Text Analytics

Ch6 : Dimensionality Reduction

서수원

Business Intelligence Lab.

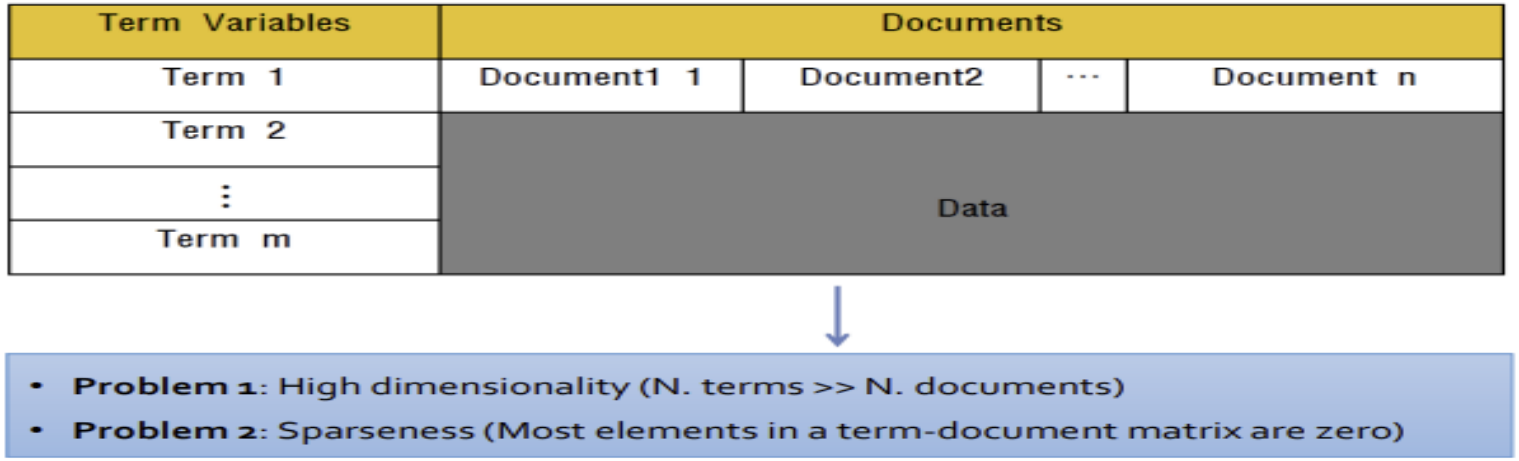
산업경영공학과, 명지대학교

01

Dimensionality Reduction

Dimensionality Reduction

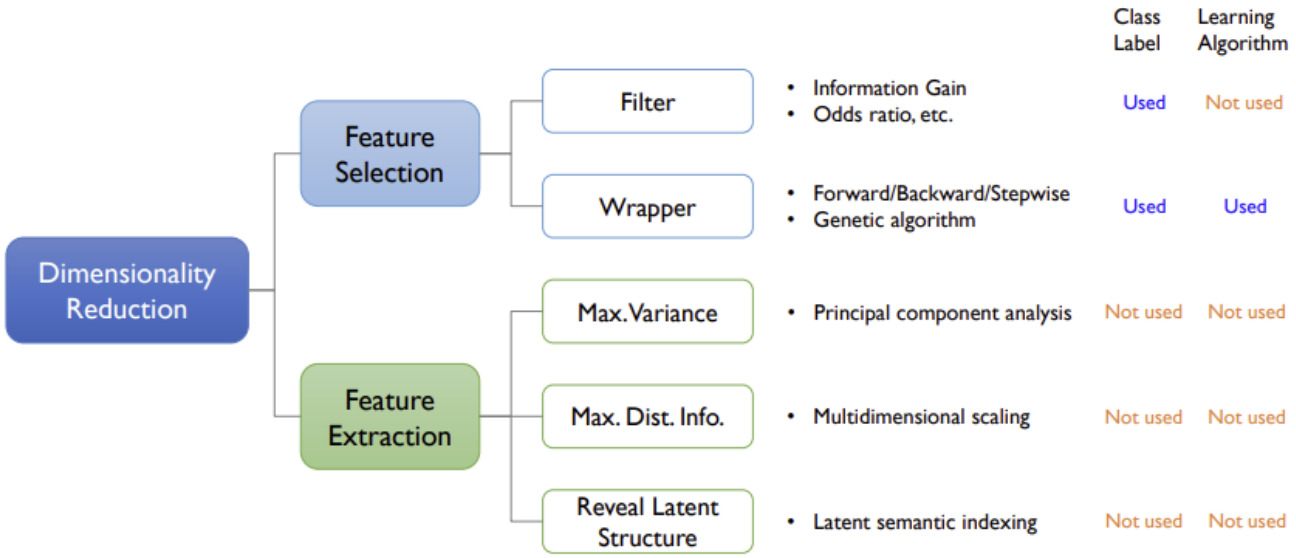
- 텍스트 데이터의 일반적인 특징(Bag of Words 를 가정)
  - 단어의 수(Terms)가 문서의 수(Documents)보다 많다.
    - ✓ 변수가 관측치 보다 많으면 기존 통계의 방법이 사용 불가능 하다.
      - ❖ 과적합의 문제가 발생하기 때문이다.
    - ✓ 희소성의 문제가 발생한다.



- 왜 dimensionality reductio이 필요할까?
  - 계산 효율성을 높이기 위함이다.
  - 텍스트 마이닝의 정확도를 높이기 위함이다.

