3rd week

목차

- Ensemble
- Unsupervised Learning?
- Clustering
- Dimensionality Reduction

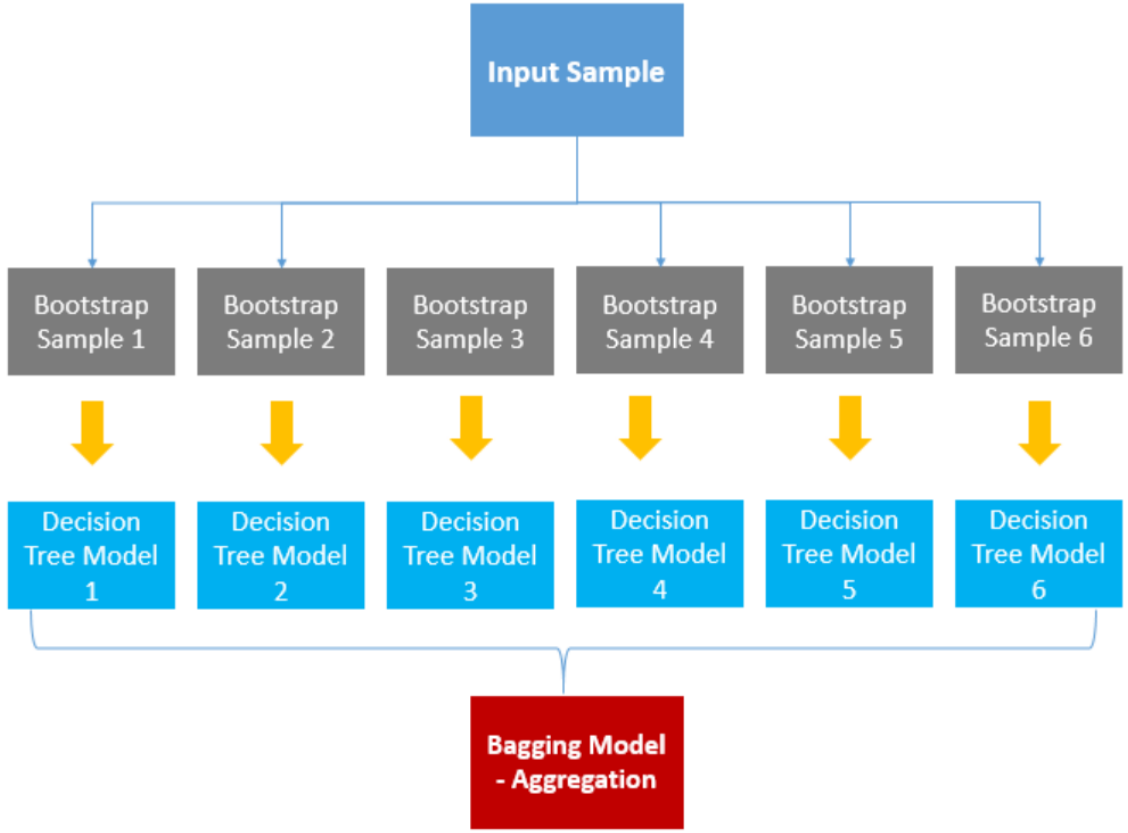
Ensemble

한 데이터셋을 쪼개서 학습 여러 번 하는 방식.

- Bagging (Bootstrap Aggregation)
- Boosting
- RandomForest

Bagging

샘플 여러 개 쪼개서 그걸 따로 학습. 마지막에 Avg 내는 구조



bootstrap:

