인공지능 4주차 2차시

1 가중 회귀분석

가중 회귀분석

- 단순한 최소 회귀법은 특이값에 취약
- 취약함을 보완하기 위해 가중치를 변화시켜 유연성을 높임
  - LOWESS(Locally Weighted Scatterplot Smoothing)
    - 이변량 자료를 작은 윈도우로 나누어 각각의 구간에서 가중선형회귀(Weighted Linear Regression)를 하여 곡선을 구함
  - L2 정규화
  - L1 정규화

1 최소 제곱법 수정

최소제곱법

- 실제 답과 결과 값의 오차의 제곱의 합이 최소가 되는 해를 구하는 방법

알고 싶은 값 y_i,

예측 방정식 $y = \beta_0 - \beta_1 X$ 로 설정한 경우,

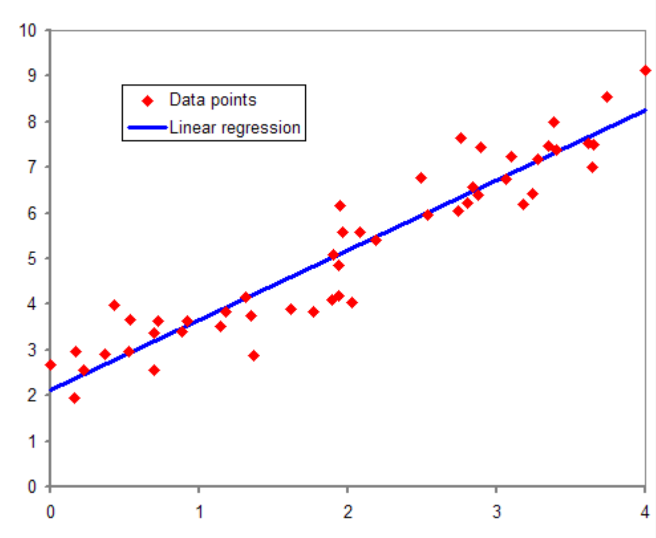
오차 제곱의 합 $\sum_{i=1}^{n}\left(y_i - (\beta_0 - \beta_1 X)\right)^2$, 일 경우

β_0, β_1을 편미분하여 편미분 값이 0이 되도록 하는 것

그래프 상에서 오차의 제곱이 제일 작은 함수를 얻음

1 최소 제곱법 수정

최소제곱법



※ 출처 : Linear Regression - wikipedia

1 최소 제곱법 수정

- 특이값에 취약
  - 데이터에 특이값이 포함되면 회귀 식의 예측 결과가 왜곡되는 현상이 발생
  - 왜곡 현상으로 인해 앞으로 수집할 데이터 예측 시 정확성을 잃음
  - 특이값에 페널티 부여나 제외를 통한 회귀 식을 수정
  - 데이터의 개수가 많은 경우 계산량이 매우 많이 증가함

- 오차 제곱의 합을 Cost Function이라 함

2 LOWESS 분석

- 어떤 한 지점에 가중회귀 함수를 사용해 평활화를 실행한 회귀 식 도출 방법
- 유연한 비선형 함수들을 적합하는 다른 기법
- 목표점 x_0에서 그 주변의 훈련 관측치들만을 사용하여 적합을 계산
- 전체적으로 모든 데이터는 적합하지 않음
- 최근 데이터가 중요할 경우 사용하며 최근 데이터에 중요도를 높여서 적합함

1 가중 회귀분석

2 LOWESS 분석

- 목표점에서 먼 데이터들을 제외시켜
  계산 상의 부담을 줄임

- 목표점 근처의 데이터를 어떻게 수집하는 지가 관건

✓ **평활화(Smoothing)**
: 연속성 있는 데이터와 관련 없는 데이터를 제거하거나 변화해서 데이터 전체를 연속성 있게 유지하는 작업

2 LOWESS 분석

- 임의로 설정한 폭 d(x)에서 x_i 최소값부터 차례로 값을 증가시켜 x에 가장 가까운 x_i 값이 되도록 가중치 w_i를 산출

[LOWESS 분석의 가중치 식]

$$w_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3$$

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

2 LOWESS 분석

- 평활화 실행을 위해 특이값을 없앨 수 있도록 가중치 w를 설정하는 방법으로 로버스트 평활화 사용
- 중위 절대편차를 산출했을 때 6배 이상의 잔차 r_i가 존재하면 w_i를 0으로 설정