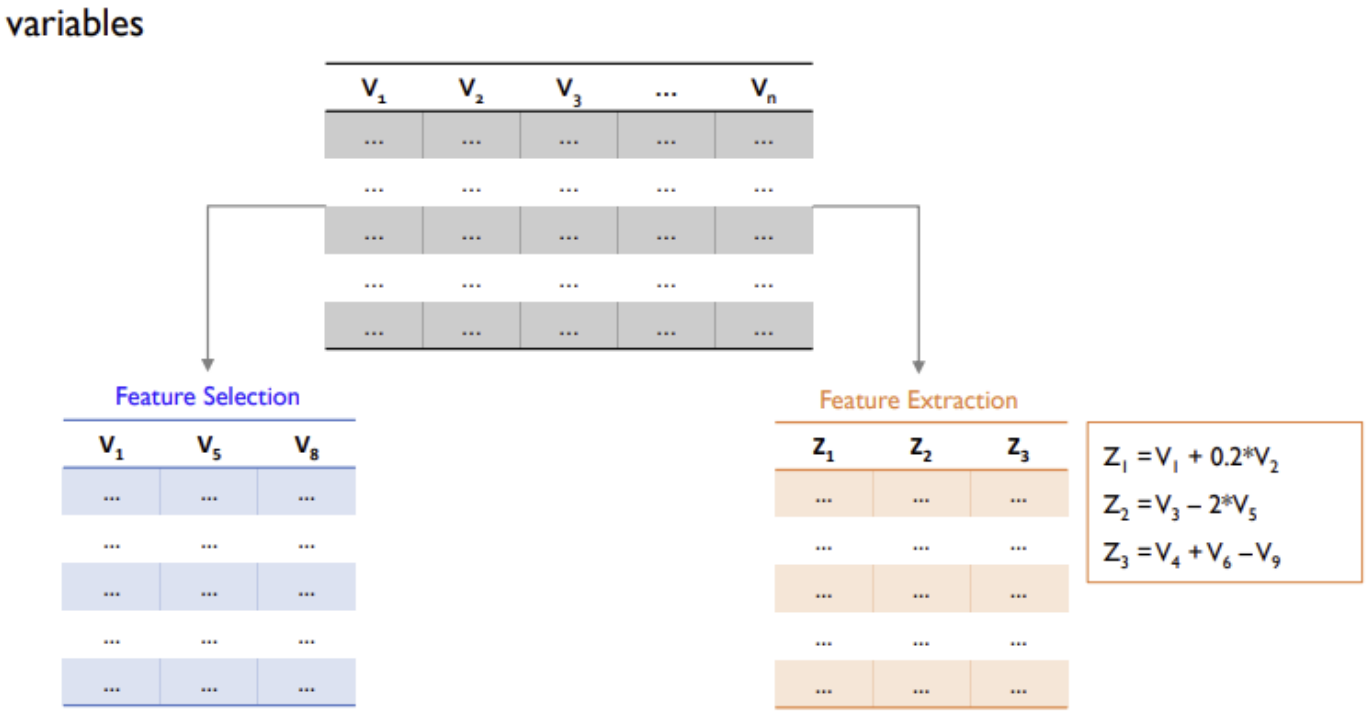
Dimensionality Reduction

- Dimension Reduction의 테크닉
  - 크게 보면 Feature Selection과 Feature Extraction이 있다.
    - ✓ 둘의 차이점은 차원을 단순히 축소하느냐, 데이터셋의 특징을 가지고 새로운 셋을 만드느냐에 따라 차이가 있다.
  - Feature Selection
    - ✓ 지도학습이다.
    - ✓ 알고리즘 사용 유무에 따라 Filter와 Wrapper가 있다.
  - Feature Extraction
    - ✓ 비지도 학습이다.



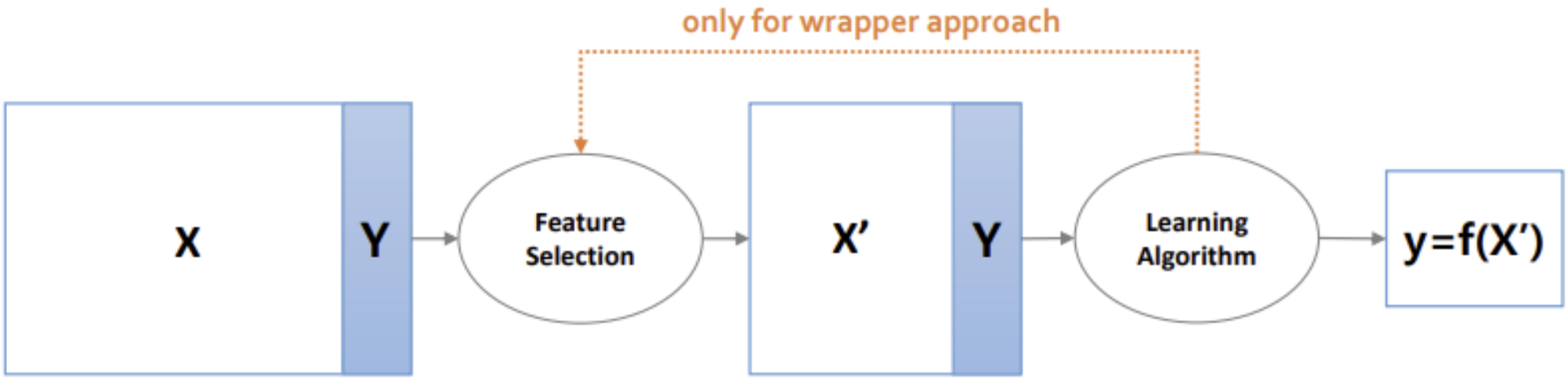
Dimensionality Reduction

- Dimension Reduction의 테크닉
  - 크게 보면 Feature Selection과 Feature Extraction이 있다.
    - ✓ 둘의 차이점은 차원을 단순히 축소하느냐, 데이터셋의 특징을 가지고 새로운 셋을 만드느냐에 따라 차이가 있다.



Dimensionality Reduction

- Dimension Reduction의 테크닉
  - Feature Selection
    - ✓ 지도학습이다.
    - ✓ 알고리즘 사용 유무에 따라 Filter와 Wrapper가 있다.



02

Feature Selection

Feature Selection

- 10개의 문서와 10개의 단어
  - 긍부정 판별 문제이다.
  - 6개의 긍정 문서와 4개의 부정 문서가 있다.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1
Term 4	1	1	1	1	1	1	1	1	0	0
Term 5	0	0	0	1	1	1	1	1	1	1
Term 6	1	1	1	0	0	0	0	0	0	0
Term 7	0	0	0	0	0	0	1	1	0	0
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0
Term 10	1	0	0	0	0	0	0	0	1	1
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg

Feature Selection

- Feature Selection Metric
  - Document frequency(DF)
    - ✓ 단어가 나타나는 횟수를 센다.
      - Term1 : DF(w) = 6
      - Term2 : DF(w) = 4
      - Term3 : DF(w) = 10
    - ✓ Accuracy(Acc)
      - 정확도를 본다. $Acc(w) = N(Pos, w) - N(Neg, w)$
        - » Term1 : N(Pos,w) = 6,N(Neg,w) = 0, ACC(w) = 6
        - » Term2 : N(Pos,w) = 0,N(Neg,w) = -4, ACC(w) = -4
        - » Term3 : N(Pos,w) = 6,N(Neg,w) = 4, ACC(w) = 2

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1
Term 4	1	1	1	1	1	1	1	1	0	0
Term 5	0	0	0	1	1	1	1	1	1	1
Term 6	1	1	1	0	0	0	0	0	0	0
Term 7	0	0	0	0	0	0	1	1	0	0
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0
Term 10	1	0	0	0	0	0	0	0	1	1
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg

Feature Selection

- Feature Selection Metric
  - Accuracy ration(AccR)

$$AccR(w) = \left| \frac{N(Pos,w)}{N(Pos)} - \frac{N(Neg,w)}{N(Neg)} \right|$$

  - For Term 1: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{0}{4} = 0, AccR(w) = 1$
  - For Term 2: $\frac{N(Pos,w)}{N(Pos)} = \frac{0}{6} = 0, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 1$
  - For Term 3: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 0$

  ✓ 성능이 나타나는 것을 확인할 수 있다.

