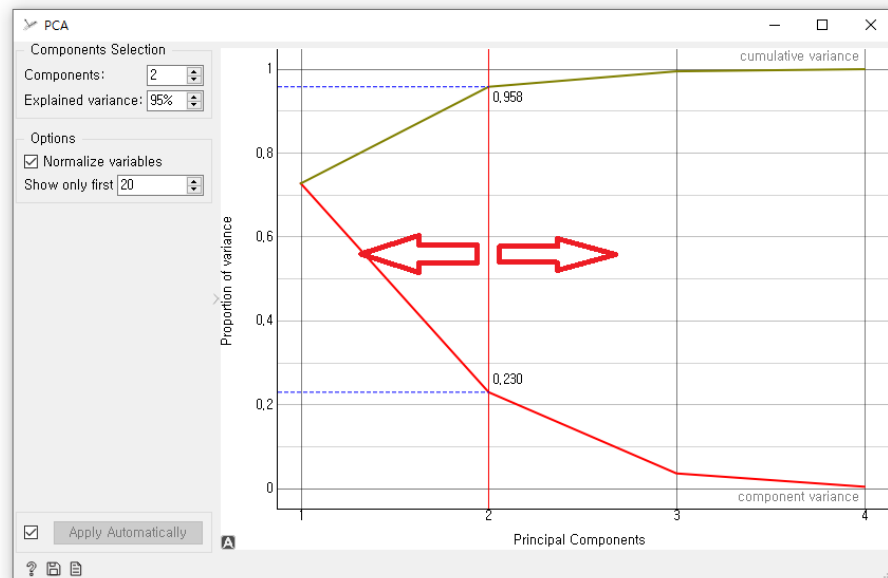
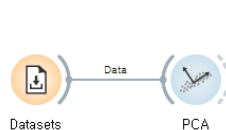
Iris 데이터 활용하기(PCA)

- 먼저 출력에서 원하는 주성분 수를 선택
- 기존 변수보다 차원을 축소하여 새로운 변수를 만들 수를 정할 수 있음. 현재는 2개
- 다음으로 주성분으로 포함할 분산의 양을 %로 정함. 현재는 95%



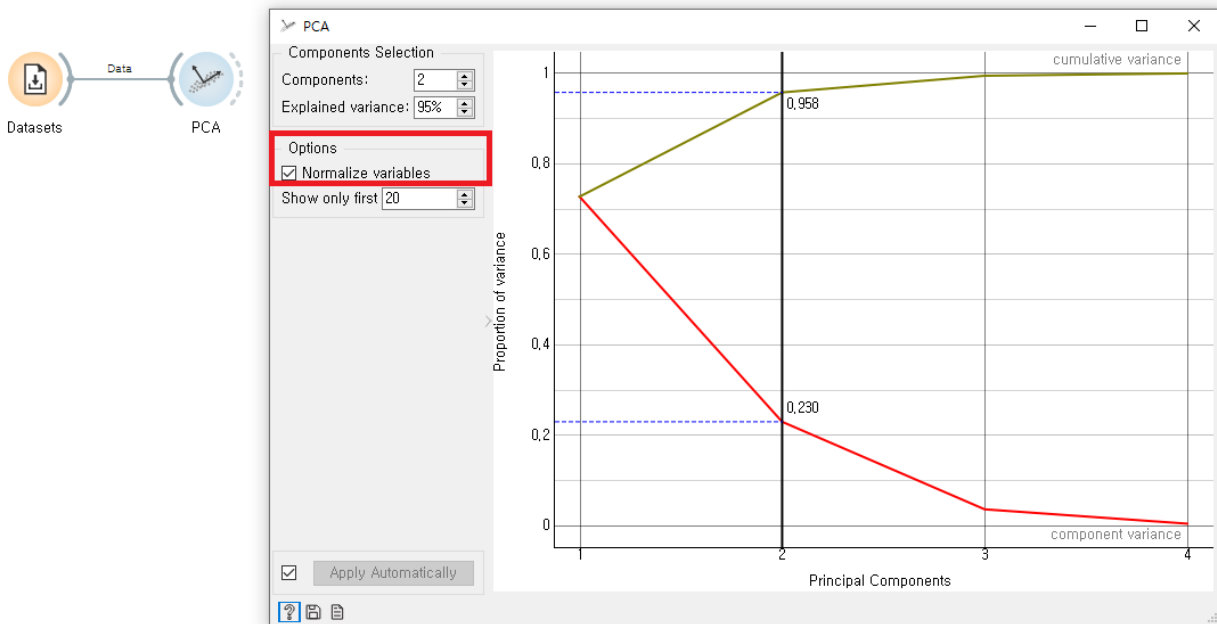
Iris 데이터 활용하기(PCA)