[로버스트 평활화의 가중치 식]

$$w_i = \begin{cases} \left(1 - \left(\dfrac{r_i}{6MAD}\right)^2\right)^2 & |r_i| < 6MAD \\ 0 & |r_i| \geq 6MAD \end{cases}$$

$$MAD = \mathrm{median}(|r|)$$

MAD는 잔차의 중위 절대편차

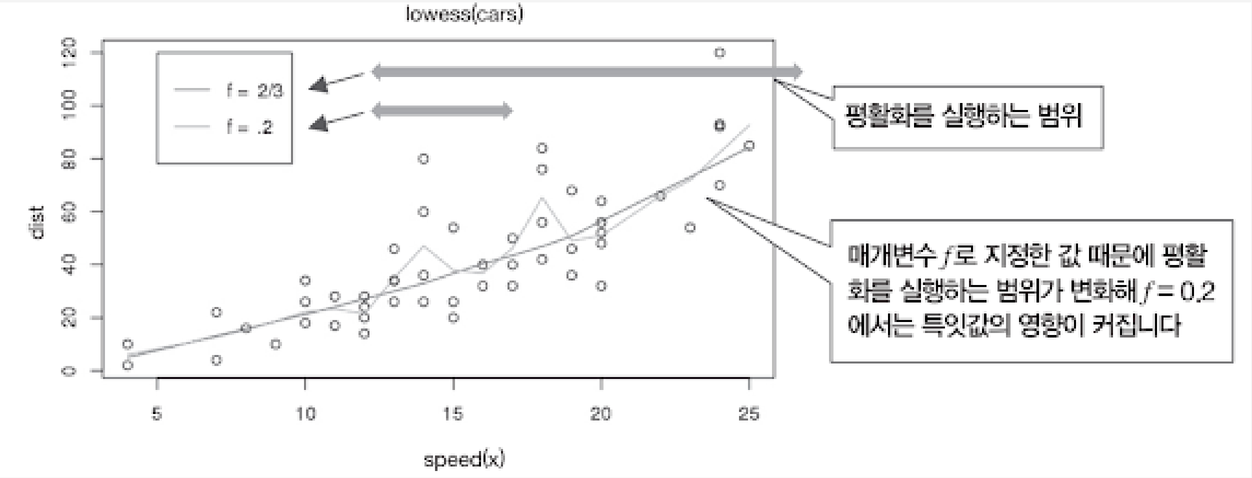
※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

2 LOWESS 분석

- 앞서 구한 가중치 계수 w와 독립 변수 x의 내적을 구했을 때 알 수 있는 종속 변수 y를 보정
- 독립 변수의 값에서 멀어져 있는 점의 기울기를 조정하여 특이점으로 인한 영향을 무시하도록 보정
- 단순 회귀를 반복해 실행하지만 실제 직선이 나타나지는 않음

2 LOWESS 분석

[독립 변수 x를 움직였을 때 가중치 변화 예]



※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

3 L2 정규화, L1 정규화

- 최소제곱법으로 구성한 방정식에 페널티를 부여
- L2 정규화, L1 정규화, 일래스틱(Elastic) 넷
- 벌칙 항(Penalty Term),
  정규화 항(Regularization Term)
  : 페널티를 부여하는 항

3 L2 정규화, L1 정규화

[L2 정규화 L1 정규화, 앨래스틱 넷 수식]

- L2 정규화

$$E = (Y - \omega^T X)^T (Y - \omega^T X) + \lambda \|\omega\|^2$$

$$\|\omega\|^2 = \sum_i \omega_i^2$$

$$\omega^T = (X^T X + \lambda I)^{-1} X^T Y$$

- L1 정규화

$$E = (Y - \omega^T X)^T (Y - \omega^T X) + \lambda |\omega|$$

$$|\omega| = \sum_i |\omega_i|$$

- 일래스틱 넷

$$E = (Y - \omega^T X)^T (Y - \omega^T X) + \lambda \sum_i (\alpha |\omega_i| + (1 - \alpha)\omega_i^2)$$

