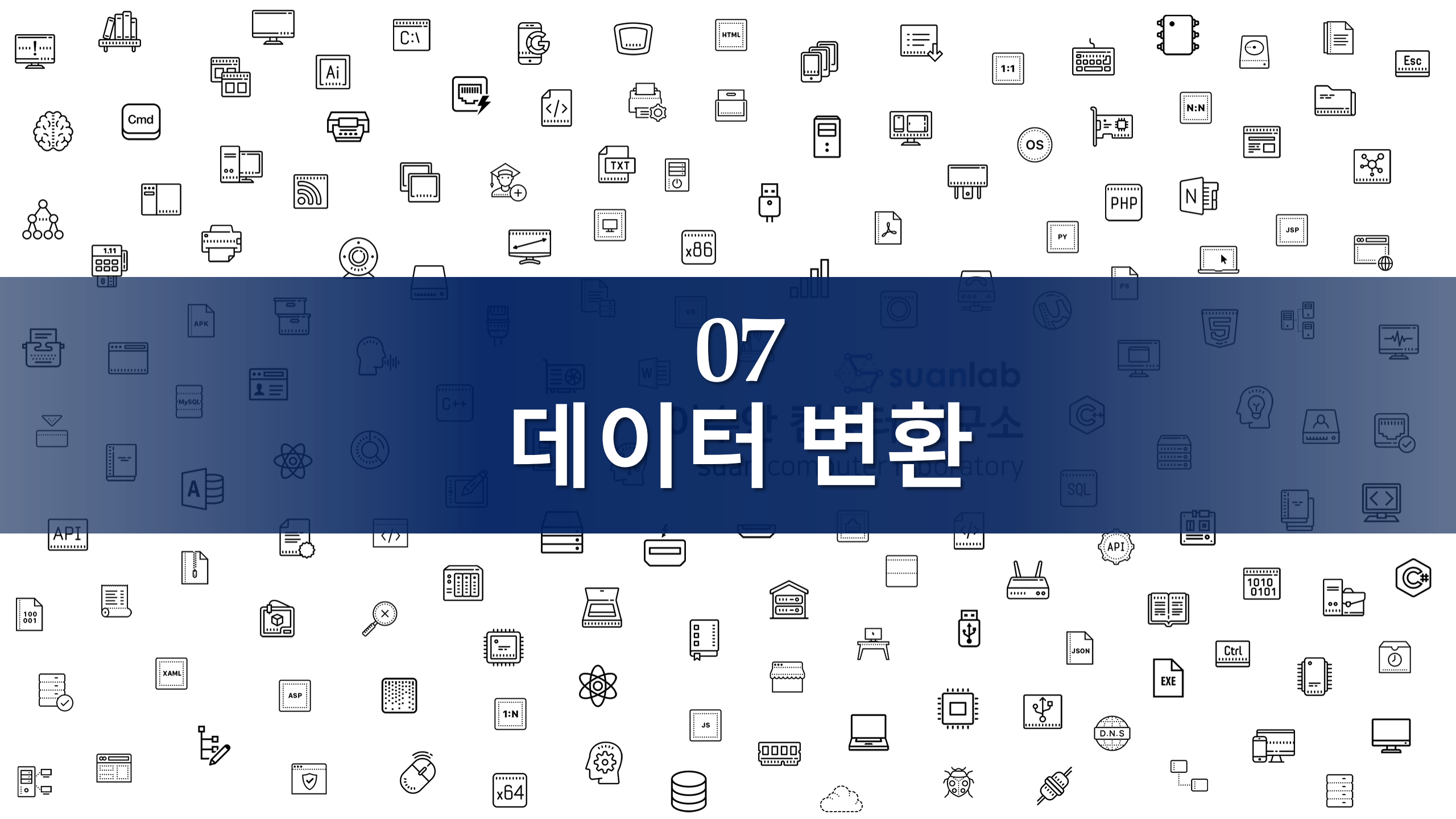
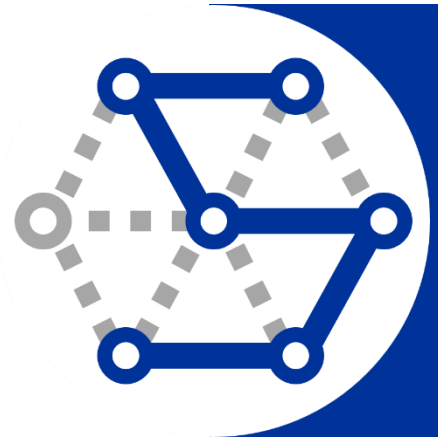
데이터 전처리
Data Preprocessing