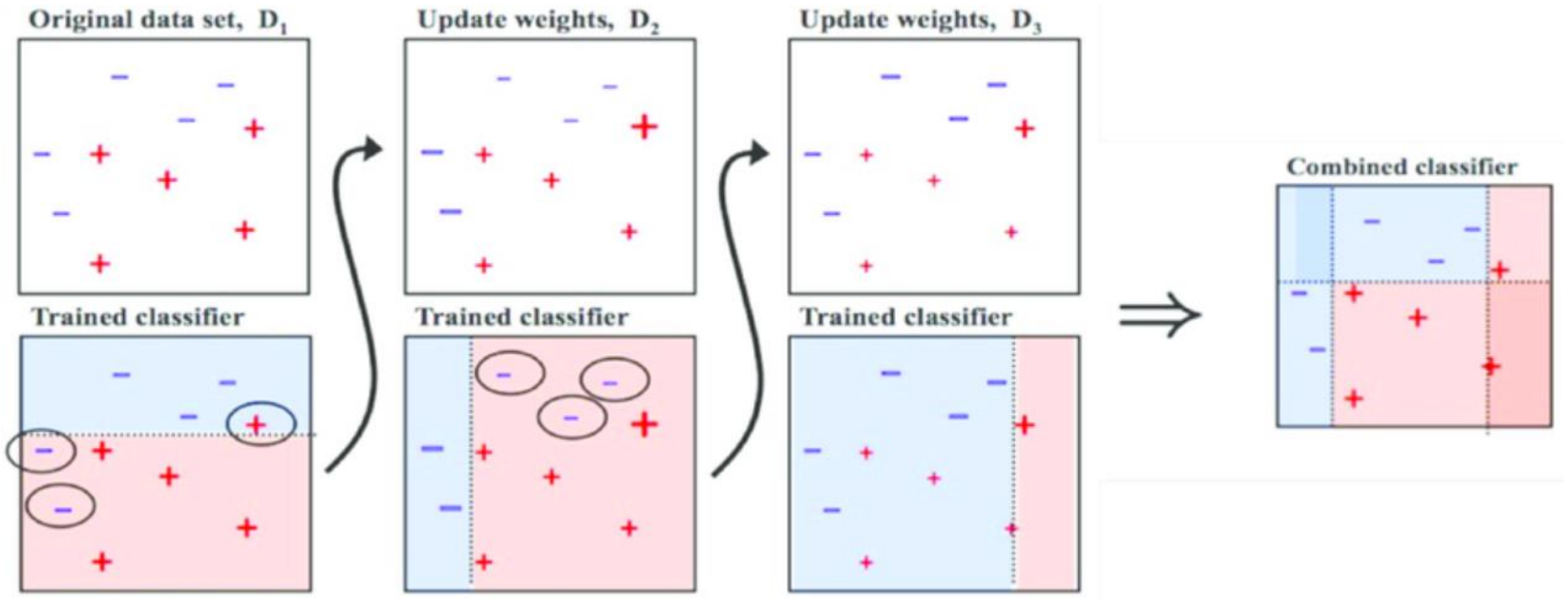
데이터를 쪼개서 샘플링 한다는 말

RandomForest

Extra Trees는 Bagging이라 할 수 없다.

Boosting

이전 학습에서 잘못 분류한 것에 더 큰 가중치 부여 후 학습.



overfitting에 취약

- AdaBoost
- Gradient Boosting
- XGBoost

등등

Bagging vs Boosting

배깅은 병렬적 학습.

부스팅은 순차적 학습.

이런 차이 존재.

그럼 언제 뭘 써야 하는가?

정확도가 너무 낮다... ➔ Boosting

과적합인거 같은데? ➔ Bagging

Unsepervised Learning

여태까지 0,1값 억지로 만들어 줬던거 기억나는가?

이 데이터가 뭐다~ 라는걸 의미해요.

그래서 기존까지는 전부 지도학습.

그럼 비지도 학습은 뭔데?

라벨링, 어떤 정답 지표가 없는 데이터의 패턴 찾는 방식.

