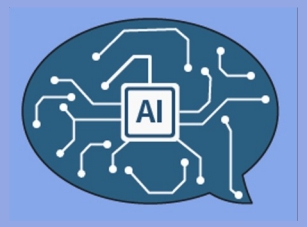


비지도학습

황현준

차원축소	주성분 분석, Feature Selection, Feature Extraction …
군집화	k-평균 군집, 계층 군집, …

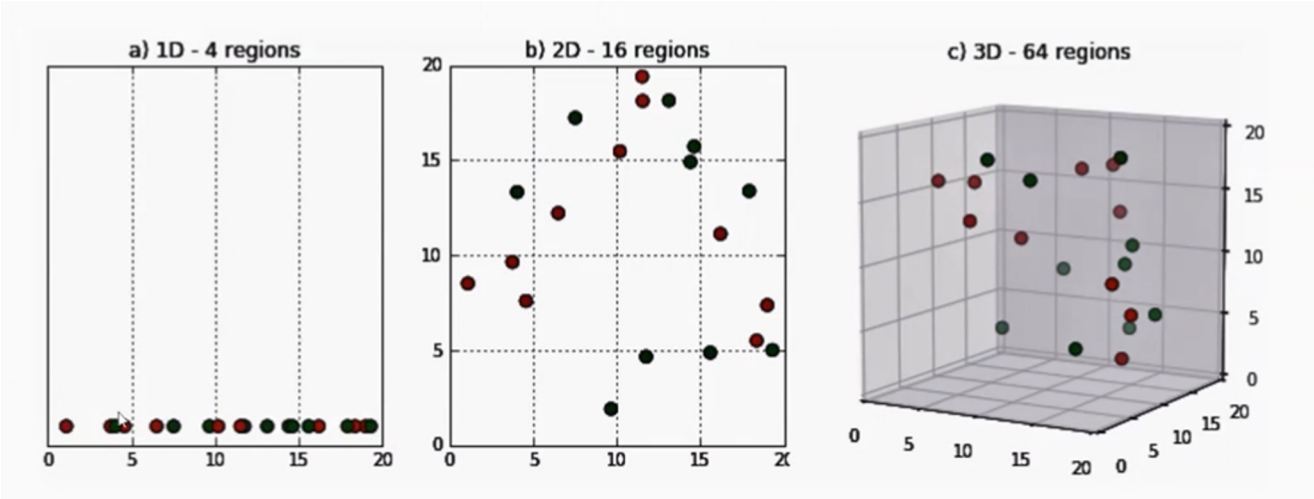
차원축소

OUTTA

차원 축소 (Dimensionality Reduction)

차원의 저주

- 피처의 종류 수가 학습 데이터 수 보다 더 많아지면서 성능이 저하되는 현상
- 차원에 존재하는 데이터들이 희박해지는(sparse) 현상
- 희박하다 = 빈 공간이 많다 = 정보가 없는 공간이 많다



차원 축소 (Dimensionality Reduction)

- 성능 좋은 인공지능, 머신 러닝을 위해선 많은 데이터가 필요

데이터가 많아질수록 계산 성능 / 저장 공간의 한계,

잡음 데이터 (noise) 존재 가능성 높아짐

데이터 전처리 단계에서 차원 축소를 많이 이용

차원 축소 (Dimensionality Reduction)

- 모델 학습에 불필요한 피처(속도 향상)나 방해되는 피처(성능 향상)를 제거

Feature Selection

Feature Extraction

Feature Selection

종속변수와 가장 관련성이 높은 피처만을 선택하고, 나머지를 제외시키는 것.
독립변수들(피처) 사이의 상관관계가 너무 높아서 한 쪽이 의미없는 경우 또한 제외

무게를 킬로그램과 파운드 두 개의 단위로 나타냈을 때, 두 피처는 서로 단위만 다르지 같은 특성을 말하고 있으므로 하나를 제외해도 무방함.
상관계수 히트맵(heatmap)으로 여러 피처의 공분산을 분석하면 피처 사이의 상관관계를 분석할 수 있다.