  - Probability Ratio(PR)

  ✓ P(Pos,w|Pos)/P(Neg,w|Neg)

  - For Term 1: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{0}{4} = 0, PR(w) = \infty$
  - For Term 2: $\frac{N(Pos,w)}{N(Pos)} = \frac{0}{6} = 0, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 0$
  - For Term 3: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 1$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1
Term 4	1	1	1	1	1	1	1	1	0	0
Term 5	0	0	0	1	1	1	1	1	1	1
Term 6	1	1	1	0	0	0	0	0	0	0
Term 7	0	0	0	0	0	0	1	1	0	0
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0
Term 10	1	0	0	0	0	0	0	0	1	1
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg

Feature Selection

- Feature Selection Metric

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	DF	Acc	AccR	PR
Term 1	1	1	1	1	1	1	0	0	0	0	6	6	1.00	Inf
Term 2	0	0	0	0	0	0	1	1	1	1	4	-4	1.00	0.00
Term 3	1	1	1	1	1	1	1	1	1	1	10	2	0.00	1.00
Term 4	1	1	1	1	1	1	1	1	0	0	8	4	0.50	2.00
Term 5	0	0	0	1	1	1	1	1	1	1	7	-1	0.50	0.50
Term 6	1	1	1	0	0	0	0	0	0	0	3	3	0.50	Inf
Term 7	0	0	0	0	0	0	1	1	0	0	2	-2	0.50	0.00
Term 8	1	0	1	0	1	0	1	0	1	0	5	1	0.00	1.00
Term 9	1	1	1	0	0	0	1	0	0	0	4	2	0.25	2.00
Term 10	1	0	0	0	0	0	0	0	1	1	3	-1	0.33	0.33
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg				

Feature Selection

- Feature Selection Metric
  - Odds ratio(OddR)
    - ✓ 동일 집단 내에서 어떤 사건이 발생할 확률과 발생하지 않을 확률의 비교값이다.
      - ❖ 예시로 흡연자가 폐암에 걸릴 확률이 20%, 비흡연자가 폐암에 걸릴 확률이 1%라고 하면 OddR은 20/80,1/99이다.
    - For Term 8: $\frac{N(Pos,w)}{N(Neg,w)} = \frac{3}{2}, \frac{N(Neg,\bar{w})}{N(Pos,\bar{w})} = \frac{2}{3}, OddR(w) = 1$
    - For Term 9: $\frac{N(Pos,w)}{N(Neg,w)} = \frac{3}{1}, \frac{N(Neg,\bar{w})}{N(Pos,\bar{w})} = \frac{3}{3}, OddR(w) = 3$
    - ✓ Odds ratio Numerator(OddN)
      - ❖ 분자만 가져왔다.
    - For Term 8: $N(Pos,w) = 3, N(Neg,\bar{w}) = 2, OddN(w) = 6$
    - For Term 9: $N(Pos,w) = 3, N(Neg,\bar{w}) = 3, OddN(w) = 9$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1
Term 4	1	1	1	1	1	1	1	1	0	0
Term 5	0	0	0	1	1	1	1	1	1	1
Term 6	1	1	1	0	0	0	0	0	0	0
Term 7	0	0	0	0	0	0	1	1	0	0
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0
Term 10	1	0	0	0	0	0	0	0	1	1
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg

Feature Selection

- Feature Selection Metric
  - F1 – Measure

$$F1(w) = \frac{2 \times Recall(w) \times Precision(w)}{Recall(w) + Precision(w)}$$

$$Recall(w) = \frac{N(Pos,w)}{N(Pos,w) + N(Pos,\bar{w})}, \qquad Precision(w) = \frac{N(Pos,w)}{N(Pos,w) + N(Neg,w)}$$

$$F1(w) = \frac{2 \times N(Pos,w)}{N(Pos) + N(w)}$$

- For Term 1: $F1(w) = \frac{2 \times 6}{6+6} = 1$
- For Term 2: $F1(w) = \frac{2 \times 0}{6+4} = 0$
- For Term 3: $F1(w) = \frac{2 \times 6}{6+10} = 0.75$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1
Term 4	1	1	1	1	1	1	1	1	0	0
Term 5	0	0	0	1	1	1	1	1	1	1
Term 6	1	1	1	0	0	0	0	0	0	0
Term 7	0	0	0	0	0	0	1	1	0	0
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0
Term 10	1	0	0	0	0	0	0	0	1	1
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg

Feature Selection

- Feature Selection Metric

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	OddR	OddN	FI
Term 1	1	1	1	1	1	1	0	0	0	0	24.00	24	1.00
Term 2	0	0	0	0	0	0	1	1	1	1	0.00	0	0.00
Term 3	1	1	1	1	1	1	1	1	1	1	0.00	0	0.75
Term 4	1	1	1	1	1	1	1	1	0	0	4.00	12	0.86
Term 5	0	0	0	1	1	1	1	1	1	1	0.00	0	0.46
Term 6	1	1	1	0	0	0	0	0	0	0	4.00	12	0.67
Term 7	0	0	0	0	0	0	1	1	0	0	0.00	0	0.00
Term 8	1	0	1	0	1	0	1	0	1	0	1.00	6	0.55
Term 9	1	1	1	0	0	0	1	0	0	0	3.00	9	0.60
Term 10	1	0	0	0	0	0	0	0	1	1	0.20	2	0.22
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg			