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

3 L2 정규화, L1 정규화

L2 정규화

- 최소제곱법의 종속 변수인 잔차 제곱의 합에 가중치 계수인 w_i 제곱의 합을 페널티로 추가
- 능형회귀
- 페널티 항을 L2 **노름(Norm)**이라 함

> ✓ **노름**
> : 벡터 공간의 원소들에 일종의 길이나 크기를 부여하는 함수

3 L2 정규화, L1 정규화

L2 정규화

- λ 값은 다양한 값을 적용하면서 교차검증법으로 최적값 검색
- 값이 크면 페널티도 강함
- L2 정규화 실행 시 정규방정식에서 $X^T X$ 항에 λI(I는 단위 행렬) 추가

3 L2 정규화, L1 정규화

L1 정규화

- Lasso(Least Absolute Shrinkage Selection Operator)
- w_i의 절대값을 페널티로 더해 줌
- 페널티 항을 L1 Norm이라 함
- L1 정규화를 실행하면 일부 w는 0이 되어 밀도가 낮아질 수 있음
- 모델을 구축할 때 특징량 선택에 이용할 수 있음을 의미

3 L2 정규화, L1 정규화

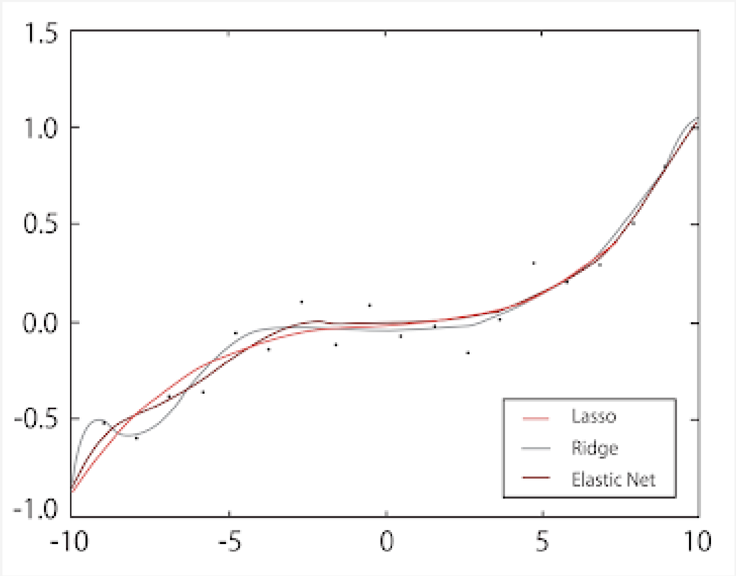
L1 정규화

- 신호 처리와 패턴 인식에 사용하기 쉽고
  다중 공선성 문제에도 대응 가능
- L2 정규화는 일반적인 회귀 모델로 계산
- L1 정규화는 볼록 최적화의 추정 알고리즘 사용

3 L2 정규화, L1 정규화

- 데이터 생성에 사용한 함수 : $y=0.001(x^3+x^2+x)$
- 함수 y에 N(0, 0.1) 을 따르는 무작위 값을 더해준 데이터 20개 이용
- L2 정규화 사용 시 약간의 학습 효과가 있으며, L1 정규화는 최적의 모델링을 얻었으며 일라스틱 넷은 L1과 L2 정규식의 중간 정도 모델링 결과를 보임

3 L2 정규화, L1 정규화



※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

2 유사도

1 유사도

- 회귀 분석 시 상관관계(상관계수)를 확인한다는 것은 유사도를 확인하는 것과 같음
- 유사도 확인 방법
  - 코사인 유사도
  - 상호 상관함수
  - 자기 상관함수
  - 자카드 계수
  - 편집 거리

1 유사도

- 추천 시스템이나 군집화 등 다양한 영역에서 유사도 사용
- 다양한 측정법이 존재하는 만큼 시스템 특성에 맞는 적절한 방법 사용 시 좀 더 나은 결과를 얻을 수 있음

