Big data
알아보 '조'

조원 : 이동희 권형준 이송은 김주영
이동희 이희범 오선민 이상훈

Contents

1 Supervised & Unsupervised

2 Principal Component Analysis

3 Random Forest

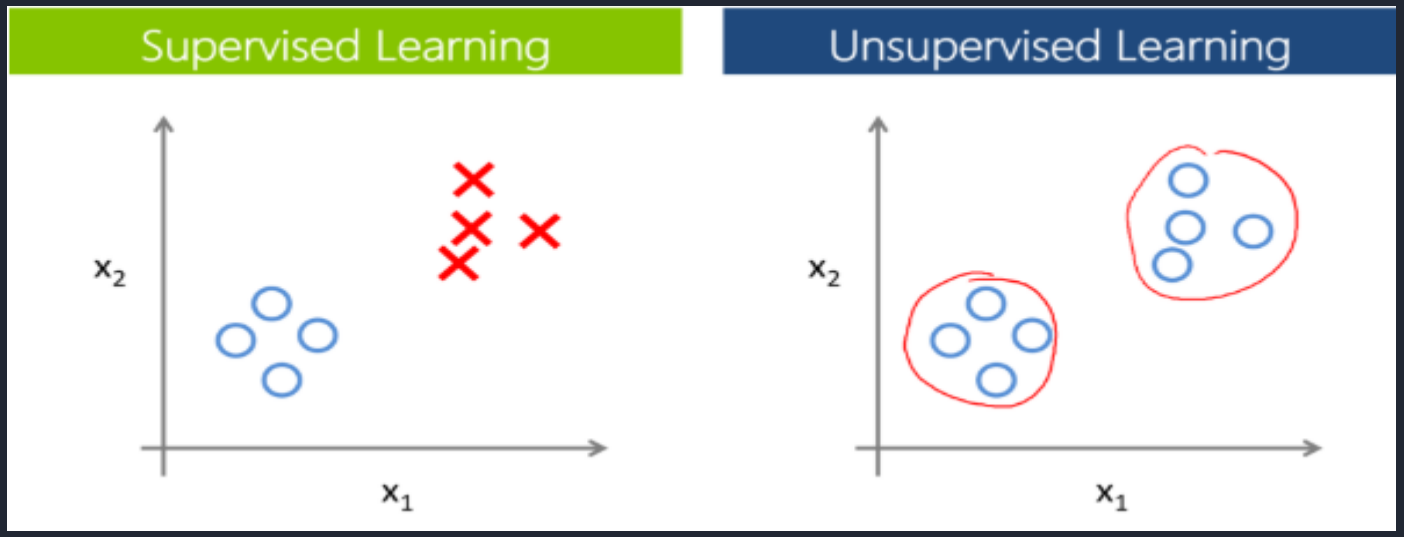
4 Compare & Conclusion

5 Discussion

6 QnA

Supervised & Unsupervised

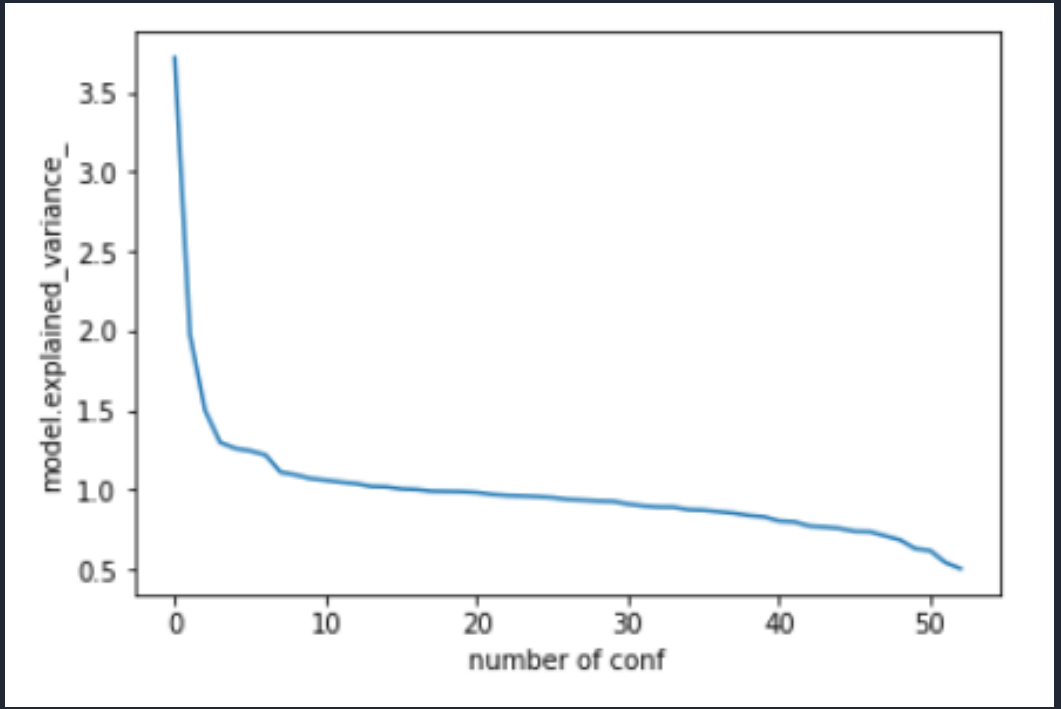
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Dimension Reduction