Unsepervised Learning

- Clustering(군집화)

- Dimensionality Reduction(차원 축소)

등등

Clustering

- **K-Means Clustering**
  (거리 기반)

- **Hierarchical Clustering (계층적 클러스터링)**
  (거리 기반)

- **DBSCAN(Density-Based Spatial Clustering of Applications with Noise)**
  (밀도기반)

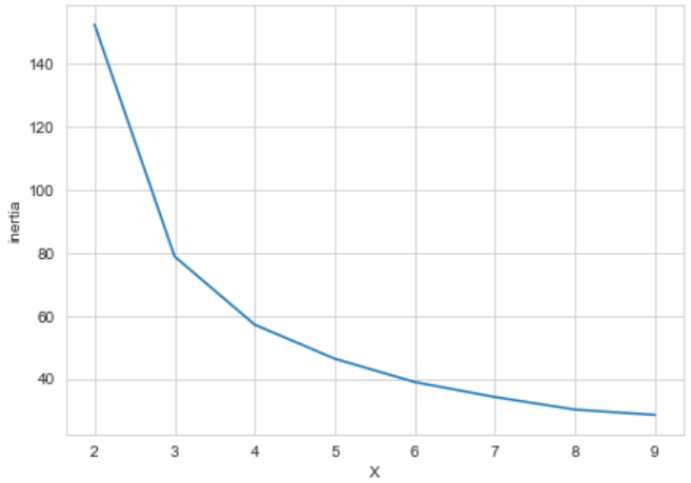
• K-Means Clustering

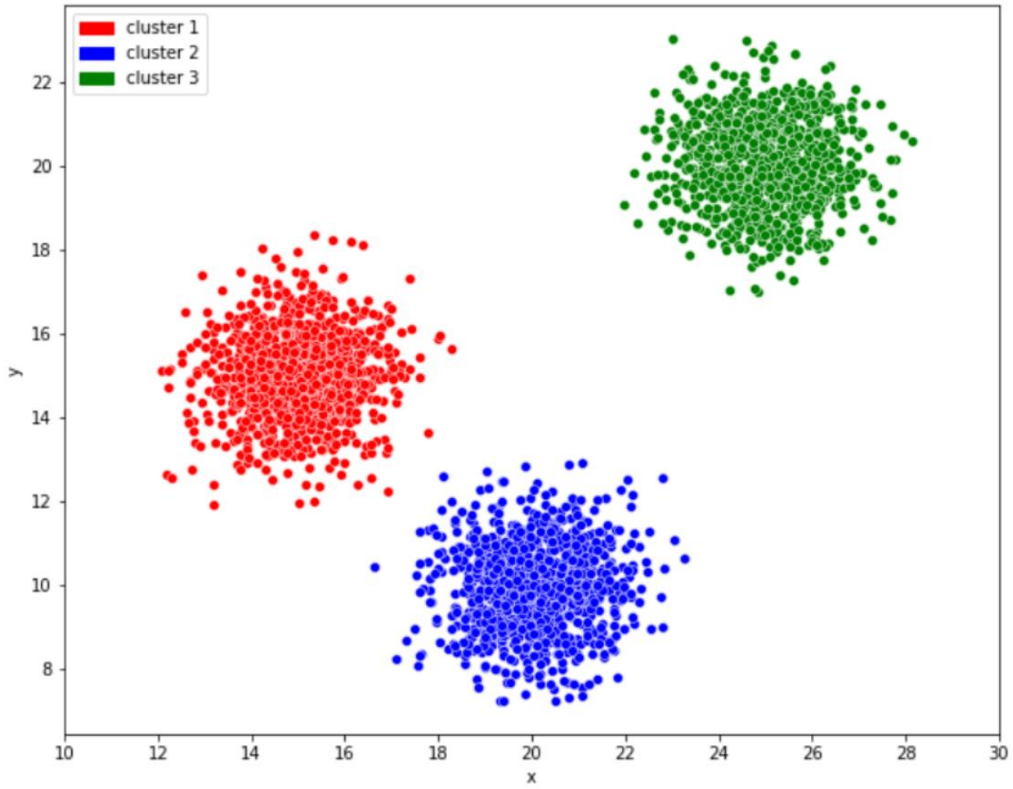
말 그대로 k개의 묶음.

k값을 지정해줘야 함.