1 유사도

- 변수값 쌍이 얼마나 비슷한가를 측정
- 컴퓨터에서 자동으로 답을 추측하는 과정에서 매우 중요

1 유사도

수학적인 유사도 개념

- 코사인 유사도
  - 내적 공간의 두 벡터간 각도의 코사인값을 이용하여 측정된 벡터 간의 유사한 정도를 의미
- 상관계수
  - 상관관계의 정도를 파악
- 상관 함수
  - 상관관계의 정도를 파악하는 수식

1 유사도

수학적인 유사도 개념

- 편집 거리(Edit Distance)
  - 단어의 유사도를 측정하는 방법 중 하나

- 자카드 계수
  - 두 집합 사이의 유사도 측정 방법으로 두 집합이 동일하면 1, 공통 원소가 하나도 없으면 0

2 코사인 유사도

- 변수값 x, y 가 주어졌을 때 cosθ의 값이 유사도로 표시
- 유사도는 0에서 1사이의 값
- 유사도가 높을 수록 1에 가까운 값
- 변수 x와 변수 y가 같은 방향성을 갖는 것으로 유사도 측정
- 벡터 길이와는 상관 없음

2 코사인 유사도

- 어떤 개수의 차원에도 적용이 가능
- 다차원의 양수 공간에서의 유사도 측정에 자주 이용

2 코사인 유사도

- 예
  - 정보 검색 및 텍스트 마이닝 분야
    - 단어 하나 하나는 각각의 차원을 구성하고
      문서는 각 단어가 문서에 나타나는 회수로
      표현되는 벡터값
    - 이러한 다차원 공간에서 코사인 유사도는
      두 문서의 유사도를 측정하는 매우 유용한 방법
  - 데이터 마이닝 분야의 클러스터 간의 응집도 측정

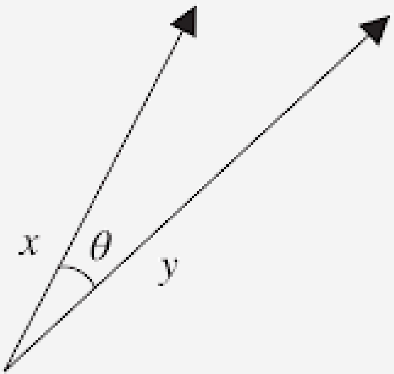
2 코사인 유사도

$$cos\theta = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

$\|x\|$를 x의 노름(Norm)이라 함
분자 : x와 y의 내적
분모 : x의 벡터 크기 및 y의 벡터 크기

$$cos\theta = \frac{(x_1y_1 + x_2y_2 + x_3y_3 + \cdots + x_ny_n)}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

2 코사인 유사도



x와 y의 방향성을 고려하여
cos 값을 구하면

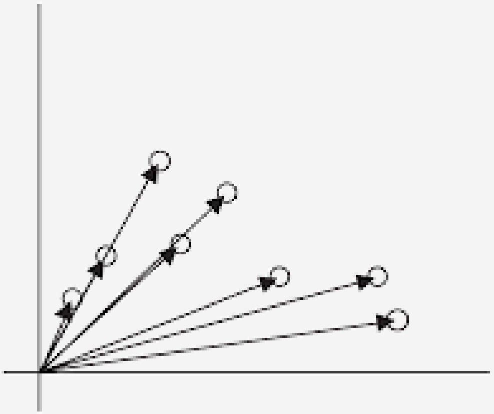
- 같은 방향을 가리킬 경우 각도가 0이 되어 cos 값이 1이 됨
- 비슷한 방향을 가리킬 경우 각도가 0에 가까워 cos 값이 1에 가까움
- 다른 방향을 가리킬 경우 각도가 90도에 가까워 cos의 값이 0에 가까움

2 코사인 유사도

- 문서 간의 유사도 계산 시 사용
- 문서에 나타나는 단어의 출현 빈도를 구해 유사도 계산 식에 적용
- 변수 값 쌍은 산포도를 사용해 점의 집합으로 표현

2 코사인 유사도

[변수값 쌍의 산포도와 각 점으로의 벡터]



단어 목록 n
: 유사도를 요구하는 문서 1과
문서 2의 모든 단어로 구성

- x : 문서 1의 단어가 나오는 빈도
  (i= 1, 2, 3, ⋯, n)
- y : 문서 2의 단어가 나오는 빈도
  (i= 1, 2, 3, ⋯, n)