Principal Component Analysis

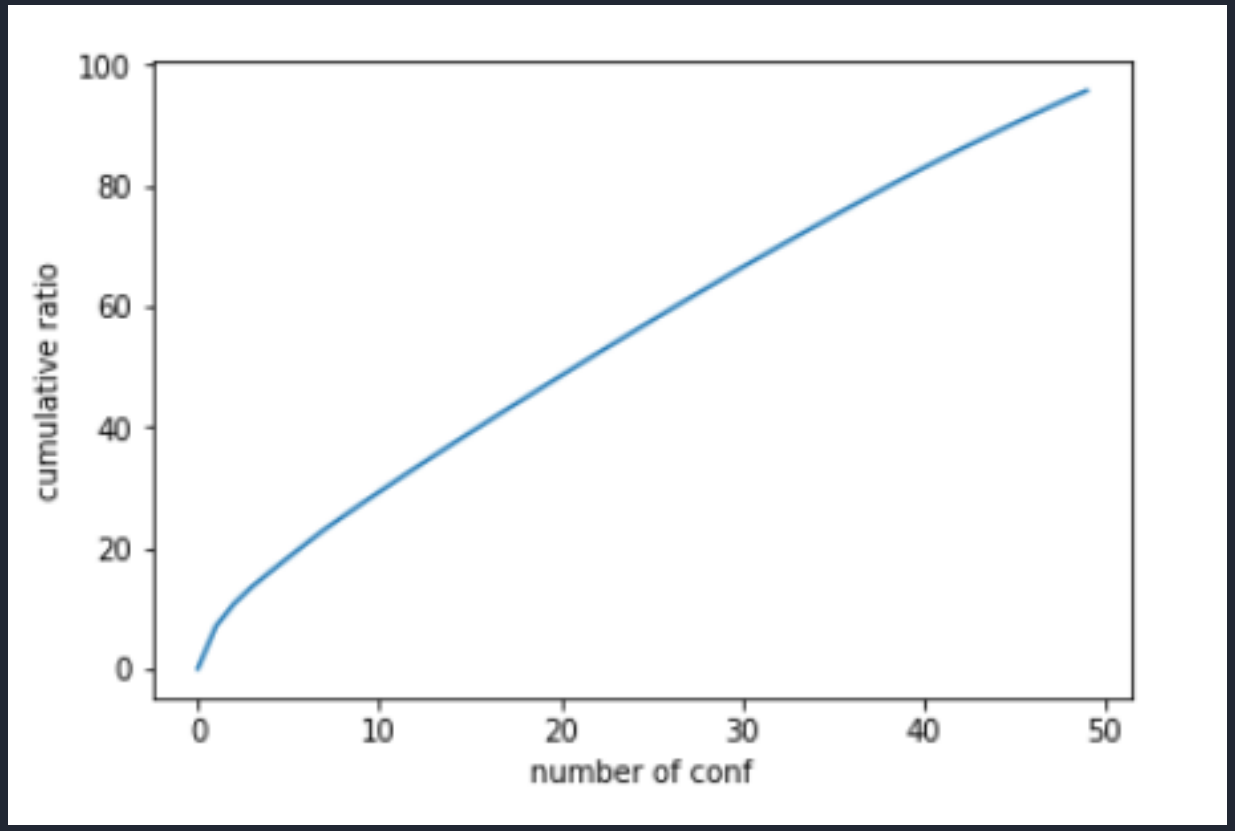
- PCA
  - Dimension Reduction
  - Covariance