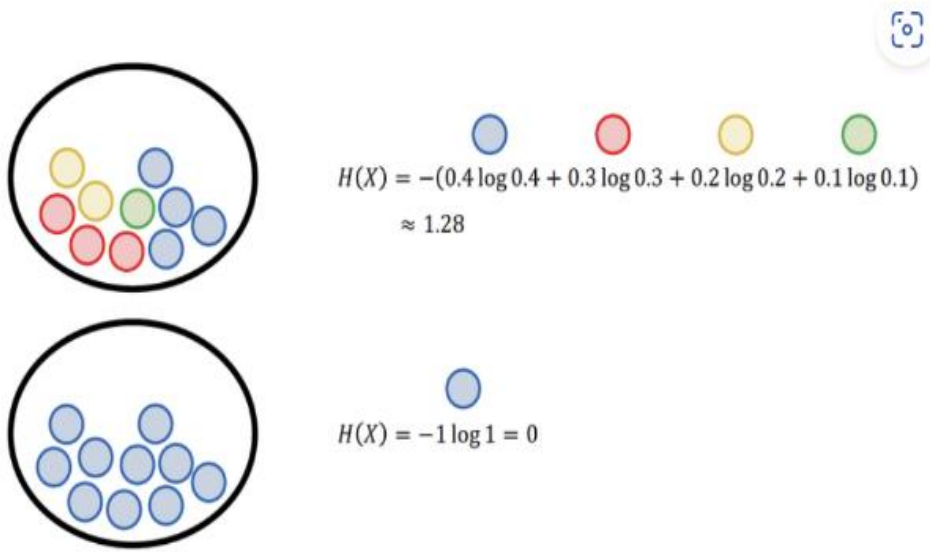
Feature Selection

- Feature Selection Metric
  - Information Gain(IG)
    - ✓ 엔트로피 감소량을 측정한다.

$$Entropy(absent\ w) = \sum_{C \in \{Pos,Neg\}} -P(C) \times \log(P(C))$$

$$Entropy(given\ w) = P(w)\left[\sum_{C \in \{Pos,Neg\}} -P(C|w) \times \log(P(C|w))\right] + P(\overline{w})\left[\sum_{C \in \{Pos,Neg\}} -P(C|\overline{w}) \times \log(P(C|(\overline{w}))\right]$$

$$IG(w) = Entropy(absent\ w) - Entropy(given\ w)$$



Feature Selection

- Feature Selection Metric
  - Information Gain(IG)
    - ✓ 엔트로피 감소량을 측정한다.

$$Entropy(absent\ w) = -P(Pos) \times \log(P(Pos)) - P(Neg) \times \log(P(Neg))$$
$$= -0.6 \times \log(0.6) - 0.4 \times \log(0.4)$$
$$= 0.29$$

$$Entropy(given\ w) = P(w)[-P(Pos|w) \times \log(P(Pos|w)) - P(Neg|w) \times \log(P(Neg|w))]$$
$$+ P(\bar{w})[-P(Pos|\bar{w}) \times \log(P(Pos|\bar{w})) - P(Neg|\bar{w}) \times \log(P(Neg|\bar{w}))]$$
$$= 0.6[-1 \times \log(1) - 0 \times \log(0)] + 0.4[-0 \times \log(0) - 1 \times \log(1)]$$
$$= 0$$

Convert log(o) to zero

$$IG(w) = 0.29 - 0 = 0.29$$

Feature Selection

- Feature Selection Metric
  - 카이제곱 통계량
    - ✓ 클래스와 독립적이다.

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

k : 범주의 수, O_i: 실제 도수, E_i: 기대 도수

. 데이터의 분포와 사용자가 선택한 기대값 또는 가정된 분포 사이의 차를 나타낸 측정값
. (기대도수) = (열의 합계) × (행의 합계) / (전체 합계)

Term 1	Pos	Neg	Total
w	6	0	6
$\bar{w}$	0	4	4
total	6	4	10

Term 4	Pos	Neg	Total
w	6	2	8
$\bar{w}$	0	2	2
total	6	4	10

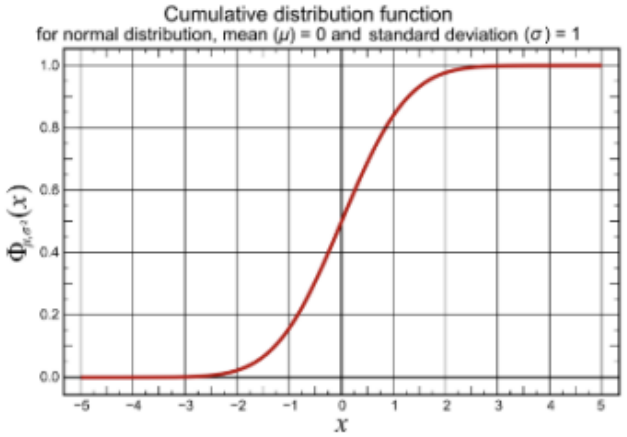
$$\chi^2(T1) = \frac{10 \times [0.6 \times 0.4 - 0 \times 0]^2}{0.6 \times 0.4 \times 0.6 \times 0.4} = 10.00 \qquad \chi^2(T4) = \frac{10 \times [0.6 \times 0.2 - 0.2 \times 0]^2}{0.8 \times 0.2 \times 0.6 \times 0.4} = 3.75$$

Feature Selection

- Feature Selection Metric
  - Bi-Normal Separation(BNS)
    - ✓ 정규분포를 가정하고, 누적 분포함수를 따른다고 한다.

$$BNS(w) = \left| F^{-1}\left(\frac{N(Pos,w)}{N(Pos)}\right) - F^{-1}\left(\frac{N(Neg,w)}{N(Neg)}\right) \right|$$

F: c.d.f of the standard normal distribution



Term 4	Pos	Neg	Total
w	6	2	8
$\bar{w}$	0	2	2
total	6	4	10

$$BNS(w) = |F^{-1}(1) - F^{-1}(0.5)|$$
$$\approx |F^{-1}(0.9995) - F^{-1}(0.5)|$$
$$= |3.29 - 0| = 3.29$$

Feature Selection

- Feature Selection Metric
  - IG랑 x^2으로 충분하다.