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

2 코사인 유사도

▶ 코사인 유사도 식의 변형

$$cos\theta = \frac{(x_1y_1 + x_2y_2 + x_3y_3 + \cdots + x_ny_n)}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

↓

$$cos\theta = \frac{((x_1 - x_0)(y_1 - y_0) + (x_2 - x_0)(y_2 - y_0) + (x_3 - x_0)(y_3 - y_0) + \ldots + (x_n - x_0)(y_n - y_0))}{\sqrt{\sum_{i=1}^{n} (x_i - x_0)^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - y_0)^2}}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

x_0, y_0는 x, y의 원점

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

2 코사인 유사도

- 코사인 유사도 예
  - 2주차 강의에서 언급한 경주 지진의 신문 기사 별 유사도 분석
  - 코사인 유사도 식을 이용하여 단어의 빈도 표의 값을 적용, 유사도 산출

2 코사인 유사도

▶ 코사인 유사도 예

[기사에서 출현한 단어의 빈도 표]

	기사A	기사B	기사C	기사D
경주	0.5	0.4	0	0.5
지진	0.2	0	0.4	0.3
지층	0.1	0	0.4	0
단층	0	0.3	0.2	0.1
비	0.2	0	0	0
연휴	0	0.3	0	0.1

[코사인 유사도 식을 이용한 유사도 분석표]

	기사A	기사B	기사C	기사D
기사A	1.000			
기사B	0.588	1.000		
기사C	0.343	0.171	1.000	
기사D	0.886	0.743	0.389	1.000

- 기사 A와 유사한 기사는 기사 D로 계산 됨
- 기사 B와 유사한 기사는 기사 D로 계산 됨

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

2 코사인 유사도

- 상관 관계
  - 2개의 확률 변수 사이의 분포 규칙의 관계
  - 분포 규칙
    - 한쪽이 증가하면 다른 한쪽도 증가하고 한쪽이 감소하면 다른 한쪽도 감소하는 것
  - 대부분 선형 관계의 정도를 의미

- 상관 계수
  - 보통 피어슨 상관 계수를 의미

2 코사인 유사도

▶ 피어슨 상관 계수

- 두 변수 간의 관련성을 구하기 위해 보편적으로 사용
- r : $\dfrac{x\text{와 } y\text{가 함께 변하는 정도}}{x\text{와 } y\text{가 각각 변하는 정도}}$
  - 1 : x와 y가 완전히 동일함
  - 0 : x와 y가 전혀 다름
  - -1: x와 y가 반대방향으로 완전히 동일함
  - 1 또는 -1 에 가까울 수록 강한 상관 관계
  - 보통 0.7 이상이면 상관 관계가 있다고 판단

2 코사인 유사도

▶ 피어슨 상관 계수

x와 y가
함께 변하는 정도

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x와 y가
각각 변하는 정도

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

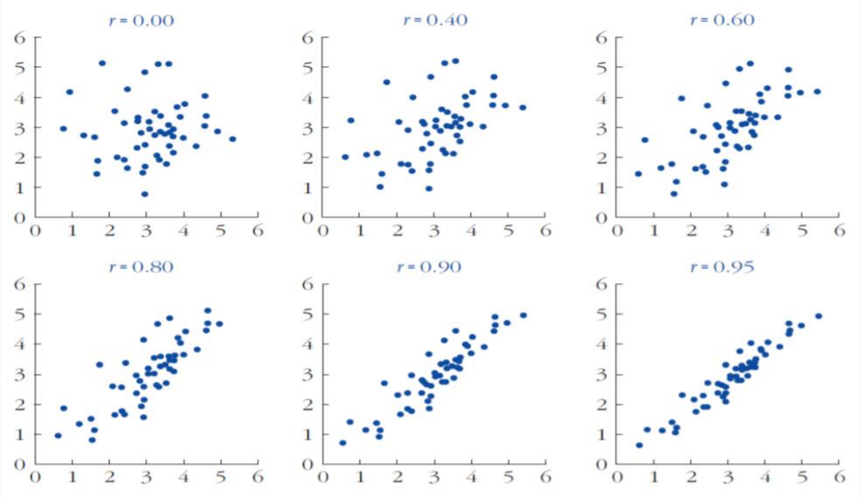
3 상관계수

- 상관계수 의미
  - 상관계수의 절대값이 1에 가까움
    - 단순 회귀의 점 분포에 불규칙성이 작음을 의미
  - 점 분포에 불규칙성이 작아도 상관계수는 0에 가까울 수 있음
  - 불규칙성이 전혀 없어 표준 편차가 0일 경우 상관 계수를 산출 할 수 없음