- PCA Analysis



Principal Component Analysis

- PCA Analysis



Principal Component Analysis

- PCA Data (Columns = 50)

-0.663158	0.678356	-0.033575	-0.782518	-0.499932	-1.186263	-0.309962	-0.063758	0.286009	-0.430581	...
-1.308782	-0.974940	0.009729	0.622225	-0.278115	-0.092327	-0.245752	0.748239	-0.281921	-1.397178	...
4.875452	-2.109782	-0.711727	1.184342	-0.035985	0.142653	0.456267	0.070398	-1.151501	-1.021532	...
3.626298	3.234188	1.028590	-5.425468	10.004434	0.355383	-3.771828	2.786469	-0.632918	-0.650267	...
-0.049022	0.497845	0.983549	0.056858	-0.083463	-0.502875	-0.661337	0.525266	-0.143771	0.145604	...
-1.015137	-0.435495	0.039161	-0.083400	-0.313453	-0.309702	-0.137740	0.010962	0.206832	-0.130738	...
-1.060045	-0.386080	-0.109975	0.015473	0.609282	-0.016394	0.158291	0.303236	-0.135463	0.094517	...
-0.793660	-0.467950	0.189265	-0.011924	-0.143351	-0.169061	-0.248473	0.205356	-0.118697	0.007187	...
-0.794567	-0.077408	-0.137458	-0.175391	0.596502	-0.366291	-0.041669	0.047191	0.050634	0.273177	...
-0.456001	-0.616192	-0.071286	-0.089501	0.170934	-0.140204	-0.219108	0.063431	0.242190	0.290138	...
-0.637180	0.385284	-0.665826	0.653875	0.766670	-0.468042	0.698846	1.222156	-0.848033	0.951623	...
1.187273	0.795225	-3.300769	-1.008055	-3.299496	3.738458	-4.537206	3.195876	-0.958527	-2.140704	...

Random Forest

- Random ForestClassifier
  - Bootstrap
  - Overfitting

- Random ForestClassifier Analysis

Random Forest

- Random ForestClassifier Analysis

pca + random Frest	n_estimators	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split	max_features	accuracy
pca.columns = 40	500	50	200	100	5	6	0.6186
	500	50	200	100	5	7	0.61996
	500	50	200	100	5	8	0.61948
	500	50	200	100	5	9	0.62111
	500	50	200	100	5	10	0.62045

pca + random Frest	n_estimators	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split	max_features	accuracy
pca.columns = 40	250	50	200	100	5	9	0.6223
	500	50	200	100	5	9	0.62111
	1000	50	200	100	5	9	0.62133

Random Forest

- Random ForestClassifier Analysis

pca + random Frest	n_estimators	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split	max_features	accuracy
pca.columns = 50	500	50	200	100	5	7	0.62137
	500	50	200	100	5	8	0.62164
	500	50	200	100	5	9	0.62313
	500	50	200	100	5	10	0.62287
	500	50	200	100	5	11	0.62172

pca + random Frest	n_estimators	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split	max_features	accuracy
pca.columns = 50	250	50	200	100	5	9	0.62283
	500	50	200	100	5	9	0.62313
	1000	50	200	100	5	9	0.62494

Compare

- Random ForestClassifier

random Frest	n_estimators	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split	max_features	accuracy
	1000	50	200	100	5	9	0.503414
	10000	50	200	100	5	9	0.50491

- PCA and Random ForestClassifier

pca + random Frest	n_estimators	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split	max_features	accuracy
pca.columns = 53	1000	50	200	100	5	9	0.62608
	2000	50	200	100	5	9	0.62653
	10000	50	200	100	5	9	None

Discussion

개선사항

1. 여러 가지 모델 비교를 하지 못한 점
2. 결과 데이터가 낮은 점
3. 시간이 오래걸려서 accuracy 값 못구한점
4. 그외 기타.....

QnA