Bigdata Intelligence

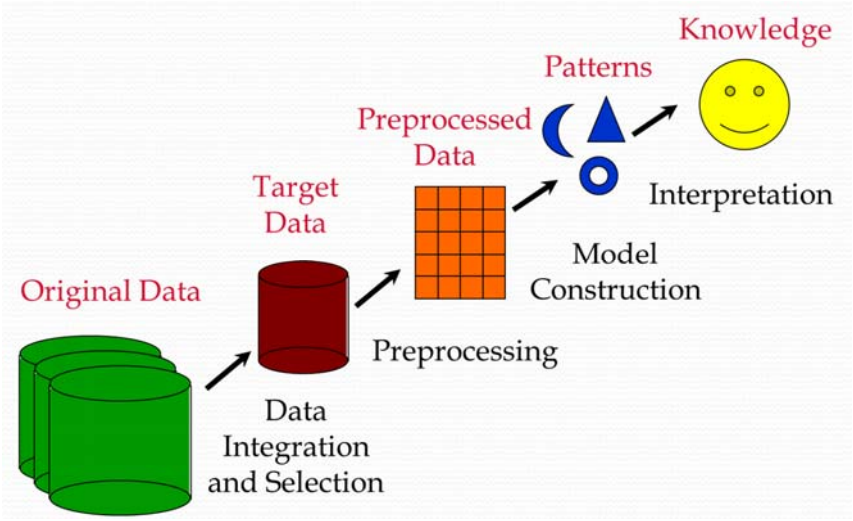
Python과 SQL을 활용한 데이터 전처리

Ch.1 데이터전처리 개요

데이터전처리 개요

- 데이터 전처리 정의

  데이터 분석 작업을 하기 전에 데이터를 분석하기 좋은 형태로 만드는 과정을 총칭하는 개념



* 출처 : Ramakrishnan and Gehrke. Database Management Systems, 3rd Edition

[그림 1.1] 데이터마이닝 과정에서의 데이터 전처리

데이터전처리 개요

- 데이터 전처리가 필요한 이유
  - 실무 데이터는 분석 기법을 바로 적용하기 힘든 형태
  - 비어있음(missing value), 잡음(noise), 적합하지 않은 데이터구조
  - 낮은 품질의 데이터로는 좋은 분석결과를 얻기 힘듬
- 데이터 품질 저하의 원인
  - 불완전(incomplete)
    - 데이터가 비어 있어 있는 경우로 DB 테이블의 속성값이 NULL인 경우
  - 잡음(noisy)
    - 데이터에 오류(error)가 포함된 경우. 예) 나이 = -20
  - 모순된(inconsistent)
    - 데이터 간의 정합성(일관성)이 없는 경우. 예) 성별은 남자인데, 주민번호 뒷 7자리 중 첫 번째 자리가 2인 경우
- 고품질 데이터라 하더라도 전처리 필요
  - 실무에서 존재하는 데이터의 구조적 형태(format)가 분석목적이나 분석기법에 적합한 경우가 드물기 때문

데이터전처리 예 – 웹로그 분석

- 분석 환경 및 주제
  - 웹로그 : Apache 사의 access log ([그림 1.2])
  - 사이트 : 사진을 서비스하고 앨범을 만들어 주는 사이트로 가정
  - 분석주제 : 사진 조회 빈도 및 사진 간 연관조회 현황 분석

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2004:16:06:51 -0800] "GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2 HTTP/1.1" 200 4523
64.242.88.15 - - [07/Mar/2004:16:07:03 -0800] "GET /07T2KZone/PostPhotos/photo_read.asp?oid=3110&category=1&theme=t01&page_num=11&konum=3&kosm=m7_3 HTTP/1.1" 200 76137 Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.7) Gecko/20060909 Firefox/1.5.0.7
```

[그림 1.2] Apache access log의 예

데이터마이닝의 이론적 사례

다음과 같은 itemset을 갖는 트랜잭션 t_1, t_2, t_3가 있다.

t_1 = {a, b, c}, t_2 = {a, b}, t_3 = {c}

트랜잭션 t_1, t_2, t_3에 대해서, minimum support 0.6 이상인 itemset을 구하고, 이들을 대상으로 minimum confidence 0.8 이상인 연관규칙을 찾으시오.

[그림 1.3] 연관규칙에 대한 이론적 사례

- 실무에서는
  - t1, t2, t3와 같은 트랜잭션의 개념을 어떻게 잡아야 할까?
  - 사례처럼 간단명료한 항목집합(itemset)에 존재할까?

➔ 실무에서 존재하는 [그림 1.2]와 같은 데이터를 분석이 용이하도록 [그림 1.3]과 같은 데이터 형태로 바꾸는 데이터전처리 필요

웹로그 분석을 위한 데이터 전처리 단계

1) 웹로그와 사진정보 DB 간의 매개 현황 파악

- 웹로그의 클라이언트 요청 URL에 존재하는 모듈과 패러미터 파악
  - ✓ 사진조회 모듈 : photo_read.asp
  - ✓ 패러미터 : oid, category, theme

2) 사진조회와 직접적인 관련이 없는 로그 제거

64.242.88.15 - - [07/Mar/2004:16:07:03 -0800] "GET /07T2KZone/PostPhotos/photo_read.asp?oid=3110&category=1&theme=t01&page_num=11&konum=3&kosm=m7_3 HTTP/1.1" 200 76137 Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.7) Gecko/20060909 Firefox/1.5.0.7

[그림 1.4] 로그 필터링을 통해 추출한 분석대상 로그

웹로그 분석을 위한 데이터 전처리 단계

3) 로그 파싱(parsing)을 통한 패러미터 값 추출

- 사진정보 조회에 필요한 패러미터 값을 얻기 위해 클라이언트 요청 URL(CLIENT_FULL_RESQ 항목)을 다시 파싱

SEQ	1814718
CLIENT_IP	84.222.150.222
SERVER_IP	-
AUTH_NM	-
DATE_TIME	2008-04-01 17:06
CS_METHOD	GET
CLIENT_FULL_RESQ	http://chinese.tour2korea.com/07T2KZone/PostPhotos/photo_read.asp?oid=3110&category=1&theme=t01&page_num=11&konum=3&kosm=m7_3