3 상관계수

▶ 상관계수 의미

[양의 상관 관계를 보이는 산포도]



※ 출처 : http://ezstat.snu.ac.kr/textbook_sources/chapter_05.pdf

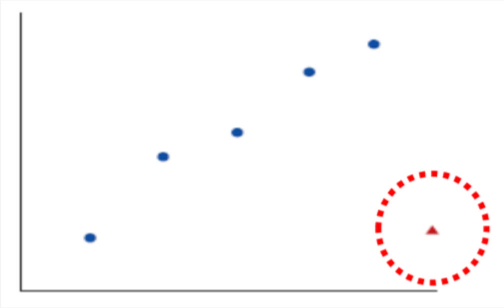
3 상관계수

- 상관계수 특징
  - 상관계수는 단위가 없음, 측정 단위와 독립적
  - 상관계수는 방향성을 갖지 않음
  - 상관계수의 값이 0.8일 경우 상관 계수가 0.4인 경우보다 2배의 상관 관계를 갖는 것은 아님

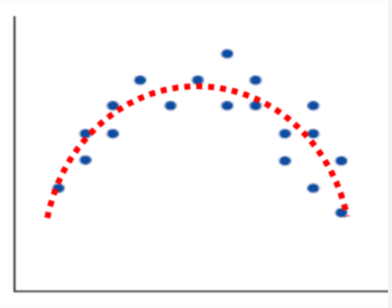
3 상관계수

- 상관 관계를 사용하지 못하는 경우
  - 경향을 따르지 않는 이탈 값들이 존재할 경우
  - 두 변수 간의 관계가 비선형일 경우

[이탈 값 예]



[비선형 관계]



※ 출처 : http://ezstat.snu.ac.kr/textbook_sources/chapter_05.pdf

3 상관계수

- 스피어만의 순위 상관 계수
  - 피어슨 상관 계수의 특별한 경우
  - 데이터가 서열 척도인 경우 자료의 값 대신 순위를 이용하는 경우의 상관 계수
  - 자료에 이상점이 있거나 표본 크기가 작을 경우 유용
  - 피어슨 상관 계수 식을 이용하여 확률 변수들의 값을 순위로 계산

3 상관계수

- 스피어만의 순위 상관 계수
  - 피어슨 상관계수 식(a)을 이용하지만 변수들의 값을 순위(b)로 사용하면 스피어만의 순위 상관계수 식(c) 생성

(a)

$$r_{xy}=\frac{S_{xy}}{S_x S_y} \qquad (-1<r_{xy}<1)$$

$$=\frac{\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2\times\frac{1}{n}\Sigma(y_i-\bar{y})^2}}$$

$$=\frac{\Sigma x_i y_i-\bar{x}\bar{y}}{\sqrt{\Sigma(x_i^2-n\bar{x}^2)\times\Sigma(y_i^2-n\bar{y}^2)}}$$

$\Sigma x_i = \Sigma y_i = n(n+1)/2$

$\Sigma x_i^2 = \Sigma y_i^2 = n(n+1)(2n+1)/6$

$x - = y \sim = (n+1)/2$

$x \sim {}^2 = y \sim {}^2 = (n+1)^2/4$

(b)

(c)

$$= 1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(x_i-y_i)^2$$

※ 출처 : http://www.tamagaki.com/math/Statistics609.html

3 상관계수

- 스피어만의 순위 상관 계수 예
  - 예) 수학을 잘하는 학생이 영어도 잘하는 것과의 상관 관계를 알아보고자 할 경우
  - 학생 10명의 표본을 이용하여 수학 순위와 영어 순위를 작성
  - 학생 별로 수학을 잘하는 학생이 영어도 잘하는지 스피어만의 순위 상관 계수 활용
  - 결과가 0.703으로 상관 관계가 있는 것으로 분석됨