07
데이터 변환

목차

데이터 변환Data Transformation

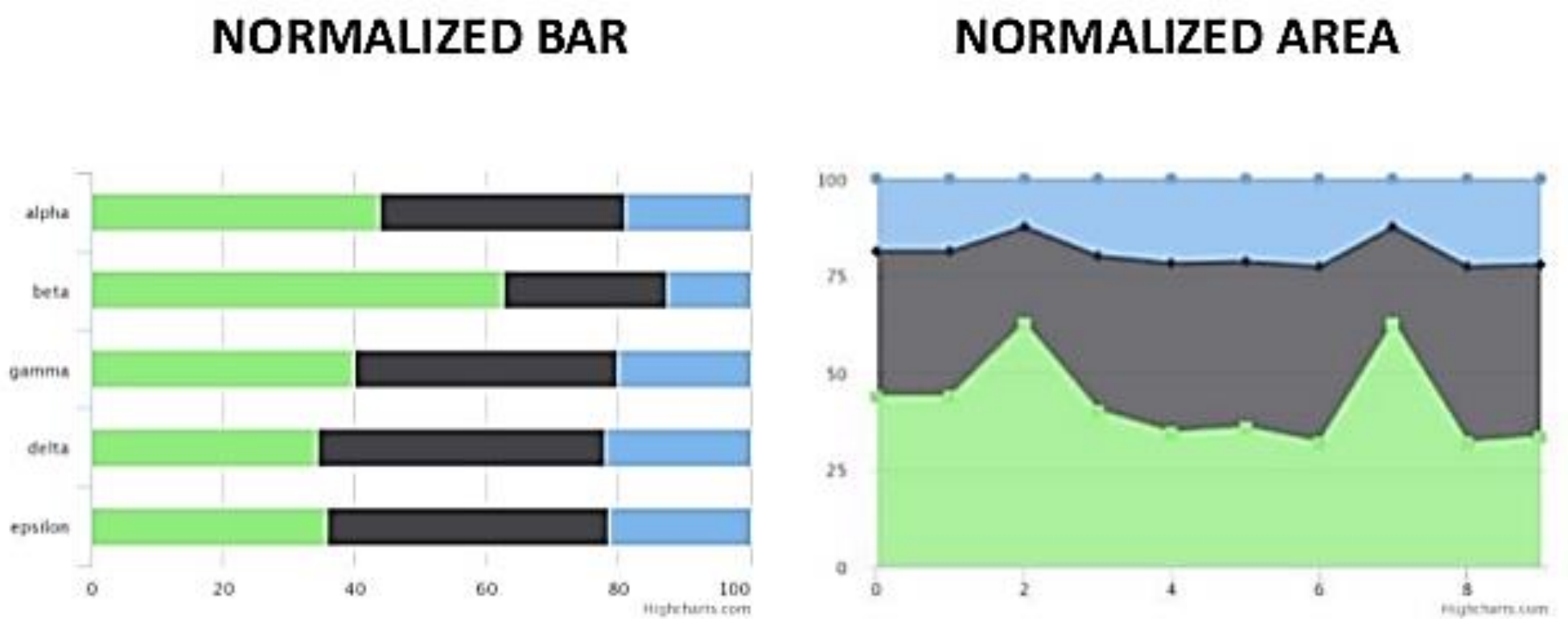
- 데이터 변환은 데이터 분석에 적절한 형태로 데이터를 바꾸는 전처리 작업을 의미
- 데이터 변환 방법
  - 평활Smoothing: 데이터로부터 잡음을 제거. 구간화binning, 회귀regression, 군집화clustering 등이 대표적 기법
  - 집계Aggregation: 그룹화 연산을 데이터에 적용 (일일 판매 데이터를 월별 또는 연도별로 그룹화하는 기법으로 주로 데이터 큐브를 구축하는 데 사용)
  - 속성 구성Attribute Construction: 주어진 속성 집합으로부터 새로운 속성을 구성하는 기법
  - **정규화Normalization: 데이터를 정해진 구간 내에 존재하도록 변환시키는 기법**
  - **이산화Discretization: 수치형 속성의 원시값raw-value을 구간 라벨로 대체하거나 개념적인 라벨로 대체하는 기법**
  - **개념 계층Conceptual Hierarchy: 도로명과 같은 속성을 시나 국가와 같은 상위 레벨 개념으로 일반화시키는 기법 (많은 명목형 속성들이 암시적인 개념 계층을 가지고 있음)**