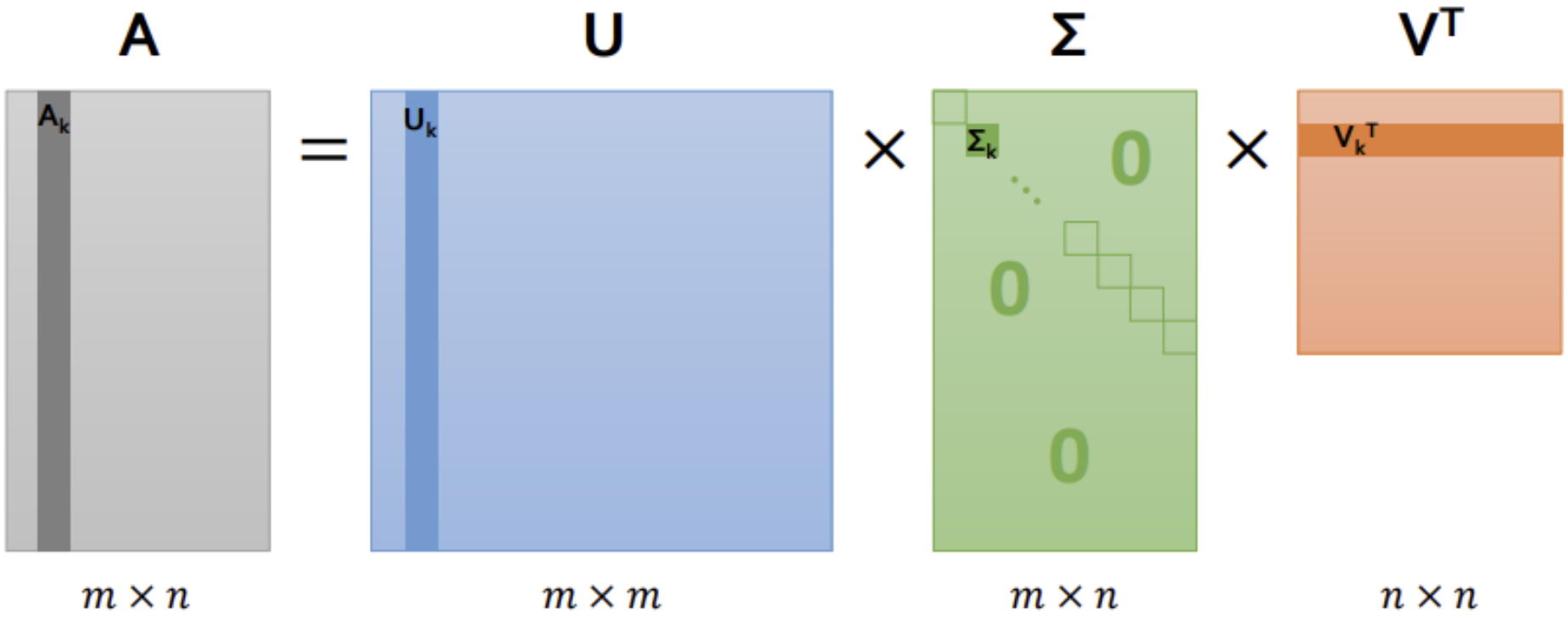
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	IG	χ^2	BNS
Term 1	1	1	1	1	1	1	0	0	0	0	0.29	10.00	6.58
Term 2	0	0	0	0	0	0	1	1	1	1	0.29	10.00	6.58
Term 3	1	1	1	1	1	1	1	1	1	1	0.00	0.00	0.00
Term 4	1	1	1	1	1	1	1	1	0	0	0.10	3.75	3.29
Term 5	0	0	0	1	1	1	1	1	1	1	0.08	2.86	3.29
Term 6	1	1	1	0	0	0	0	0	0	0	0.08	2.86	3.29
Term 7	0	0	0	0	0	0	1	1	0	0	0.10	3.75	3.29
Term 8	1	0	1	0	1	0	1	0	1	0	0.00	0.00	0.00
Term 9	1	1	1	0	0	0	1	0	0	0	0.01	0.63	0.67
Term 10	1	0	0	0	0	0	0	0	1	1	0.03	1.27	0.97
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg			

03

Feature Extraction

Feature Extraction

- Feature Extraction
  - SVD
    - ✓ 특이값 분해이다.
    - ✓ U_k T U_k = 1, U_i T U_j = 0



Feature Extraction

- Feature Extraction
  - SVD의 장점
    - ✓ 하나의 차원에서 직교 하는 것은 항상 항등원이다.

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$$

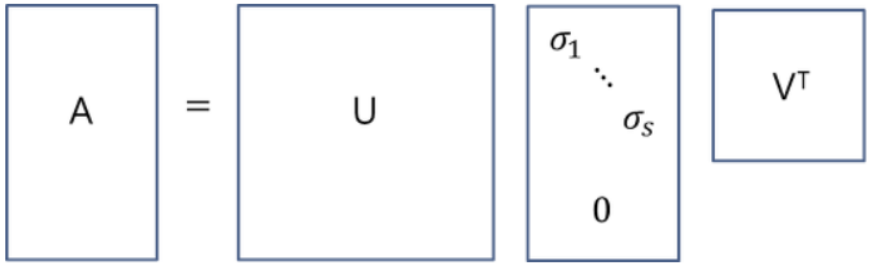
    - ✓ Rank(A)는 0이 아닌 최대의 차원수를 의미한다. zero singular value(시그마 벡터에서)

Feature Extraction

- Feature Extraction
  - Reduces SVDs

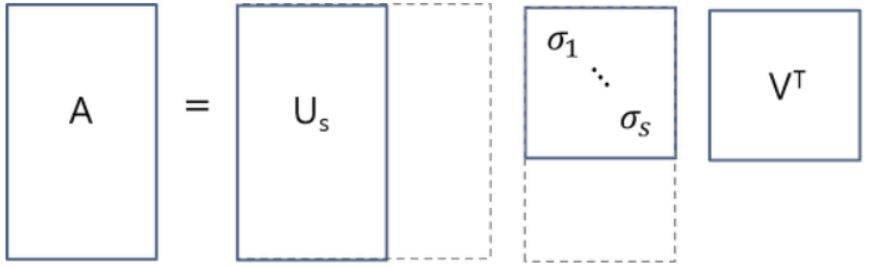
1) full SVD

위에서 설명된 내용들은 모두 full SVD이며 A행렬이 SVD를 진행하였을때 얻는 그대로의 과정을 의미합니다.



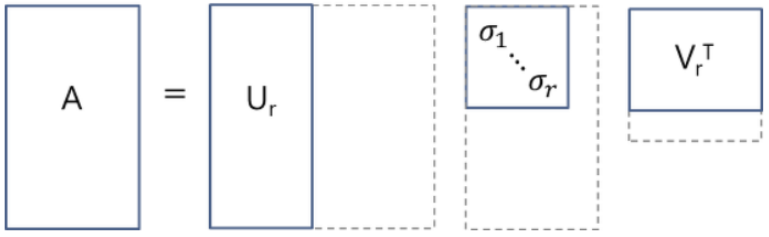
2) Thin SVD

Thin SVD는 Σ 행렬에서 대각 원소가 아닌 0으로 구성된 부분이 제거되었으며 이에 따라서 U에서 제거된 부분과 대응되는 열 벡터들이 제거된 U_s 로 이뤄지는 SVD를 의미합니다.



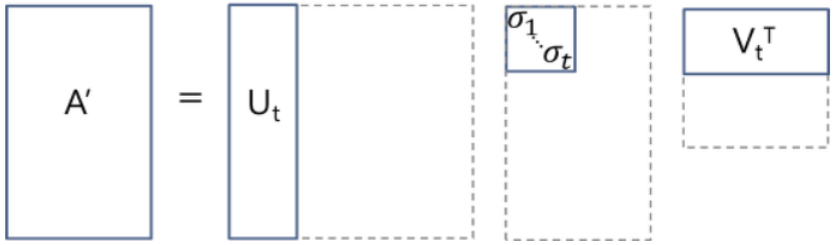
3) Compact SVD

Compact SVD는 대각선에 위치하지 않은 원소들을 제거할 뿐만 아니라 0인 singular value까지 제거된 SVD를 의미합니다.