FLAG	
URI_PROTOCOL	HTTP/1.1
SERVER_STAT	200
CONTENT_LENGTH	76137
REFERER	-
USER_AGENT	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.7) Gecko/20060909 Firefox/1.5.0.7
COOKIE	-

[표 1.1] 웹로그 텍스트 파싱(parsing) 결과

웹로그 분석을 위한 데이터 전처리 단계

4) 추출한 패러미터 값을 조건으로 사진정보 DB를 조회하고 조회된 사진의 키 값으로 조회사진 항목집합(itemset)을 구성한다

```
-- t_picture: 사진정보 저장 테이블, t_pic_class: 사진 분류정보 저장 테이블
SELECT a.pic_no -- t_picture 테이블의 key 속성
FROM t_picture a, t_pic_class b
WHERE a.oid = :v_oid -- :v_oid = 웹로그 패러미터 oid 값 (3110)
AND a.pic_no = b.pic_no -- 조인조건 (b.pic_no는 a.pic_no를 참조하는 외래키
AND b.category = :v_category -- :v_category = 웹로그 패러미터 category 값 (1)
AND b.theme = :v_theme -- :v_theme = 웹로그 패러미터 theme 값 (t01)
```

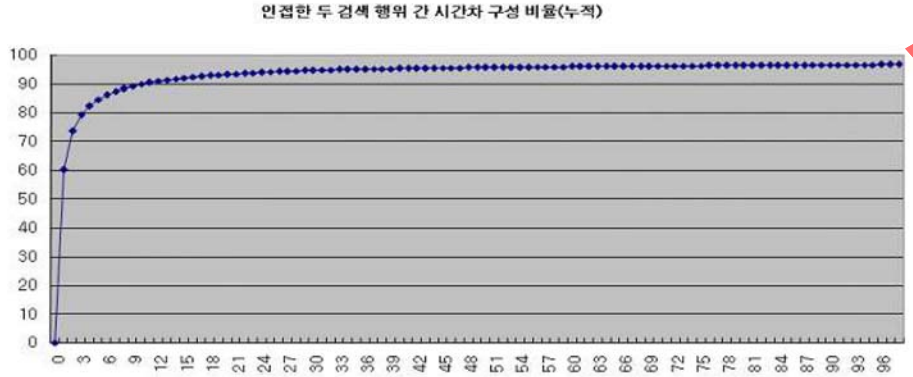
[그림 1.5] 웹로그 클라이언트 요청 URL의 패러미터 값을 통해 DB에 접근하는 사례

웹로그 분석을 위한 데이터 전처리 단계

5) 트랜잭션 설계

- 사진의 조회 연관성 분석을 위해서는 트랜잭션 설계가 필수적
  - ✓ 트랜잭션 내 조회사진 itemset에 대한 빈발항목을 구함
- 트랜잭션을 어떻게 설계할 것인가?
  - ✓ 트랜잭션 : 특정 사용자가 동시에 조회하는 사진 itemset
  - ✓ But, 웹로그는 트랜잭션의 개념이 없는 개별적인 웹서버 사용 기록
  - ➔ 로그 간 시간차 분석을 바탕으로 가상의 트랜잭션을 도출

예)

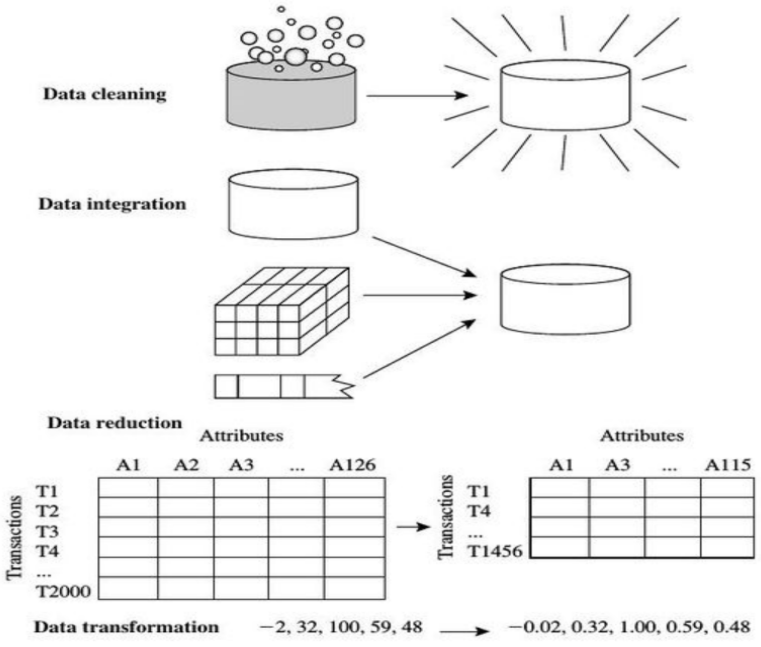


인접한 두 검색행위(로그) 시각 차가 OO분 이내이면 동일 트랜잭션으로 봄

데이터 전처리 주요 기법

- 데이터 정제(Data Cleansing)
  없는 데이터(missing values)는 채우고, 잡음(noisy data)는 제거하며, 모순된 데이터(inconsistent data)는 정합성이 맞는 데이터로 교정하는 작업
- 데이터 통합(Data Integration)
  여러 개의 데이터베이스(databases), 데이터큐브(data cubes), 또는 파일(files)을 통합하는 작업
- 데이터 축소(Data Reduction)
  샘플링(sampling) 등을 통해 데이터 볼륨(volume)을 줄이거나 분석대상 속성(차원)을 줄이는 작업
- 데이터 변환(Data Transformation)
  데이터 정규화(normalization) 또는 집단화(aggregation) 하는 작업
- 데이터 이산화(Data discretization)
  데이터 축소(data reduction)의 일종으로 연속적인 수치 데이터에 대한 구간화 작업 (예, 실제 나이를 10대, 20대, 30대 등으로 변환)

데이터 전처리 주요 기법



* 출처 : Jiawei Han and Micheline Kamber. Data Mining: Concepts & Techniques

[그림 1.6] 데이터 전처리 주요 기법의 개념 (concept)

학습의 방향

- 데이터 전처리 유형
  - 실무에 있는 데이터는 매우 다양한 형태로 존재하기 때문에 이를 특정 분석 목적에 맞게 가공하는 일은 사안마다 다름
  - 데이터 전처리를 개념적인 몇가지 분류 내로 국한시켜서는 분석 대상 데이터를 만들어내기 힘듬
  - 데이터 전처리는 개념적·이론적으로 제시되는 범위보다 훨씬 넓고 포괄적
- 학습의 방향
  - 모든 데이터 전처리 유형을 다루는 것은 애초부터 불가능 → 이론적으로도 제시되어 있으면서 실무에 자주 등장하는 전처리 유형을 중심으로 학습
  - 이론적 이해 + 실무 사례 적용을 통한 활용능력 배양
- 활용 도구(tools)
  - Python : 플랫폼 독립적이며 인터프리터식, 객체지향적, 동적 타이핑(dynamically typed) 대화형 언어로서 데이터프레임, 기계학습 등의 데이터분석을 위한 다양한 함수를 제공하여 데이터전처리에 적합한 도구
  - Oracle : Python의 SQL은 오라클의 SQL에서 제공하는 윈도우(window) 함수와 같은 데이터의 집합적 처리를 위한 강력한 기능은 제공되지 않은 부분이 있어서, 일부 사례에서는 오라클 SQL을 통해 해답을 제시

Ch.2 데이터 정제(Data Cleaning)

실무에서의 데이터는 비어 있거나(missing value), 오류값이 들어 있거나 정합성이 맞는 않는 경우가 많다. 불완전한 데이터로 분석 작업을 수행했을 때는 분석 결과 또한 신뢰성이 반감된다. 그러므로, 본격적인 분석 작업을 수행하기 전에 데이터의 불완전성을 최대한 제거하는 것이 필요하다.