OUTTA

Feature Extraction

개별 피처를 제거하는 것이 아니라 저차원 공간으로 투영시켜 데이터와 모델을 단순화

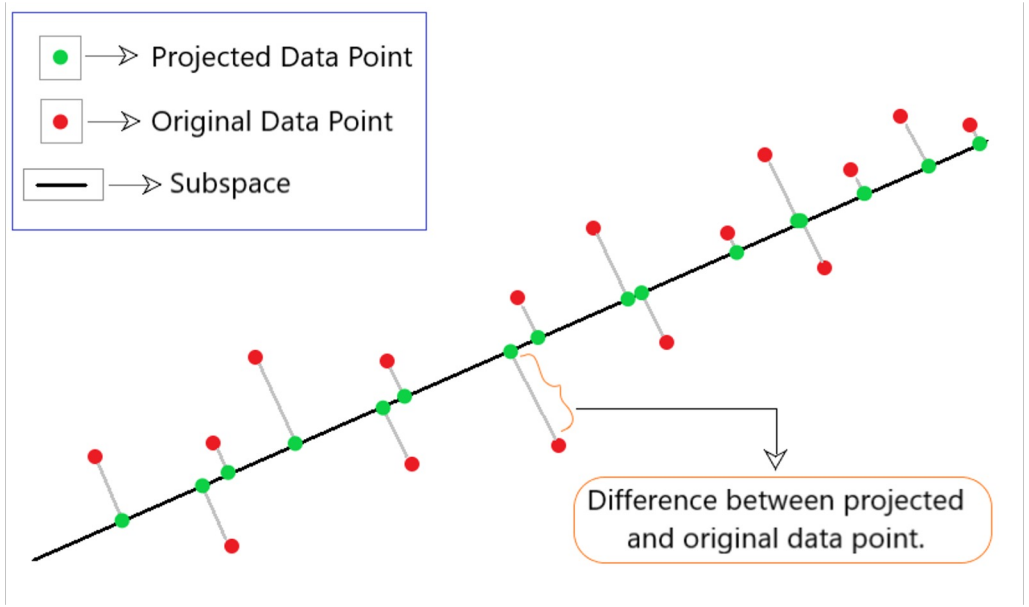
- 주성분분석(PCA) : Principal Component Analysis
- Kernel PCA

주성분 분석 (PCA) : Principal Component Analysis

- 전체 데이터의 분포를 가장 잘 설명할 수 있는 주성분을 찾아주는 방법론

데이터의 공분산 행렬의 고유값 분해를 통해 얻음.

- 고유값 분해는 대칭행렬에서 가능한데, 공분산 행렬은 항상 대칭행렬임
  - 주성분 방향벡터를 찾아 그 위로 투영(projection)시켜 차원을 한 단계 낮춤
  - 여기서 투영된 선(방향벡터)가 새로운 x축이 됨



공분산 행렬 (covariance matrix) ?

공분산(covariance): 두 변수의 상관관계를 나타내는 척도

양의 상관관계: 양수 // 음의 상관관계: 음수 // 상관관계 없음: 0

공분산의 정의는 다음과 같다.

정의 $\operatorname{Cov}(X, Y) \equiv \operatorname{E}[(X - \operatorname{E}[X])\,(Y - \operatorname{E}[Y])]$

covariance matrix: square matrix, symmetric matrix

고유값 분해 (eigen decomposition) ?

$$Av = \lambda v$$

양의 상관관계: 양수 // 음의 상관관계: 음수 // 상관관계 없음: 0

공분산의 정의는 다음과 같다.

정의 $\mathrm{Cov}(X, Y) \equiv \mathrm{E}[(X - \mathrm{E}[X])\,(Y - \mathrm{E}[Y])]$

covariance matrix: square matrix, symmetric matrix

군집화

군집화(Clustering)

데이터 샘플들을 별개의 군집 (Cluster) 으로 묶는 것
분류 알고리즘에 해당

전세계 음악을 장르/가수명 등등의 카테고리로 묶는 것이 해당함.