4) Truncated SVD

Truncated SVD는 Σ 행렬의 대각원소 가운데 상위 t개만 골라낸 형태입니다. 해당 방법은 행렬 A를 원복하지 못하게 되지만 데이터를 상당히 압축해도 행렬 A에 근사할 수 있는 장점이 있습니다.



Feature Extraction

- Feature Extraction
  - Reduces SVDs

✓ SVD decomposition

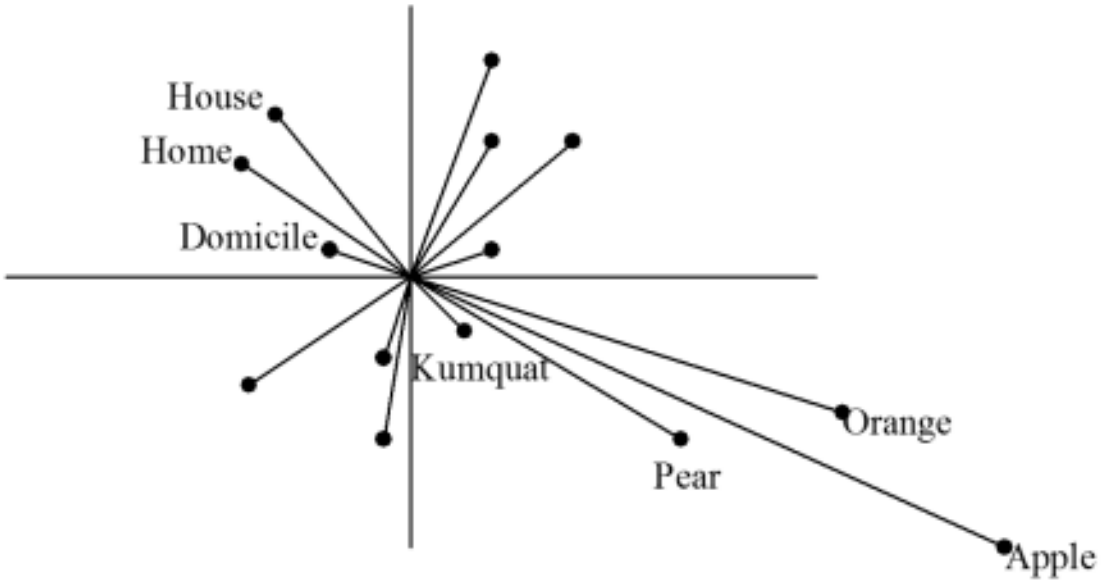
$$A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad U = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad S = \begin{bmatrix} 5.47 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad V = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

✓ Truncated SVD

$$A' = U_1 \times S_1 \times V_1^T = \begin{bmatrix} 0.82 \\ 0.58 \\ 0 \\ 0 \end{bmatrix} \times [5.47] \times \begin{bmatrix} 0.40 & 0.91 \end{bmatrix} = \begin{bmatrix} 1.79 & 4.08 \\ 1.27 & 2.89 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Feature Extraction

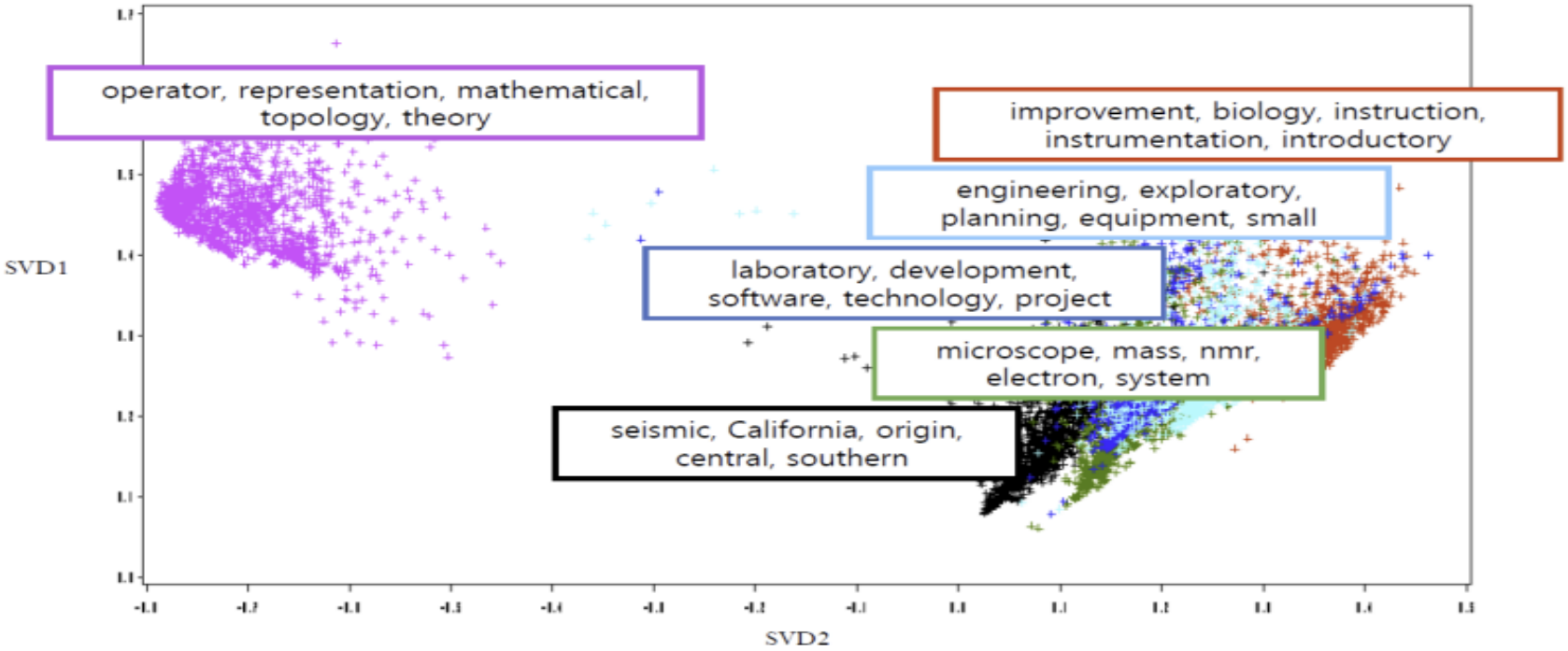
- Feature Extraction
  - Reduces SVDs
    - ✓ 불분명한 관계도 찾을 수 있다.
    - ✓ 차원을 축소 해도 거리 관계가 보존된다.