2.1 결측값(missing value)의 처리

- 결측값(missing value)
  - 존재하지 않고 비어있는 상태
  - DB에서의 NULL값
- 결측값을 채우는 방법
  ① 해당 투플을 무시한다 (row-wise deletion)
  ② 결측값을 수동으로 채워넣는다
  ③ 전역상수(global constant)를 사용하여 결측값을 채워 넣는다
  ④ 속성의 평균값을 사용하여 결측값을 채워 넣는다
  ⑤ 주어진 투플과 같은 클래스(분류)에 속하는 투플 들의 속성 평균값을 사용한다
  ⑥ 가장 가능성이 높은 값(예측)으로 결측값을 채워 넣는다 (회귀분석, 베이지안기법, 의사결정트리 기법 등)

2.1 결측값(missing value)의 처리

결측값 처리 예제

A	B
A01	10
A01	
A01	20
A02	30

① 해당 투플을 무시 (row-wise deletion)

A	B
A01	10
A01	20
A02	30

③ 전역상수(global constant) 사용

A	B
A01	10
A01	**0**
A01	20
A02	30

④ 속성의 평균값 사용

A	B
A01	10
A01	**20**
A01	20
A02	30

2.1 결측값(missing value)의 처리

결측값 처리 예제

A	B
A01	10
A01	
A01	20
A02	30

A	B
A01	10
A01	15
A01	20
A02	30

⑤ 주어진 투플과 같은 클래스(분류)에 속하는 투플 들의 속성 평균값 사용

A	B
A01	10
A01	?
A01	20
A02	30

⑥ 가장 가능성이 높은 값(예측)으로 결측값을 채워 넣는다

2.1 결측값(missing value)의 처리 - 실무예제

다음은 부산광역시 사상구 약수터 수질 현황 표(검사일: 2015년 11월 5일)이다. 표상에 나타난 결측 속성값(일반세균, 질산성질소)을 채우시오.

- 데이터 파일: ch2-1(약수터 수질 현황).csv
- 원본 투플 수: 24개

연번	약수터명	동명	총대장균군	일반세균	질산성질소	적합
1	백수	모라	양 성	10	6.7	부적합
2	이칠	모라	음 성	20	0.9	적합
3	윤수사	모라	음 성	10	1.1	적합

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch2-1.ipynb 참고**

2.2 잡음(noisy data) 제거

- 잡음(noise)
  - 변수(속성)에서의 오류나 오차 값
  - 오류나 오차에 의한 값의 경향성 훼손을 줄이기 위해서 데이터 평활화 기법(smoothing technique)을 적용
- 데이터 평활화 기법
  - 구간화(Binning)
    정렬된 데이터 값들을 몇 개의 빈(혹은 버킷)으로 분할하여 평활화하는 방법
    ① 평균값 평활화(smoothing by bin means)
    ② 중앙값 평활화(smoothing by bin medians)
    ③ 경계값 평활화(smoothing by bin boundaries)
  - 구간화(Binning) 방식
    ① 동일 너비 방식
    ② 동일 높이 방식

2.2 잡음(noisy data) 제거

평활화 예제 [동일 너비(범위) 방식]

2.2 잡음(noisy data) 제거

평활화 예제 [동일 높이(개수) 방식]

	19	8	3	9	12	7	15	5	6	10	2	11

정렬(sorting)

	2	3	5	6	7	8	9	10	11	12	14	19

구간화(binning) [구간개수 4로 가정]

	2	3	5	6	7	8	9	10	11	12	14	19

평활화(smoothing)

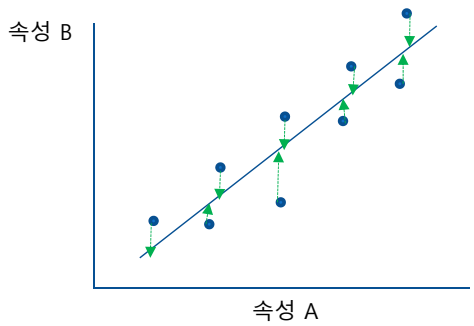
① 평균값 평활화	4	4	4	4	8.5	8.5	8.5	8.5	14	14	14	14
② 중앙값 평활화	4	4	4	4	8.5	8.5	8.5	8.5	13	13	13	13
③ 경계값 평활화	2	2	6	6	6	6	10	10	11	11	11	19
경계	2				6				10			19

2.2 잡음(noisy data) 제거

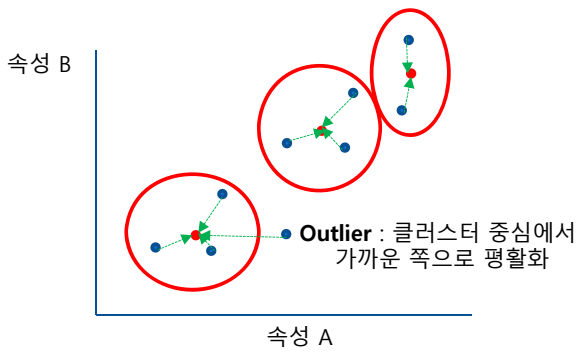
- 데이터 평활화 기법
  - 회귀(Regression)

    회귀 함수에 의한 데이터 평활화 기법
    - ✓ 선형회귀분석 : 하나의 속성값으로 다른 하나의 속성값 예측
    - ✓ 다중회귀분석 : 두 개 이상의 속성값으로 다른 속성값을 예측
  - 군집화(Clustering)

    유사한 값들끼지 그룹화하는 기법 (outlier는 평활화 대상)



< 회귀에 의한 평활화 >



< 군집화에 의한 평활화 >

2.2 잡음(noisy data) 제거 - 실무예제

다음은 2016년 항만별 선박 입출항 현황이다. 선박 수 속성은 구간화(Binning) 기법으로 평활화하고, 선박 톤수 속성은 선박 수 속성과의 회귀(regression) 분석을 통하여 평활화시키시오.