추정

모수적 (parametric) 추정

- 주어진 데이터가 특정 분포를 따른다고 가정
- Gausian Mixture Model (GMM): 데이터가 정규분포를 따른다고 가정

비모수적 (non-parametric) 추정

- 데이터가 특정 분포를 따르지 않는다는 가정 하에서 확률 밀도 추정
- K-means, Mean Shift, DBSCAN
- K-means Clustering
  - 중심점(centroid)을 기반으로 클러스터링 진행
  - 몇 개의 군집으로 나눌 것인지 하이퍼 파라미터로 제공 필요.

계층적 군집화 (Hierarchical Clustering)

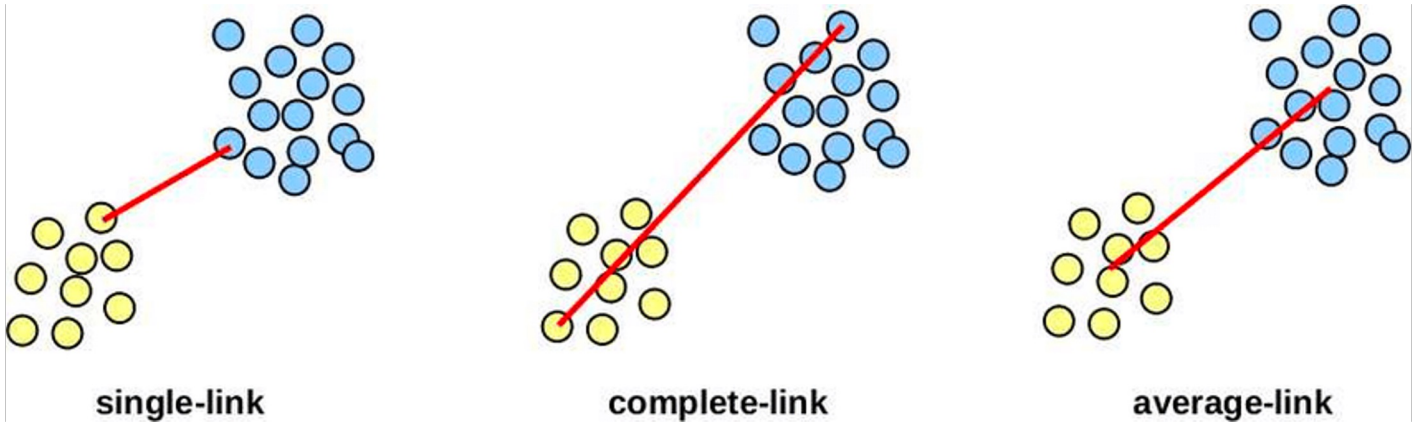
1. Divisive (top-down approach) 분할계층군집 : 전체 샘플을 포함하는 하나의 클러스터에서, 클러스터 안에 샘플이 하나만 남을 때 까지 클러스터를 작게 나누는 방법
2. Agglomerative (bottom-up approach) 병합계층군집 : 각 샘플이 하나의 독립적인 클러스터가 되고, 하나의 클러스터가 될 때 까지 가장 가까운 클러스터를 합침

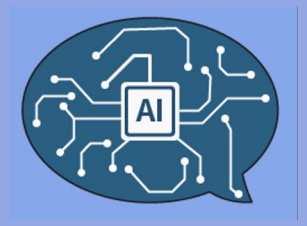
- 군집 간 거리 계산을 통해 클러스터를 나누거나 합침
- 사전에 클러스터 수를 정하지 않아도 학습 가능, 덴드로그램(dendrogram)으로 시각화 가능

- min distance(single link): 군집간 가까운 원소끼리의 거리
- max distance(complete link): 군집간 가장 먼 원소끼리의 거리
- average distance(average link): 군집간 원소끼리 거리의 평균
- centroid distance: 군집 중심간 거리

Single / Complete

- 단일 연결 (single linkage) : 클러스터 간 샘플 중 가장 비슷한 (거리가 가까운) 샘플의 거리를 계산하여 병합
- 완전 연결 (complete linkage) : 클러스터 간 샘플 중 가장 비슷하지 않은 (거리가 먼) 샘플의 거리를 계산하여 병합





감사합니다