Feature Extraction

- Feature Extraction
  - Reduces SVDs
    - ✓ 불분명한 관계도 찾을 수 있다.
    - ✓ 차원을 축소 해도 거리 관계가 보존된다.

✓ Visualize the project in the reduced 2-D space



Feature Extraction

- Stochastic Neighbor Embedding
  - 차원을 줄이더라도 데이터 간의 거리(이웃)는 유사해야 한다.
  - 확률적 결정을 한다.
  - 데이터 간의 거리는 유클리드 거리를 이용해 정의한다.

$$D_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|$$

  - 데이터의 분산 정도에 따라, 달라짐으로 정규화를 해야한다.

$$d_{ij}^2 = \frac{\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2}$$

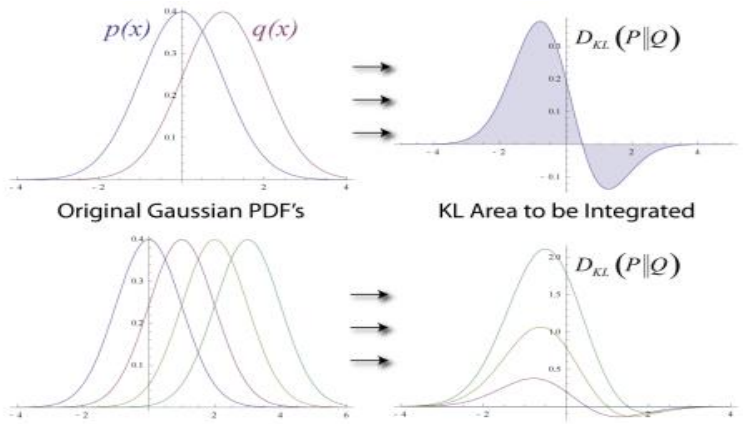
    - ✓ 표준편차가 크면, 멀리 있어도 이웃으로 정의 할 확률이 커진다.
      - ❖ 표준편차가 크면, 엔트로피 값이 커진다. 작다면 엔트로피 값이 낮아진다.
    - ✓ q는 우리가 찾아야 하는 것을 의미한다.(축소된 공간)

$$p_{j|i} = \frac{e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}}}{\sum_k e^{-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\sigma_i^2}}} \qquad q_{j|i} = \frac{e^{-||\mathbf{y}_i - \mathbf{y}_j||^2}}{\sum_k e^{-||\mathbf{y}_i - \mathbf{y}_k||^2}}$$

Feature Extraction

- Stochastic Neighbor Embedding
  - q의 y값을 찾는것이 목표이다.
  - 쿨백 라이블러 발산(KL Divergence)를 이용한다.
    - ✓ 비용함수를 밑과 같이 정의한다.

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$



✓ Gradient

$$\frac{\partial C}{\partial \mathbf{y}_i} = 2 \sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$$

Feature Extraction

- Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다.

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

$$C = \sum_i \sum_j p_{j|i} log p_{j|i} - \sum_i \sum_j p_{j|i} log q_{j|i}$$

$$C' = -\sum_i \sum_j p_{j|i} log q_{j|i} \qquad \left(\frac{\partial C}{\partial y_t} = \frac{\partial C'}{\partial y_t}\right)$$

$$C' = \underbrace{-\sum_i p_{t|i} log q_{t|i}}_{①} \underbrace{-\sum_j p_{j|t} log q_{j|t}}_{②} \underbrace{-\sum_{i \neq t} \sum_{j \neq t} p_{i|j} log q_{i|j}}_{③}$$

Feature Extraction

- Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다.

$$p_{j|i} = \frac{e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}}}{\sum_k e^{-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\sigma_i^2}}} \qquad q_{j|i} = \frac{e^{-||\mathbf{y}_i - \mathbf{y}_j||^2}}{\sum_k e^{-||\mathbf{y}_i - \mathbf{y}_k||^2}}$$

$$d_{ti} = exp(-||\mathbf{y}_t - \mathbf{y}_i||^2) = d_{it}$$

$$\frac{\partial d_{ti}}{\partial \mathbf{y}_t} = d'_{ti} = -2(\mathbf{y}_t - \mathbf{y}_i)exp(-||\mathbf{y}_t - \mathbf{y}_i||^2) = -2(\mathbf{y}_t - \mathbf{y}_i)d_{ti}$$

$$q_{t|i} = \frac{exp(-||\mathbf{y}_i - \mathbf{y}_t||^2)}{\sum_{k \neq i} exp(-||\mathbf{y}_i - \mathbf{y}_k||^2)} = \frac{d_{it}}{\sum_{k \neq i} d_{ik}}$$

$$q_{j|t} = \frac{exp(-||\mathbf{y}_t - \mathbf{y}_j||^2)}{\sum_{k \neq t} exp(-||\mathbf{y}_t - \mathbf{y}_k||^2)} = \frac{d_{tj}}{\sum_{k \neq t} d_{tk}}$$

$$q_{i|j} = \frac{exp(-||\mathbf{y}_j - \mathbf{y}_i||^2)}{\sum_{k \neq j} exp(-||\mathbf{y}_j - \mathbf{y}_k||^2)} = \frac{d_{ji}}{\sum_{k \neq j} d_{jk}}$$

Feature Extraction

- Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다. $= -\sum_i p_{t|i} log q_{t|i}$

• Gradient of the cost function ① (Optional)

$$\frac{\partial}{\partial y_t}\Big(-\sum_i p_{t|i} log q_{t|i}\Big) = -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \frac{\partial q_{t|i}}{\partial y_t}$$

$$= -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \frac{d'_{it} \cdot (\sum_{k \neq i} d_{ik}) - d_{it} \cdot d'_{it}}{(\sum_{k \neq i} d_{ik})^2}$$

$$= -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \frac{-2(\mathbf{y}_t - \mathbf{y}_i) \cdot d_{it} \cdot (\sum_{k \neq i} d_{ik}) + 2(\mathbf{y}_t - \mathbf{y}_i) \cdot d_{it}^2}{(\sum_{k \neq i} d_{ik})^2}$$

$$= -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \Big(-2(\mathbf{y}_t - \mathbf{y}_i) \cdot q_{t|i} + 2(\mathbf{y}_t - \mathbf{y}_i) \cdot q_{t|i}^2\Big)$$

$$= \sum_i p_{t|i} \cdot 2(\mathbf{y}_t - \mathbf{y}_i)(1 - q_{t|i})$$

①

$$\frac{\partial d_{ti}}{\partial \mathbf{y}_t} = d'_{ti} = -2(\mathbf{y}_t - \mathbf{y}_i) exp(-||\mathbf{y}_t - \mathbf{y}_i||^2) = -2(\mathbf{y}_t - \mathbf{y}_i) d_{ti}$$

$$d_{ti} = exp(-||\mathbf{y}_t - \mathbf{y}_i||^2) = d_{it}$$

$$q_{t|i} = \frac{exp(-||\mathbf{y}_i - \mathbf{y}_t||^2)}{\sum_{k \neq i} exp(-||\mathbf{y}_i - \mathbf{y}_k||^2)} = \frac{d_{it}}{\sum_{k \neq i} d_{ik}}$$