⊙ 데이터 파일 : ch2-2(선박입출항).csv

⊙ 원본 투플 수 : 30개

항만	입항선박수	입항선박톤수	출항선박수	출항선박톤수
부산	7,301	105,138,280	7,409	103,857,903
인천	2,715	30,716,710	2,716	30,779,186
평택.당진	1,558	23,153,226	1,536	22,778,109
경인항	28	126,236	27	128,344
동해.묵호	552	4,202,603	546	4,039,929

출처 공공데이터포털(www.data.go.kr)

☞ **해답은 ch2-2.ipynb 참고**

[여기서 잠깐!] PyCharm 실습환경 구축

- 왜 갑자기 PyCharm?
  - PyCharm : Python의 통합개발환경(IDE) 제공
  - PyCharm의 장점 : 디버깅(Debugging)
    - ✓ Python 코딩 시, 변수값의 변화를 보거나 오류를 잡을 때 유용
    - ✓ 복잡한 Python 로직(logic) 이해 시 도움
  - PyCharm 다운로드 및 설치
    - ① https://www.jetbrains.com/pycharm/ 사이트 접속
    - ② PyCharm Community ver2018.3.5 다운로드 (무료)
    - ③ PyCharm 설치
  - 데이터전처리 실습환경 구축
    - ① PyCharm 실행
    - ② 프로젝트(Project) 생성
    - ③ 예제 소스코드(*.py)와 데이터 파일(*.csv)을 프로젝트에 삽입
    - ④ 패키지(Package) 설치

2.3 데이터 정제 절차

1) 데이터의 모순점 포착

- 데이터의 모순점이 발생할 수 있는 주요 원인
  - ✓ 잘못 설계된 데이터 입력 폼이 존재
  - ✓ 데이터 입력에서 사람의 실수로 발생
  - ✓ 응답자가 자신의 정보가 누설되는 것을 원하지 않기에 발생하는 의도적인 오류
  - ✓ 바뀐 주소와 같이 만료된 데이터
  - ✓ 데이터 표현의 모순
  - ✓ 일치하지 않는 코드의 사용
  - ✓ 데이터를 기록하는 계측 장치의 오류나 시스템 오류
  - ✓ 원래의 의도와 다른 목적으로 데이터를 부적절하게 사용
  - ✓ 데이터 통합 과정에서 주어진 속성이 다른 데이터베이스에서 다른 이름을 갖고 있을 때
- 데이터의 모순점 파악을 위한 데이터 특성 파악 : 메타데이터의 활용
  - ✓ 속성의 데이터 타입과 도메인(속성값의 범위)
  - ✓ 속성값의 분포 특성(대칭, 비대칭 등)
    - 대칭 / 비대칭 분포
    - 실제값의 주요 분포 범위
    - 값의 표준편차

2.3 데이터 정제 절차

  - ✓ 속성 간의 의존성
    - 속성 A의 값이 같은 투플의 속성 B 값이 반드시 같다면, 속성 A와 속성 B 간의 함수적 종속성이 존재한다고 말하며, A → B로 표시
- 필드 오버로딩(Field Overloading)에 대한 검토 원칙
  - ✓ 유일 규칙(Unique Rule)
  - ✓ 일관 규칙(Consecutive Rule)
  - ✓ 무 규칙(Null Rule)
- 모순점 포착을 위한 상용 도구
  - ✓ 데이터 세정 도구 (Data Scrubbing Tool)
  - ✓ 데이터 감사 도구(Data Auditing Tool)
  - ✓ 데이터 이관 도구(Data Migration Tool)
  - ✓ ETL(Extraction/Transformation/Loading) 도구

2.3 데이터 정제 절차

2) 모순점 포착과 데이터 변환의 반복

- 모순점이 발견된 데이터에 대한 변환 과정의 반복
- 방식
  - ✓ 일괄 처리(batch process)
  - ✓ 상호 작용(interactive process)
  - ✓ SQL과 같은 선언형 언어 활용

연습문제

1. 다음은 서울시 아파트단지별 관리비 정보이다. 표 상에 나타난 결측 속성값(총공용관리비용, 총세대사용비용, 총장기구선충당금)을 채우시오.

◪ 데이터 원본 : (연습2-1)서울시아파트관리비정보.xlsx ◪ 원본 튜플수 : 111개

단지 명	년	월	총공용관리비용	총세대사용비용	총장기구선충당금
개포우성9차	2014	6	31093597	20222590	3675900
개포우성9차	2014	5	30249840	19212810	3675900
개포우성9차	2014	4	26256330	29810000	3675900
개포우성9차	2014	3	29351185	35986240	3675900
개포현대3차	2015	9	44356086	18932950	6757560
개포주공7단지	2015	9	116398370	58013730	9171050
개포주공3단지	2015	9	61083000		372000
개포주공4단지	2015	8	96033080		
개포주공7단지	2015	8	104489670	49358300	9171050
개포주공3단지	2015	8	32953800		372000
개포현대3차	2015	7	45746367	23308910	6757560
개포주공7단지	2015	7	105618070	54129240	9171050

(1)

* 출처 : 서울시 열린데이터광장(data.seoul.go.kr)

2. 1번문제에서 결측값을 채운 후에 총공용관리비용 속성은 구간화(Binning) 기법으로 평활화시키고, 총세대사용비용 속성은 총공용관리비용 속성과의 회귀(regression) 분석을 통하여 평활화시키시오. (2)

Ch.3 데이터 통합(Data Integration)

데이터 통합 (Data Integration)은 여러 데이터 저장소로부터 온 데이터들의 합병을 의미한다. 데이터웨어하우스(data warehouse)나 데이터마이닝과 같은 데이터 분석 작업은 다수의 데이터 원천으로부터 데이터를 하나의 통일된 데이터 저장소로 결합시키는 데이터 통합 작업을 필요로 한다. 데이터 원천은 데이터베이스나 데이터 큐브, 플랫 파일(flat file) 등 다양한 형태로 존재한다. 이제부터 데이터 통합에서 고려해야 할 몇가지 사항에 대하여 살펴보도록 한다.

3.1 개체의 식별

- 개체 식별 문제(Entity Identification Problem)
  - 데이터 통합 시, 동일한 의미의 개체들이 서로 다르게 표현되어 있는 문제 → 어떻게 일치시킬 것인가?
    - ✓ 예) A 데이터베이스에서 customer.customer_id vs. B 데이터베이스에서 cust.cust_number
  - 메터데이터의 역할이 중요
    - ✓ customer_id는 customer 테이블의 PK(Primary Key, 기본키)이고 cust_number는 cust 테이블의 PK이며, 두 속성 다 동일한 데이터 타입과 도메인을 가지고 있다고 한다면, 두 속성은 이름을 다르지만 동일한 속성으로 판단할 수 있음
    - ✓ 일반적으로, 속성의 데이터 타입이나 도메인 뿐만 아니라 기본키 여부, 참조무결성(외래키) 관계, 함수적 종속 관계(functional dependancy) 등을 종합적으로 고려하여 속성의 동일성 여부를 판단해야 함
  - 메터데이터는 데이터변환에도 도움을 줄 수 있음
    - ✓ 예) A 데이터베이스의 성별코드 'M', 'F'vs. B 데이터베이스의 성별코드 '1', '2'
    - ✓ 메터데이터 정보를 이용하여 어느 한 쪽의 데이터변환 필요

3.2 중복

- 중복(Redundancy)
  - 유도속성(derived attribute)
    - ✓ 예) 생년월일과 연령, 월소득과 연간소득, 과목점수와 총점
    - ✓ 월소득 속성값 100만원 vs. 연간소득 속성값 1000만원 ? → 어느 쪽이 틀린 것인가?