1. 정규화

정규화Normalization

- 정규화는 속성값으로 -1.0 ~ 1.0과 같이 정해진 구간 내에 들도록 하는 기법
- 최단 근접 분류와 군집화와 같은 거리 측정 등을 위해 특히 유용



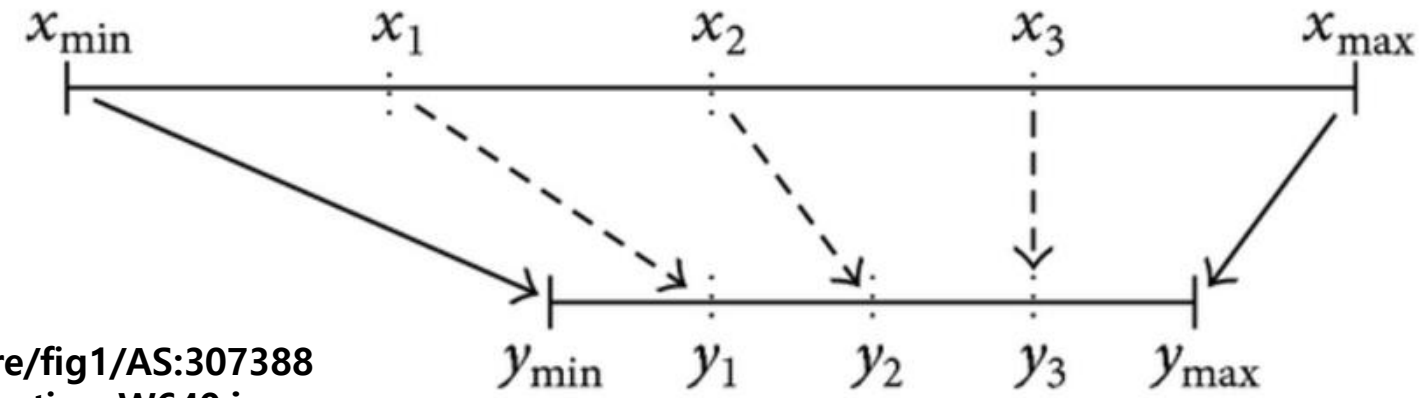
https://www.slideshare.net/nsrivast/data-analytics-club-data-visualization-workshop

최소-최대 정규화min-max normalization

- 원본 데이터에 대하여 선형 변환 수행
- 속성 A에 대한 최소값과 최대값을 min_A와 max_A라고 가정
- A의 값은 다음 계산식에 의해 v를 구간 에서의 값 v'으로 사상

$$v' = \frac{v - min_A}{max_A - min_A}(max'_A - min'_A) + min'_A$$

- 최소-최대 정규화는 원본 데이터 값들 간의 관계를 보존
- 정규화를 위한 입력이 A에 대한 원본 데이터 구간에서 벗어난 경우는 범위 초과out-of-bounds 오류 발생



https://www.researchgate.net/publication/282541174/figure/fig1/AS:307388692353061@1450298583749/Min-max-method-of-normalization_W640.jpg