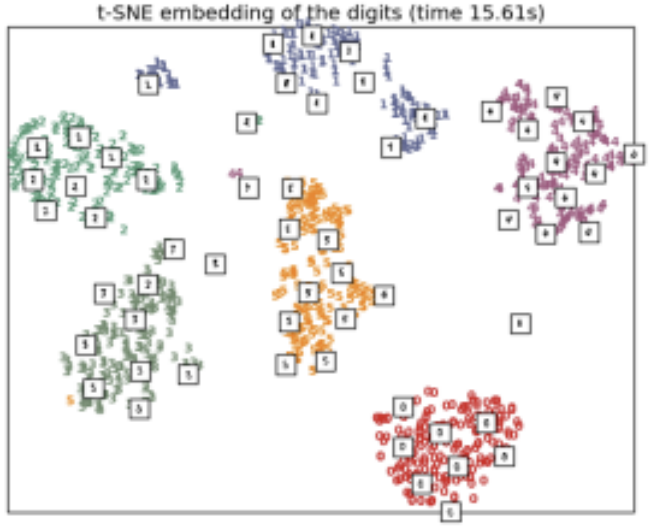
Feature Extraction

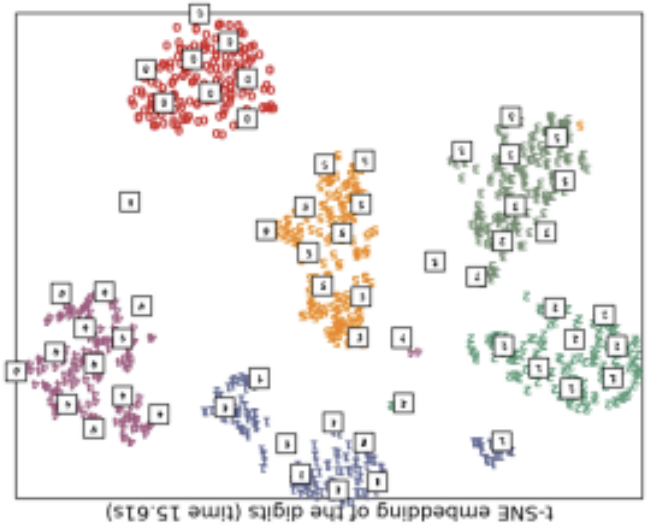
- Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다.

$$\frac{\partial C}{\partial \mathbf{y}_t} = 2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{t|j} - q_{t|j} + p_{j|t} - q_{j|t})$$

✓ Gradient

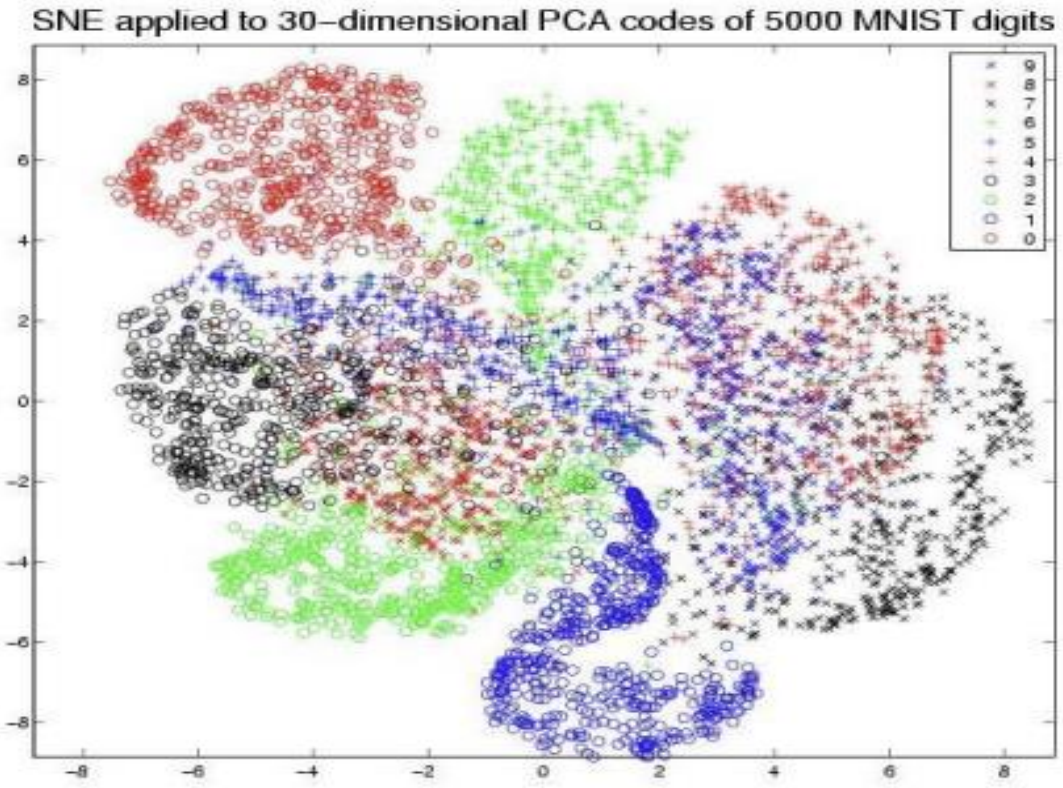
$$\frac{\partial C}{\partial \mathbf{y}_i} = 2\sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$$





Feature Extraction

- Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다.
    - ✓ 거리가 너무 가까워진다.



Feature Extraction

- T – Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다.
    - ✓ 거리가 너무 가까워진다.
      - ❖ 해결하기 위해 정규분포 대신 t분포를 사용한다.

✓ Gradient:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{ij} - q_{ij})(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}$$

Feature Extraction

- T – Stochastic Neighbor Embedding
  - 비용함수를 최소로 하는 yi값을 찾아야 한다.
    - ✓ 거리가 너무 가까워진다.
      - ❖ 해결하기 위해 정규분포 대신 t분포를 사용한다.