K값 지정 방식:

엘보우 기법 사용.





$$\text{Inertia} = \sum_{i=1}^{n} \min_{\mu_j \in C} \left(||x_i - \mu_j||^2 \right)$$

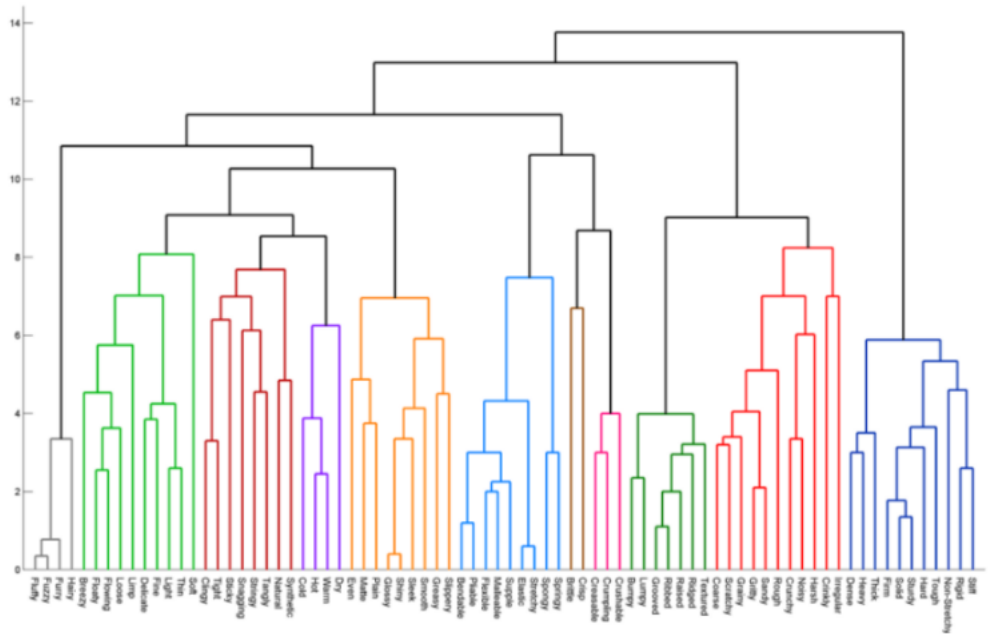
<- 이너셔

중심점과의 거리 측정.

주로 유클리디안 distance

Hierarchical Clustering

K-means와 비슷한 느낌, 거리 계산 후 클러스터링



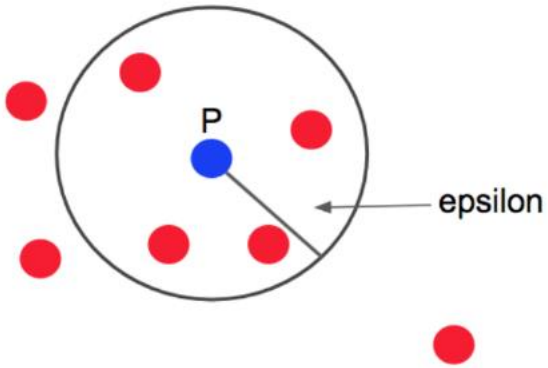
1. 거리측정: 단일 연결
   완전 연결
   평균 연결
   ward 거리

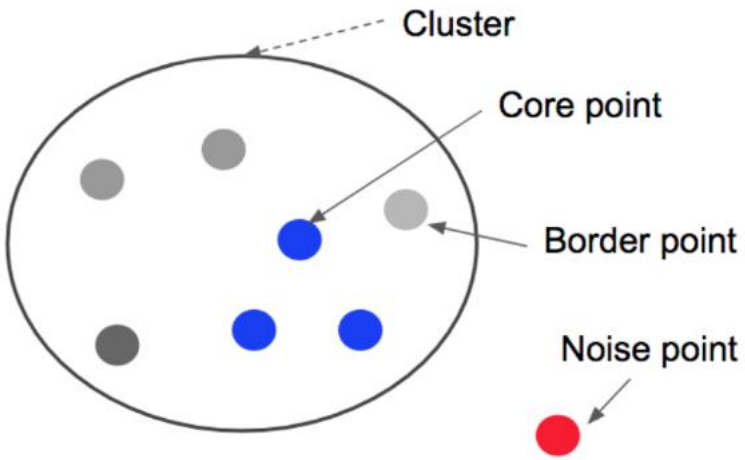
2. 위 식들 병합

3. Repeat

DBSCAN

점이 얼마나 몰려있는가. 이걸로 측정.