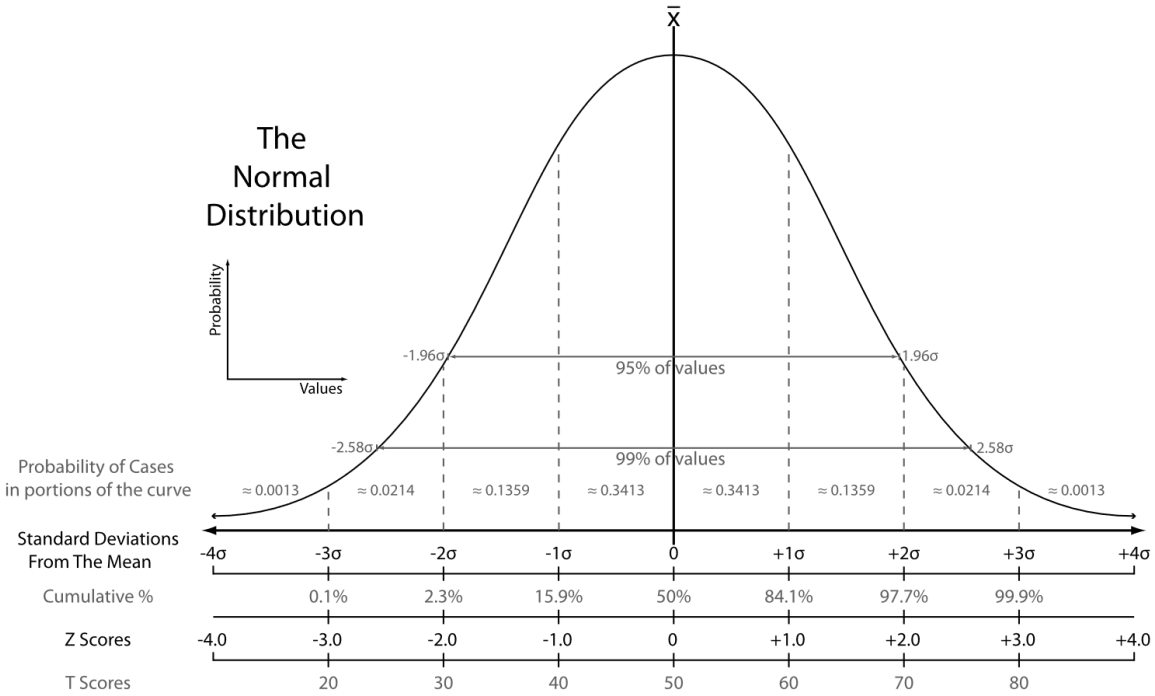
Z-score 정규화Z-score normalization

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 속성 A에 대한 값을 A의 평균과 표준편차를 기초로 정규화하는 방법
- $\bar{A}$와 σ_A를 A의 평균과 표준편차라고 가정
- A의 값 v는 다음 계산식에 의해 v'에 사상

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

- Z-score 정규화는 속성 A의 실제 최소값과 최대값이 알려져 있지 않거나, 최소-최대 정규화에 큰 영향을 주는 이상치가 존재할 때 유용

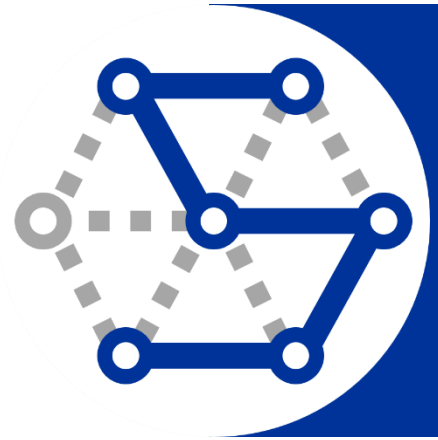


https://upload.wikimedia.org/wikipedia/commons/2/25/The_Normal_Distribution.svg

소수 척도화decimal scaling

- 속성 A 값들의 소수점을 이동해서 정규화
- 이동되는 소수점의 수는 A의 최대 절대값에 의존
- $Max(|v'|) < 1$을 만족하는 가장 작은 정수를 j라고 가정
- A의 값 v는 다음 계산식에 의해 v'로 사상

$$v' = \frac{v}{10^j}$$



2. 수치형 데이터 이산화

수치형 데이터 이산화numeric data discretization

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 구간화binning
- 히스토그램histogram
- 엔트로피-기반 이산화
- 카이제곱 χ^2 결합
- 군집 분석cluster analysis
- 직관적 분할에 의한 이산화

엔트로피-기반 이산화entropy-based discretization

- Top-down 방식의 이산화 기법
- 분할점들의 결정과 계산을 위해 클래스 분포 정보를 탐색
- 수치형 속성 A의 값들을 분할하기 위하여 분할점으로 최소 엔트로피를 갖는 A의 값을 선택
- 계층적 이산화를 달성하기 위하여 그 결과 구간을 분할
- 이러한 방식을 통해 속성 A의 개념 계층 생성

엔트로피-기반 이산화 절차

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 데이터 집합 D의 속성 A에 대한 엔트로피-기반 이산화 절차
  1. A의 각 값은 A의 범위를 분할하는 잠재적인 분할점split-point로 간주
     하나의 분할점에 의해 이항형binary 이산화 진행 가능
  2. 이항형 이산화 결과로 분리된 데이터 집합을 D_1,D_2라 하면, 속성 A의 기대 정보 요구량expected information requirement $Info_A(D)$계산식