3 상관계수

▶ 스피어만의 순위 상관 계수 예

학생	수학의 순위	영어 순위
1	6	10
2	4	1
3	5	4
4	10	9
5	2	3
6	8	8
7	3	6
8	9	5
9	1	2
10	7	7

$$= 1 - 6 * \sum \{(\text{수학 순위}) - (\text{영어 순위})\}^2 / \{10 * (10^2 - 1)\}$$

$$= 1 - 6 * \{(-4)^2 + 3^2 + 1^2 + 1^2 + (-1)^2 + (-4)^2 + 0^2 + (-3)^2 + 4^2 + (-1)^2 + 0^2\}/990$$

$$= 0.703$$

※ 출처 : http://www.tamagaki.com/math/Statistics609.html

4 거리와 유사도

- 유사도를 거리와 같은 개념으로 이해
- 거리가 가까울 수록 유사도가 높은 것으로 분석
- 종류
  - 편집 거리(Edit Distance)
  - 레벤슈타인 거리(Levenshtein Distance)
  - 해밍 거리(Hamming Distance)
  - 유클리드 거리(Euclidian Distance)
  - 마할라노비스 거리(Mahalanobis Distance)

4 거리와 유사도

편집 거리(Edit Distance)

- 치환, 삽입, 삭제의 세 요소를 각각 페널티를 설정하는 형태를 취해 페널티의 합계를 점수로 설정하여 유사도 규정
- 예: spring과 print의 유사도 계산

spring → print s, g 삭제, t 삽입	3번의 연산(2번의 삭제 + 1번의 삽입)이 필요하므로 편집 거리는 3

spring → print s삭제, g를 t로 치환	2번의 연산(1번의 삭제 + 1번의 치환)이 필요하므로 편집 거리는 2

4 거리와 유사도

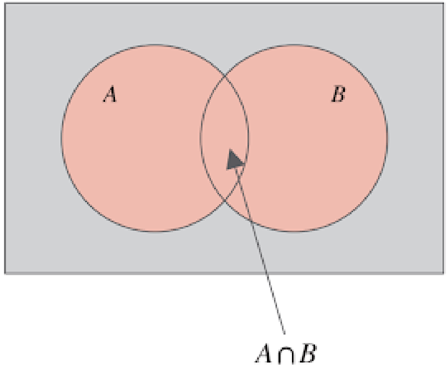
자카드 계수

- 집합 2개의 유사도를 구할 때
- 집합 2개의 공통 요소 수를 전체 요소 수로 나눈 것으로 표현
- 벤다이어그램을 이용하여 간편하게 적용 가능
- 집합 구성 요소가 수치인지 문자열인지 고려 없이 적용 가능

4 거리와 유사도

자카드 계수

[벤다이어그램과 자카드 수식]



자카드 계수 식

$$r = \frac{|\text{집합 } A \cap \text{집합 } B|}{|\text{집합 } A \cup \text{집합 } B|}$$

※ 출처 : 처음 배우는 인공지능, 다다 사토시, 한빛미디어, 2017

4 거리와 유사도

자카드 계수

- 자카드 계수 예
  - 사용자 1과 사용자 2가 구매한 상품의 유사도를 구하고자 하는 경우

	item1	Item2	Item3	Item4
User1	0	1	0	1
User2	0	1	1	1
user3	1	0	1	0

User1과 user2의 합집합은 3
User1과 user2의 교집합은 2
따라서 자카드 유사도는 0.67

※ 출처 : http://www.jidum.com/jidums/view.do?jidumId=1087

4 거리와 유사도

자카드 계수

- 자카드 계수 예
  - 사용자 1과 사용자 2가 구매한 상품의 유사도를 구하고자 하는 경우

	User1	user2	user3
User1	1	0.67	0
User2	0.67	1	0.25
user3	0	0.25	1

User1, 2, 3 간의 자카드 유사도 계산

※ 출처 : http://www.jidum.com/jidums/view.do?jidumId=1087