- 주성분의 갯수와 주성분으로 포함할 분산의 비율을 막대를 옮김으로써 설정을 변경
- 주성분의 수가 많아짐에 따라 포함된 분산의 비율 또한 늘어남을 확인



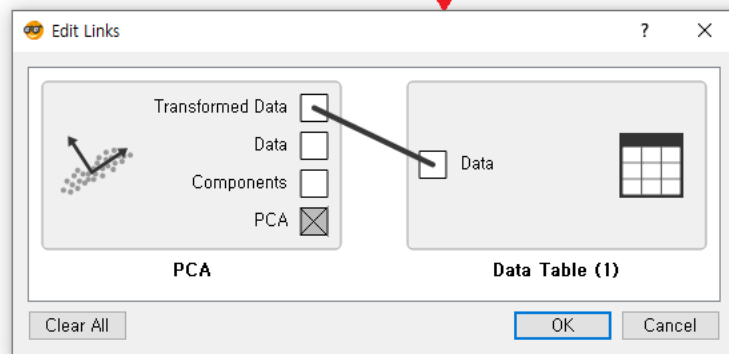
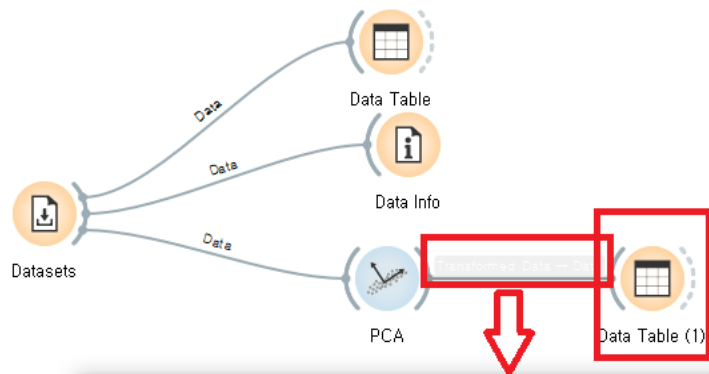
Iris 데이터 활용하기(PCA)

- 체크 박스 활성화를 통해 데이터를 정규화하여 값을 공통 척도로 조정



Iris 데이터 활용하기(PCA)

- PCA를 통해 데이터가 어떻게 변환됐는지 확인하기 위해 data table 위젯을 추가
- 두 위젯 사이의 연결은 transformed data to data로 지정



Iris 데이터 활용하기(PCA)

Data Table

Info
150 instances (no missing data)
2 features
Target with 3 values
No meta attributes

Variables
☒ Show variable labels (if present)
☒ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

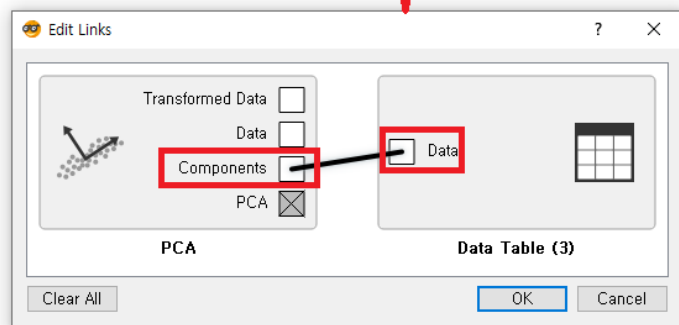
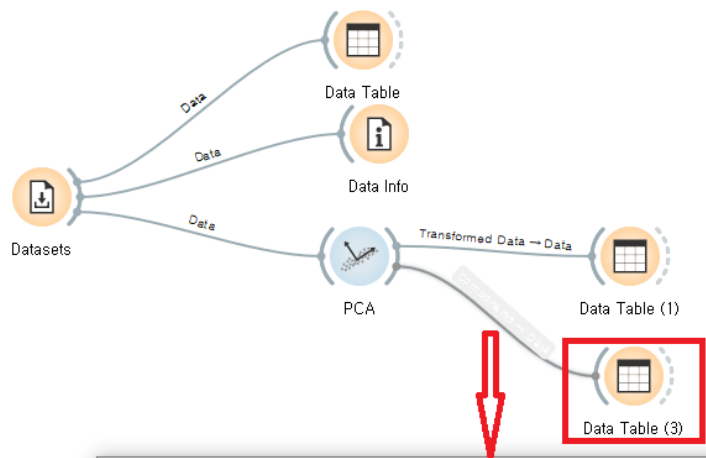
☒ Send Automatically