  - 정규화되지 않은 테이블
    - ✓ 조회 성능 향상을 위해 일부러 정규화하지 않고 중복 허용 → 일관성 저해의 문제 야기
    - ✓ 예) 구매 테이블 : {구매자번호, 구매일시, **주소**, 전화번호, 구매품목}
    - ✓ 동일 구매자번호에 대해서 다른 주소가 존재할 가능성
  - 중복 문제의 해결은 데이터 정제의 영역으로 어떠한 절대적인 해결책이 있다라기보다는 데이터 정제 시에 가장 데이터 정확성을 높이는 방향으로 정제 룰(cleansing rule)을 정의하여 일괄 적용할 수 밖에 없음

3.3 상관분석

- 상관분석을 통한 중복 탐지
  - 속성 간에 엄격한 함수적 종속 관계가 성립하지는 않지만, 상관분석을 통해서 한 속성이 다른 속성을 얼마나 강하게 암시하는지를 사용 가능한 데이터를 토대로 측정할 수 있음
  - 두 속성 간에 상관도가 높다면 두 속성을 중복으로 보고 그 중 하나의 속성을 제거할 수 있음

3.3.1 수치형 데이터 : 상관계수(correlation coefficient)

- 수치 속성에 대하여 속성 A와 B의 상관계수

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B} \quad \text{식(3-1)}$$

- N : 투플의 개수
- a_i, b_i : 투플 i에서의 속성 A, B의 값
- $\bar{A}$, $\bar{B}$: 속성 A, B의 평균값
- σ_A, σ_E : 속성 A, B의 표준편차

중복 속성 여부를 판단할 때는 해당 분야 도메인 지식(knowledge)을 충분히 고려해서 최종 판단하는 것이 바람직함

상관계수 결과 값의 범위는 −1에서 +1 사이를 만족한다.($-1 \le r_{A,B} \le +1$) 상관계수 $r_{A,B}$의 해석은 다음과 같다

- $r_{A,B} \ge 0$

속성 A, B는 양의 상관관계(positively correlated)를 가진다. 즉, B 값이 증가함에 따라서 A의 값이 증가한다.

- $r_{A,B} \le 0$

속성 A, B는 음의 상관관계(negatively correlated)를 가진다. 즉, B 값이 증가함에 따라서 A의 값은 감소한다.

- $r_{A,B} = 0$

속성 A, B는 독립적이며 둘 사이에 상관관계가 없다.

3.3.1 수치형 데이터 : 상관계수 - 실무예제

다음은 2013년 전국 주요 지점별 유동 인구 현황의 일부이다. 남자 20대 vs. 여자 20대, 남자 10대 vs. 여자 50대의 상관계수를 구하여 비교하고, 중복 속성으로 판단할 수 있을지 검토해 보시오.

⊙ 데이터 파일 : ch3-1(유동인구수).csv

⊙ 원본 투플 수 : 23,221개

조사일자	시간대	X좌표	Y좌표	행정구역명	남자 10대	남자 20대	남자 30대	남자 40대	남자 50대	여자 10대	여자 20대	여자 30대	여자 40대	여자 50대
2010-06-21	12시~13시까지	343099	417482	대전광역시 서구 월평동	2	24	68	50	31	4	37	64	44	26
2010-06-21	19시~20시까지	343099	417482	대전광역시 서구 월평동	19	44	28	33	21	14	56	49	43	18
2010-06-20	12시~13시까지	343099	417482	대전광역시 서구 월평동	13	33	34	61	55	13	32	29	28	12
2010-06-20	19시~20시까지	343099	417482	대전광역시 서구 월평동	23	33	32	547	129	12	39	13	46	4
2010-06-21	12시~13시까지	343121	417343	대전광역시 서구 월평동	0	9	27	21	6	5	24	20	10	6

출처: 공공데이터포털(www.data.go.kr)

☞ 해답은 ch3-1.ipynb 참고

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

- 범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨(Pearson)의 카이제곱(χ^2) 검정에 의해 측정

범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨(Pearson)의 카이제곱(χ^2)검정에 의해 측정될 수 있다. 속성 A가 c개의 범주 값 $a_1, a_2, ..., a_c$를 취하고, 속성 B는 r개의 범주 값 $b_1, b_2, ..., b_r$을 취한다고 가정하자. 그러면, 속성 A와 B에 의해 구성되는 투플은 c개의 열과 r개의 행으로 구성되는 분할표로 표현될 수 있다. (A_i, B_j)를 속성 A가 a_i를 취하고 속성 B가 b_j를 취하는 투플이라고 할 때, χ^2은 다음과 같이 정의된다.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} = \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad \text{식(3-2)}$$

- o_{ij} : (A_i, B_j)에 대한 관측도수(observed frequency; 실제로 존재하는 (A_i, B_j) 투플 수)
- e_{ij} : (A_i, B_j)에 대한 기대도수(expected frequency; 확률적으로 기대되는 (A_i, B_j) 투플 수)

e_{ij}는 다음과 같이 계산된다.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N} \qquad \text{식(3-3)}$$

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

식 (3-2)에서의 합계는 분할표상의 $r \times c$ 개의 모든 셀에 대하여 계산된다. 따라서 χ^2 값에 가장 크게 기여하는 칸은 실제 관측도수와 기대도수의 차이가 매우 큰 셀이다.

이러한 χ^2 통계량은 속성 A와 B가 독립이라는 가설을 검증한다. 이 검정은 자유도 $(r-1)\times(c-1)$을 갖는 유의수준에 근거하여 검정한다.

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

- 카이제곱 통계량을 이용한 범주형 속성에 대한 상관분석 사례

어떤 설문조사에서 1,500명의 사람들을 대상으로 각 사람에 대한 성별과 선호 글의 픽션 여부 간의 상관관계를 분석하고자 한다. 성별 속성값은 'male', 'female'이 있고, 픽션 여부 속성값은 'fiction'과 'non-fiction'이 있다. 성별 속성값 분포는 male:female=300:1,200이고 픽션 여부 속성값 분포는 fiction:non-fiction=450:1,050이다. 이들 속성의 조합에 대한 관측도수를 기록한 2×2 분할표는 다음과 같다.