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2)$$

     - $|D|$: 데이터 집합 D의 튜플 개수
     - $|D_1|$: m개의 클래스 $C_1, C_2, \ldots, C_m$가 주어진 상황에서 의 D_1 엔트로피는 $Entropy(D_1) = -\sum_{i=1}^{m} p_i \log^2(p_i)$ where p_i : 데이터 집합 D_1 내에서이 클래스 C_1 확률
     - 속성 A의 분할점을 선택할 때< 기대정보요구량을 최소화하는 방향으로 부할점을 선택
  3. 1. 2. 과정을 어떤 정지 규칙stopping criterion이 만족될 때까지(예를 들어, 모든 분할점 후보에 대한 최소 정보 요구량이 어떤 임계치threshold ε보다 작을 때까지, 혹은 구간의 수가 미리 설정한 최대 구간 수보다 클 때까지 구해진 각 분할에 재귀적으로 적용

카이제곱 χ^2 결합

- 대부분의 이산화 기법들이 top-down 분할 방식을 적용하는 것과는 달리 카이제곱 χ^2 결합 기법은 가장 가까운 이웃 구간을 찾아내고 이들을 결합하여 더 큰 구간을 형성시키는 bottom-up 방식을 적용
- 엔트로피-기반 이산화와 같이 클래스 정보를 활용한다는 측면에서 지도supervised 기법
- 카이제곱 χ^2 결합 기법의 핵심 개념은 정교한 이산화를 위해 클래스 상대도수가 구간에 매우 일관성이 있어야 함
- 만약 두 개의 이웃 구간들이 매우 유사한 분포를 가진다면 그 구간들은 결합되고, 그렇지 않으면 분리된 상태로 둠

(복습) 카이제곱검정Chi-Square Tests

- 범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨Pearson의 카이제곱 χ^2검정에 의해 측정 가능
- 속성 A가 c개의 범주 값 $a_1, a_2, \cdots, a_c$를 가진다고 가정
- 속성 B가 r개의 범주 값 $b_1, b_2, \cdots, b_r$를 가진다고 가정
- 속성 A와 B에 의해 구성되는 튜플은 c개의 열과 r개의 행으로 구성되는 분할표로 표현
- (A_i, B_j)를 속성 A가 a_i를 가지고, 속성 B가 b_j를 가지는 튜플이라고 할 때 χ^2 의 정의

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  - o_{ij} : (A_i, B_j)에 대한 관측도수observed frequency; 실제로 존재하는 (A_i, B_j) 튜플 수
  - e_{ij} : (A_i, B_j)에 대한 기대도수expected frequency; 확률적으로 기대되는 (A_i, B_j) 튜플 수

(복습) 카이제곱검정Chi-Square Tests

- 기대도수 e_{ij} 계산 식

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

  - N : 데이터 튜플 수
  - $count(A = a_i)$: 속성 A에 대하여 a_i를 갖는 튜플 수
  - $count(B = b_j)$: 속성 B에 대하여 b_j를 갖는 튜플 수

- χ^2 통계량은 속성 A와 B가 독립이라는 가설을 검증
- 카이제곱검정은 자유도 $(r-1) \times (c-1)$을 갖는 유의수준에 근거하여 검정

카이제곱 χ^2 결합

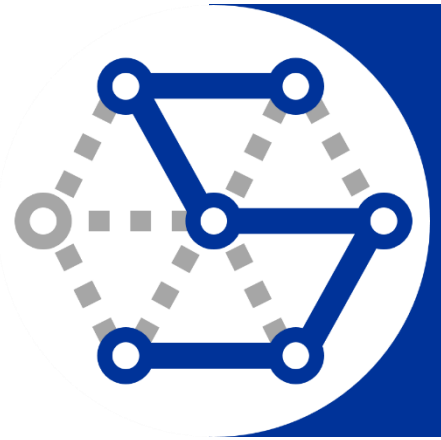
- χ^2 통계량은 주어진 속성에 대한 "두 개의 이웃 구간들의 독립이다" 라는 가설 검정
- 데이터 분할표를 구성하면 분할표는 두 개의 이웃 구간을 의미하는 두 개의 열과 m개의 행을 가짐
- 값 o_{ij}는 i번째 구간의 j번째 클래스에 해당하는 튜플들의 수
- 유사하게 o_{ij}의 기대빈도수 e_{ij}는 (구간 i의 튜플 수)x(클래스 j의 튜플 수)/N이며, 여기서 N은 데이터 튜플의 총 수
- 한 구간 쌍에 대해 낮은 χ^2 값은 '구간들의 클래스와 독립' 이라는 것을 의미하고 그 구간들이 결합할 수 있다는 것을 의미