	iris	PC1	PC2
1	Iris-setosa	-2.68421	0.326607
2	Iris-setosa	-2.71539	-0.169557
3	Iris-setosa	-2.88982	-0.137346
4	Iris-setosa	-2.74644	-0.311124
5	Iris-setosa	-2.72859	0.333925
6	Iris-setosa	-2.2799	0.747783
7	Iris-setosa	-2.82089	-0.0821045
8	Iris-setosa	-2.62648	0.170405
9	Iris-setosa	-2.88796	-0.570798
10	Iris-setosa	-2.67384	-0.106692
11	Iris-setosa	-2.50653	0.651935
12	Iris-setosa	-2.61314	0.0215206
13	Iris-setosa	-2.78743	-0.22774
14	Iris-setosa	-3.2252	-0.50328
15	Iris-setosa	-2.64354	1.18619
16	Iris-setosa	-2.38387	1.34475
17	Iris-setosa	-2.62253	0.81809
18	Iris-setosa	-2.64832	0.319137
19	Iris-setosa	-2.19908	0.879244
20	Iris-setosa	-2.58735	0.520474
21	Iris-setosa	-2.31053	0.397868
22	Iris-setosa	-2.54323	0.440032
23	Iris-setosa	-3.21586	0.141616
24	Iris-setosa	-2.30313	0.105523
25	Iris-setosa	-2.35617	-0.0312096

- 두개의 차원으로 축소된 데이터에서 각 차원이 붓꽃의 품종을 구별하는데 얼마나 영향을 미치는지 상관관계를 숫자 및 그래프로 파악

상관관계와 분산, 공분산

- 각 차원이 품종별로 얼마나 상관관계가 있는지 한눈에 파악하기 위해 data table 위젯을 추가하고 두 위젯의 연결을 components to data로 설정



상관관계와 분산, 공분산

Data Table (1)

Info
2 instances (no missing data)
4 features
No target variable.
1 meta attribute