	male	female	Total
fiction	250	200	450
non–fiction	50	1000	1050
Total	300	1200	1500

표에서 각 셀에 대한 기대도수는 식 (3-3)에 의해 구할 수 있다. 예를 들어, 셀 (female, fiction)의 기대도수는

$$e_{12} = \frac{count(femal) \times count(fiction)}{N} = \frac{1200 \times 450}{1500} = 360$$

이와 같은 방식으로 다른 셀에 대한 기대도수를 구하면 다음과 같은 기대도수 분할표를 얻는다.

	male	female	Total
fiction	90	360	450
non–fiction	210	840	1050
Total	300	1200	1500

이제 식 (3-2)에 의해서 χ^2 통계량을 구하면 다음과 같다.

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93$$

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

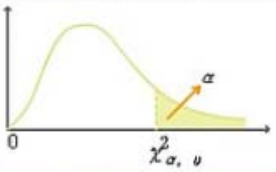
- 카이제곱 통계량을 이용한 범주형 속성에 대한 상관분석 사례

χ^2 통계량을 산출하였으므로 카이제곱 검정 방식에 의해 유의수준 0.05 수준에서의 두 속성 사이에 연관성이 있다라는 가설을 검증해 보자. 본 예제에서의 자유도는 (2−1)×(2−1)=1이 된다. 자유도 1일 때, χ^2 통계량(507.93)에 대한 유의확률(p-value)은 2.2e−16이다(이 수치는 자유도 1일 때의 카이제곱분포로 획득 가능함). 이는 유의수준 0.05보다 훨씬 작은 값으로서 대립가설(두 속성은 연관성이 있다)[7]은 채택된다. 자유도 1일 때 유의수준 0.05로 대립가설을 기각하는 데 필요한 값은 3.842로서 이 값보다 작아야 가설이 기각되는데, χ^2 통계량은 507.93으로 3.842에 비해 매우 크므로 가설은 채택되고, 실제 두 속성 사이에는 강한 연관성이 있다고 결론지을 수 있다.

7) 상관분석 검정의 가설은 크게 귀무가설(두 속성은 연관성이 없다)과 대립가설(두 속성은 연관성이 있다)로 나누어진다.

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

<카이제곱 분포표>



유의확률(p-value)

자유도(dof) ν	α=.995	α=.99	α=.975	α=.95	α=.05	α=.025	α=.01	α=.005	ν
1	.3333930	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597	2
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838	3
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860	4
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750	5
6	.676	.872	1.237	1.635	13.592	14.449	16.812	18.548	6
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.516	18
19	6.844	7.633	8.907	10.117	30.114	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정 - 실무예제

다음은 2013년 전라남도 유망중소기업 지정업체 명단의 일부이다. 시군 속성과 지정구분 속성 간의 연관성 여부를 카이제곱 검정 방법에 의해 판단해 보시오(단, 유의 수준은 0.05이다).

⊙ 데이터 파일 : ch3-2(유망중소기업현황).csv

⊙ 원본 투플 수 : 386개

연번	시군	지정구분	기업명	대표자	소 재 지	주생산품	전화번호(061)	비고 (지정번호)
1	목포시	기술유망	브로드컴㈜	이동헌	목포시 석현동 1175(벤처지원 202)	정보통신	284-0017	11월 01일
2	목포시	기술유망	㈜케이에스	김시오	목포시 연산동1236-3	융합철망외	1588-4118	11월 02일
3	목포시	기술유망	삼진물산(주)	김관석	목포시 연산동 1239-1	잡지통조림	270-6113	11월 03일
4	목포시	기술유망	㈜해성	전재두	목포시 연산동1237-3	가드레일외	1588-2811	11월 04일
5	목포시	기술유망	(유)한국메이드	이승룡	목포시 연산동 1238-4	선박블록	278-4411	11월 05일
6	목포시	수출유망	환길산업	박승남	목포시 산정동 1780-1	해조류	272-7147	11월 06일

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch3-2.ipynb 참고**

3.3.3 데이터 값 충돌의 탐지와 해결

- 데이터값 충돌
  - 서로 다른 데이터 원천으로부터 온 데이터들의 통합 시에는 동일한 개체에 대해서도 속성값들이 서로 다를 수 있음 → 데이터 통합 시에 기준을 정해서 그 기준에 따라 데이터값을 변환시켜서 통합시키는 것이 필요함
- 데이터 통합 시, 데이터구조에 주의
  - 원천 시스템의 기능적 종속성과 제약사항들이 목표 통합 시스템의 그것들과 일치하도록 해야 함 → 통합된 후에도 관련 규칙(rule)들이 적용될 수 있도록
  - 예) A 시스템에서는 할인이 총 주문금액에 적용되는 반면, B 시스템에서는 할인이 개별 항목에 적용됨

연습문제

1. 다음은 서울메트로 동대문역 기준 시간대별 승하차인원현황이다. 승차와 하차를 구분하여 비슷한 패턴을 보이는 시간대를 찾아보고 속성 통합 가능 여부를 판단해 보시오.

◪ 데이터 원본 : (연습3-1)동대문역시간대별승하차인원현황.xlsx ◪ 원본 튜플수 : 702개

날짜	호선	역명	구분	05-06	06-07	07-08	08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20
2014-02-24 (월)	1호선	동대문(155)	승차	517	407	772	1006	893	678	817	952	1072	1144	1148	1227	1232	1185	917
2014-02-24 (월)	1호선	동대문(155)	하차	74	296	515	1445	1173	1090	1192	1396	1267	1202	1172	1090	1018	1124	1320
2014-02-25 (화)	1호선	동대문(155)	승차	452	443	684	1003	879	736	798	1043	1139	1257	1240	1211	1274	1388	981
2014-02-25 (화)	1호선	동대문(155)	하차	92	278	504	1399	1181	1149	1230	1286	1464	1261	1281	1301	1104	1289	1419
2014-02-26 (수)	1호선	동대문(155)	승차	482	408	719	1030	837	759	771	1036	1138	1264	1157	1234	1310	1280	950
2014-02-26 (수)	1호선	동대문(155)	하차	60	280	537	1433	1196	1194	1237	1413	1316	1145	1308	1125	1090	1213	1400
2014-02-27 (목)	1호선	동대문(155)	승차	502	398	686	1010	858	714	827	1024	1020	1122	1175	1158	1287	1298	962
2014-02-27 (목)	1호선	동대문(155)	하차	67	337	524	1366	1204	1177	1211	1359	1363	1306	1320	1079	1084	1196	1324
2014-02-28 (금)	1호선	동대문(155)	승차	491	410	664	994	873	713	818	1081	1116	1196	1189	1205	1311	1446	1009

(3)

* 출처 : 서울시 열린데이터광장(data.seoul.go.kr)

2. 다음은 부산광역시 어린이집 현황의 일부이다. 어린이집유형 속성과 어린이집특성 속성 간의 연관성 여부를 카이제곱 검정 방법에 의해 판단해 보시오.