밀도 높은 지역 하나의 군집으로.

밀도 낮은 지역은 Noise로 간주.

```
model = DBSCAN(min_samples=6)
```

전과 동일하게 몇 개의 점을 기준으로 할지는 설정 가능.

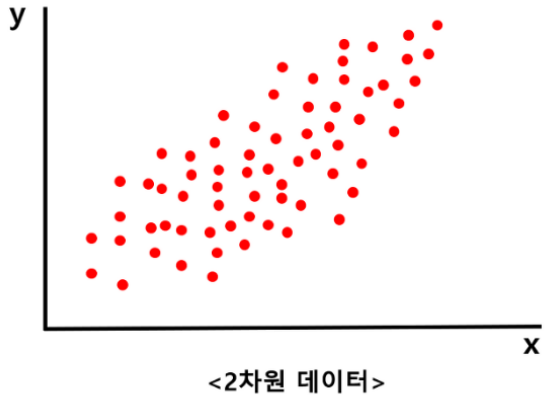
Dimensionality Reduction

고차원 데이터를 저차원으로 바꿔서 처리 효율 향상 시키는 방식.

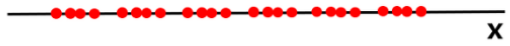
- PCA
- t-SNE

등등

PCA

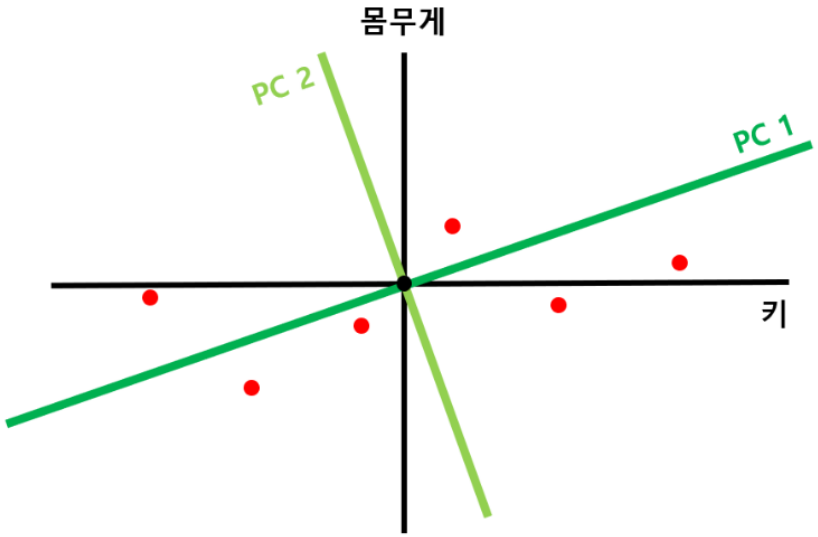


<2차원 데이터>

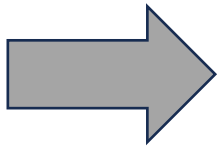


<1차원 데이터>

거리측정은
대부분 유클리디안 distance



1. 최적의 직선(거리 최소화 linear 그래프)
2. 수직이 되는 또 다른 직선(거리 최대화된 그래프)
3. 어느정도 맞다하면 PC2 제거.



결국 어느정도의 데이터
손실 감수 필요

t-SNE

PCA는 선형적 차원 축소, t-SNE는 비선형적 차원 축소 방식

고차원에서 사용.

써보면 기가막히게 잘나뉨 ㅋㅋ

쫌 심하게 고차원 ➔ 가우시안 distribution 사용

저차원 ➔ t-distribution 사용

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(t|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$