카이제곱 χ^2 결합

- 정지 규칙은 세 가지 조건 중 하나에 의해 결정
  1) 모든 인접 구간들에 대한 χ^2 값들이 유의수준에 의해 설정된 임계치를 초과
  2) 구간의 개수가 사전 정의된 최대-구간 내로 들어옴
  3) 상대도수가 구간 내에서 일치(또는 사전 허용 오차 범위 내)

직관적 분할에 의한 이산화

- 직관적 분할의 대표적인 기법 3-4-5 규칙
- 3-4-5 규칙은 데이터의 주어진 범위를 '가장 큰 유효숫자most significant digit'에서의 값의 범위를 토대로, 재귀적recursively 그리고 수준별로 3, 4, 5개의 비교적 동일한 폭의 구간들로 분할
- 어떤 구간이 가장 큰 유효숫자에서 3, 6, 7, 9개의 개별 값들을 포함한다면, 그 범위를 세 개의 동일 폭 구간들로 분할, 7에 대해서는 2-3-2로 묶인 세 개의 구간으로 분할
- 어떤 구간이 가장 큰 유효숫자에서 2, 4, 8개의 개별 값들을 포함한다면, 그 범위를 네 개의 동일 폭 구간들로 분할
- 어떤 구간이 가장 큰 유효숫자에서 1, 5, 10개의 개별 값들을 포함한다면, 그 범위를 다섯 개의 동일 폭 구간들로 분할