◪ 데이터 원본 : (연습3-2)부산시어린이집현황.csv ◪ 원본 튜플수 : 206개

No	어린이집유형	어린이집	어린이집특성	평가인증여부	정원	아동현원	보육교직원현원
1	민간	21세기어린이집	일반	Y	70	70	10
2	국공립	ymca어린이집	장애아통합	Y	60	60	11
3	가정	가담어린이집	일반	N	14	11	3
4	민간	가빈어린이집	일반	N	59	50	14
5	가정	거목어린이집	일반	Y	20	20	5
6	민간	건강한영아전담어린이집	영아전담	Y	32	32	10
7	민간	과학나라어린이집	일반	N	39	21	6
8	민간	구름나무어린이집	일반	N	32	29	7
9	법인·단체등	구포회광어린이집	일반	Y	49	45	7

(4)

* 출처 : 공공데이터포털(www.data.go.kr)

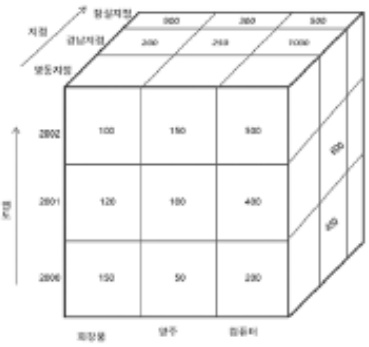
Ch.4 데이터 축소(Data Reduction)

방대한 양의 데이터를 대상으로 복잡하게 데이터를 분석하고 마이닝 기법을 적용한다면 매우 많은 시간이 소요되어 분석이 비현실적일 수 있다. 데이터 축소(data reduction) 전략에는 크게 차원적 축소(dimensionality reduction)와 수치적 축소(numerosity reduction)가 있다. 차원적 축소는 데이터 인코딩 스키마를 적용하여 압축되거나 축소된 표현을 제공함을 의미한다. 수치적 축소는 모수적 모형(parametric model)이나 비모수적 모형(non-parametric model)을 사용하여 데이터를 대체함을 의미한다. 어떠한 데이터 축소 기법을 적용하든 대전제는 축소된 데이터집합에 대한 데이터 분석 결과는 원본 데이터집합에 대한 데이터 분석 결과와 거의 동일한 결과를 산출해야 한다는 것이다.

4.1 데이터큐브 집계

- 데이터큐브(Data Cube)
  - 데이터웨어하우스(DW)에서 나오는 용어로서 다차원의 집계 정보를 의미함
  - 집단화된 데이터(aggregated data) 저장
  - 원천 데이터를 여러 관점에서의 추상화시켜 데이터 축소를 구현

연도	지점="명동지점" 품목별 총판매금액			지점="강남지점" 품목별 총판매금액			지점="잠실지점" 품목별 총판매금액		
	화장품	양주	컴퓨터	화장품	양주	컴퓨터	화장품	양주	컴퓨터
2002	100	150	500	200	250	1000	900	300	500
2001	120	100	400	140	200	600	600	290	450
2000	150	50	200	160	100	450	200	300	200



[그림 4.1] 데이터큐브의 예

4.1 데이터큐브 집계

- 데이터큐브의 다양한 추상화 레벨
  - 기본 큐보이드(base cuboid) : 최하위 추상 레벨에서 만들어지는 큐브
  - 정점 큐보이드(apex cuboid) : 최상위 추상 레벨에서 만들어지는 큐브
  - 데이터 큐브는 사전계산(precomputed)되고, 요약된(summerized) 데이터에 신속한 접근이 장점 → 데이터마이닝과 같은 분석처리 가능

4.1 데이터큐브 집계 - 실무예제

다음은 2015년 국내 대학 현황이다. 데이터큐브 집계를 이용하여 학제별/지역별/설립별로 재학생 수 합계로 데이터를 축소시키시오.

- 데이터 파일 : ch4-1(국내대학현황).csv
- 원본 투플 수 : 1,930개

학제	학교명	지역	설립	재적학생수	재학생수	휴학생수	총장및전임교원수
전문대학	한국철도대학	경기	국립	180	134	46	-
전문대학	강원도립대학	강원	공립	1,658	900	758	31
전문대학	경남도립거창대학	경남	공립	1,614	955	659	31
전문대학	경남도립남해대학	경남	공립	1,639	846	793	25
전문대학	경북도립대학교	경북	공립	1,389	930	459	32

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch4-1.ipynb 참고**

4.2 속성 부분집합 선택

- 속성 부분집합 선택(Attribute Subset Selection)
  - 연관성이 낮거나 중복되는 데이터 속성을 제거하여 데이터 집합의 크기를 줄이는 기법
  - 데이터 분석에 영향을 미치지 않거나, 중복적 성격의 속성을 충분히 제거하지 않을 경우, 분석결과 품질 저하의 우려가 있고, 분석 성능에도 악영향을 미침
  - 목표는 전체 속성에 가장 가까운 데이터 범주의 확률 분포와 최소의 속성 집합을 찾는 것
- 최소 속성집합을 찾는 기법
  - 소모적 탐색법(exhausted search) : n개의 속성에 대해 2^n개의 가능한 속성 조합을 모두 탐색 → n이 증가할수록 엄청난 비용을 발생시키지 때문에 비현실적
  - 경험적 탐색법(heuristic search) : 속성 공간을 탐색하는 동안에 매 회마다 최선으로 보이는 것을 선택 (지역적으로 최적의 해를 찾는 것으로 탐욕적(greedy)라고도 함)
  - 속성 평가척도 활용 : 분류(classification)을 위한 의사결정트리(decision tree) 생성에 사용되는 정보이득 (information gain) 등

4.2 속성 부분집합 선택

- 경험적 기법

1) 단계적 전진선택법(stepwise forward selection)

속성의 공집합으로 시작해서 최적의 속성들을 하나씩 추가하는 방법

2) 단계적 후진제거법(stepwise backward selection)

속성의 전체 집합으로 시작해서 최악의 속성들을 하나씩 제거하는 방법

3) 전진선택법과 후진제거법의 결합(combination of forward selection and backward elimination)

4) 의사결정트리 귀납법(decision tree induction)