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

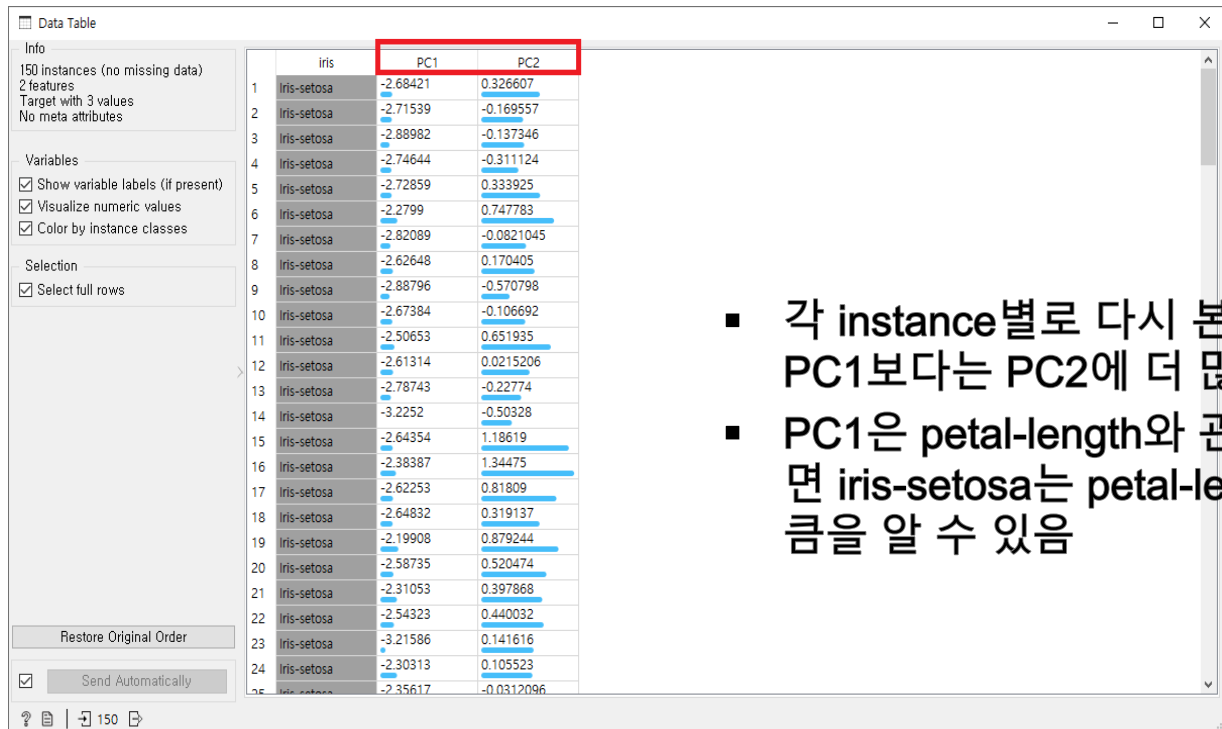
Restore Original Order

☒ Send Automatically

components	sepal length	sepal width	petal length	petal width
PC1	0.36159	-0.0822689	0.856572	0.358844
PC2	0.65654	0.729712	-0.175767	-0.0747065

- Data table 위젯창을 활성화해보면 각 차원 (PC1~PC2)가 품종별로 얼마나 영향을 미치는지 아래와 같이 파악 가능
- 예를 들어 PC1은 petal-length와는 큰 상관관계가 있지만 sepal-width와는 거리가 멀다고 할 수 있음.

상관관계와 분산, 공분산



Data Table

Info
150 instances (no missing data)
2 features
Target with 3 values
No meta attributes

Variables
☒ Show variable labels (if present)
☒ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	iris	PC1	PC2
1	Iris-setosa	-2.68421	0.326607
2	Iris-setosa	-2.71539	-0.169557
3	Iris-setosa	-2.88982	-0.137346
4	Iris-setosa	-2.74644	-0.311124
5	Iris-setosa	-2.72859	0.333925
6	Iris-setosa	-2.2799	0.747783
7	Iris-setosa	-2.82089	-0.0821045
8	Iris-setosa	-2.62648	0.170405
9	Iris-setosa	-2.88796	-0.570798
10	Iris-setosa	-2.67384	-0.106692
11	Iris-setosa	-2.50653	0.651935
12	Iris-setosa	-2.61314	0.0215206
13	Iris-setosa	-2.78743	-0.22774
14	Iris-setosa	-3.2252	-0.50328
15	Iris-setosa	-2.64354	1.18619
16	Iris-setosa	-2.38387	1.34475
17	Iris-setosa	-2.62253	0.81809
18	Iris-setosa	-2.64832	0.319137
19	Iris-setosa	-2.19908	0.879244
20	Iris-setosa	-2.58735	0.520474
21	Iris-setosa	-2.31053	0.397868
22	Iris-setosa	-2.54323	0.440032
23	Iris-setosa	-3.21586	0.141616
24	Iris-setosa	-2.30313	0.105523
25	Iris-setosa	-2.35617	-0.0312096

- 각 instance별로 다시 본다면, iris-setosa의 경우 PC1보다는 PC2에 더 많은 가중치가 있다고 나타남
- PC1은 petal-length와 관련이 있기 때문에 바꿔 말하면 iris-setosa는 petal-length보다는 다른 요인이 더 큼을 알 수 있음

질문 있나요?

hsryu13@hongik.ac.kr