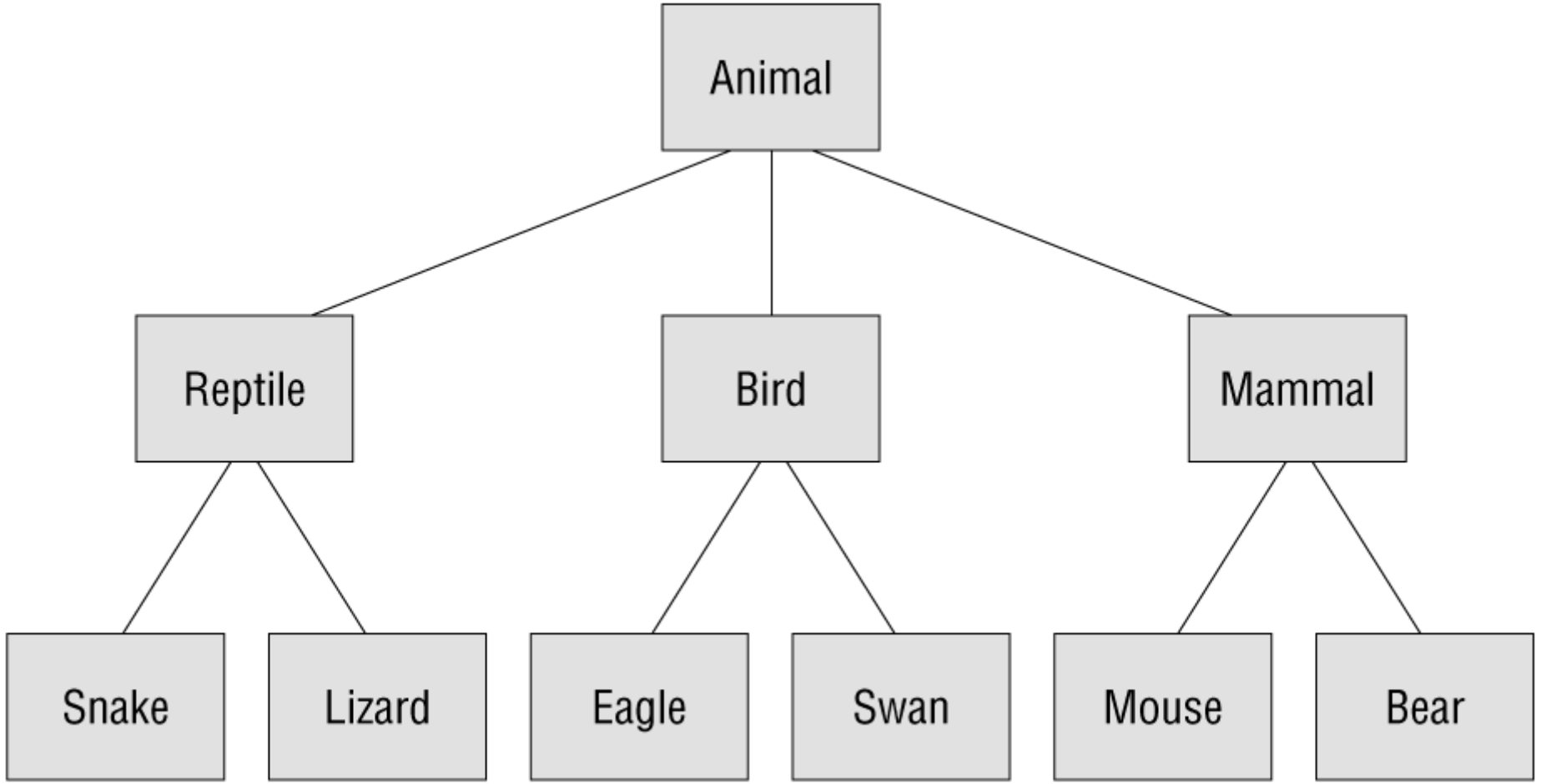


3. 범주형 데이터 개념 계층

범주형 데이터 개념 계층

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

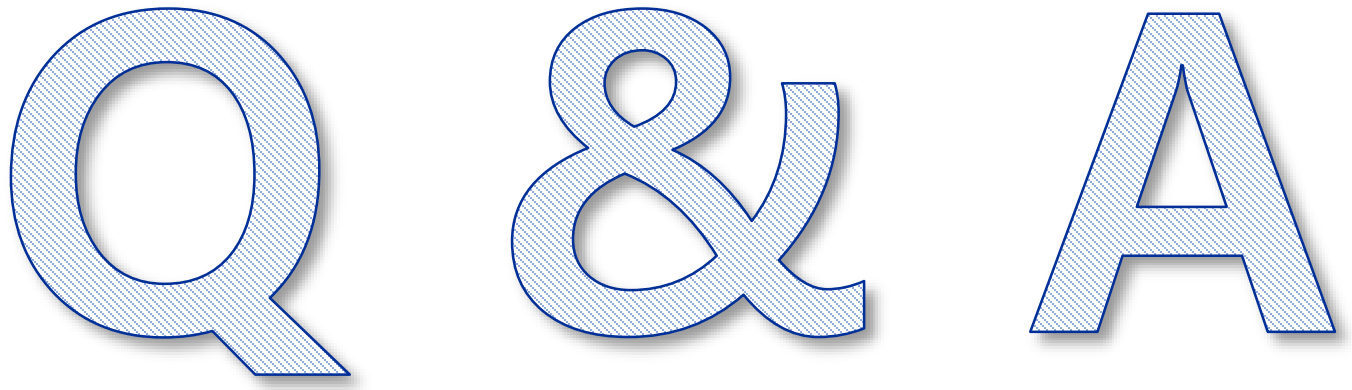
- 범주형 데이터는 기본적으로 이산 데이터로서 값들 사이에 순서가 없는 유한개의 구별 값을 가짐



범주형 데이터 개념 계층 생성 방법

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 스키마 단계에서의 명시적 생성
  - 스키마 수준에서 속성들의 개념 계층 정의
  - 주소의 경우, 스키마 수준에서 '동 < 기초시군구 < 광역시도 < 국가'를 정함
- 명시적 그룹화에 의한 일부 생성
  - 개념 계층의 일부를 수동으로 정의
  - [서울특별시, 경기도, 인천광역시] ⊂ 수도권, [전라북도, 전라남도, 광주광역시] ⊂ 호남권
- 속성들의 개념 계층을 자동 생성
  - 개념 계층은 일반적으로 상위 개념 수준이 몇 개의 하위 개념 수준을 포함하므로, 상위 수준의 개념 수가 하위 수준의 개념 수보다 적음
  - 각 수준에서의 구별 값들의 개수를 근거로 상위와 하위의 순서를 자동으로 정하여 개념 계층 생성
- 속성들의 부분 집합만을 생성
  - 개념 계층의 생성 시에 관련 속성들의 일부만 사용하는 방식
  - 주소에 대해서 동과 기초시군구 만을 사용하여 개념 계층 생성


Q & A