- ✓ 데이터마이닝 기법 중 분류(classification)을 위해 고안됨
- ✓ 트리 구성 요소
  - 비단말 노드 (non-leaf) : 속성에 대한 테스트
  - 가지 (branch) : 속성 테스트 결과에 따른 흐름
  - 단말 노드 (leaf) : 클래스 예측
- ✓ 데이터를 개별적 클래스로 분할하기 위한 최선의 속성을 각 노드에서 고르는 것
- ✓ 트리에 나타나지 않은 모든 속성은 부적절한 것으로 판단되고, 트리에 나타나는 속성들의 집합이 축소된 속성집합임

4.2 속성 부분집합 선택

- 다중공선성(Multicolinearity) 분석
  - 다중공선성 : 독립변수들 간의 상관정도가 높은 상태
  - 다중공선성의 문제점
    - ✓ 회귀분석 시 : 각 독립변수의 회귀계수가 종속변수에 미치는 영향을 제대로 설명하지 못함
    - ✓ 분류 시 : 종속변수에 중요한 영향을 미치는 독립변수가 결정트리의 분류 조건에서 누락될 가능성
  - 상관성 vs. 다중공선성

	상관성	다중공선성
특징	• 두 변수(속성) 간 상관정도 측정 • 독립변수와 종속변수를 구분하지 않음	• 두개 이상의 변수들 간의 상관정도 측정 • 독립변수들 간의 상관정도 측정
측정 방법	• Pearson Correlation	• VIF (분산팽창계수) • Tolerance (공차한계) • CN(상태지수)

  - 다중공선성 존재 판단 기준 (일반적 기준)
    - ✓ Torerance(공차한계) : 0.1 이하
    - ✓ VIF(분산팽창계수) : 10이상
    - ✓ CN(상태지수) : 100이상
    - ✓ eigen value(고유값) : 0.01이하

4.2 속성 부분집합 선택

- 다중공선성(Multicolinearity) 분석
  - 다중공선성 측정 방법 : OLS(Ordinary Least Squares) 회귀분석
    - ✓ OLS(최소제곱법) : 오차의 제곱의 합을 최소화하는 기법

• 분산팽창계수
• 10이상이면 다중공선성 의심

	VIF Factor	features
0	1338.2	Intercept
1	3.2	성별코드
2	1.7	연령대코드
3	3.4	신장
4	3.8	체중

• 다중 회귀모형의 설명력
• Recommended : 0.4 이상

• 회귀계수
• 종속변수에 대한 영향력

• 유의확률
• Recommended : 0.05 미만

• 독립변수들의 독립성
• Recommended : 1.5 ~ 2.5

OLS Regression Results

Dep. Variable:	height	R-squared:	0.589
Model:	OLS	Adj. R-squared:	0.571
Method:	Least Squares	F-statistic:	31.57
Date:	Sun, 03 Mar 2019	Prob (F-statistic):	1.20e-05
Time:	19:53:45	Log-Likelihood:	-63.655
No. Observations:	24		
Df Residuals:	22		
Df Model:	1		
ovariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	107.8624	11.816	9.128	0.000	83.357	132.368
weight	0.8904	0.158	5.619	0.000	0.562	1.219

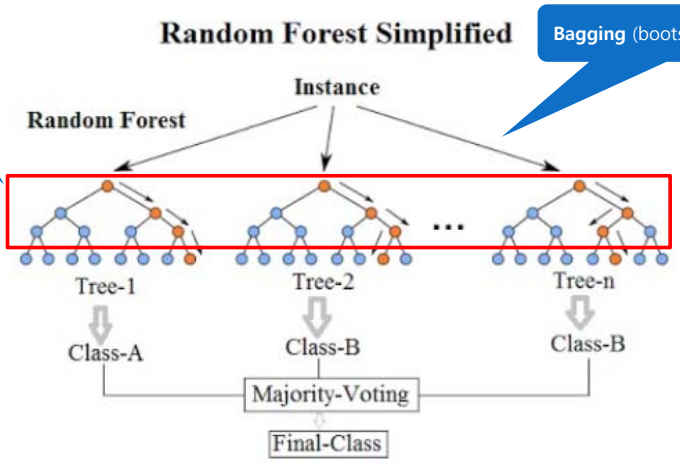
Omnibus:	0.796	Durbin-Watson:	2.201
Prob(Omnibus):	0.672	Jarque-Bera (JB):	0.829
Skew:	0.329	Prob(JB):	0.661
Kurtosis:	2.371	Cond. No.	1.20e+03

- 속성 부분집합 선택
  - ✓ 단계적 전진 선택법, 단계적 후진 제거법
  - ✓ (전진 선택법 + 후진 제거법) : 전진 선택법에 의해 독립변수를 추가한 뒤, 다중공선성을 측정하여 변수를 제거하는 방식

4.2 속성 부분집합 선택

- 의사결정트리 귀납법에 대한 Alternatives
  - Random Forest 기법에 의한 속성(feature) 중요도 측정
    - ✓ Random Forest : 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로 훈련 과정에서 구성한 다수의 결정 트리로 부터 투표(voting)을 통하여 분류하거나 예측치 산출

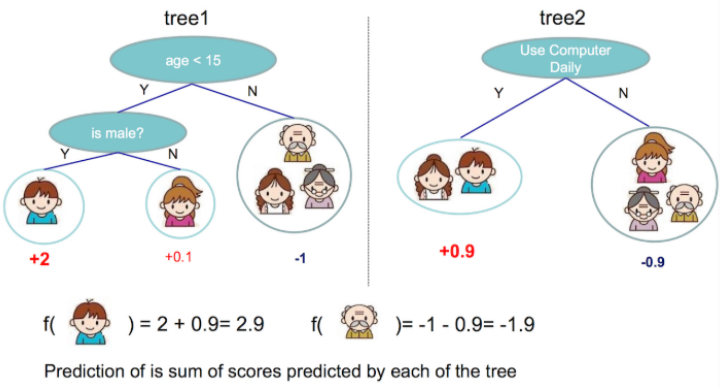
속성(feature) 중요도 측정



- ✓ 속성 부분집합 선택 : 속성 중요도를 측정하여 중요도 수치가 높은 n개의 속성 선택 (전진 선택법)

4.2 속성 부분집합 선택

- 의사결정트리 귀납법에 대한 Alternatives
  - XGBoost 기법에 의한 속성(feature) 중요도 측정
    - ✓ XGBoost (eXtreme Gradient Boost) : 분석 모델 성능 향상을 위해 gradient boosting decision tree 알고리듬 사용
      - Boosting : 이전 모델들의 오차들을 교정하기 위해 새로운 모델을 추가하는 앙상블 테크닉
      - Gradient Boosting : 새로운 모델을 더할 때, 오차 최소화를 위해 gradient descent 알고리듬 사용



    - ✓ 속성 부분집합 선택 : 속성 중요도를 측정하여 중요도 수치가 높은 n개의 속성 선택 (전진 선택법)

4.2 속성 부분집합 선택

- 의사결정트리 귀납법에 대한 Alternatives
  - RFE (Recursive Feature Elimination) – 재귀적 속성 제거
    - ✓ 모든 특성으로 시작하여 모델을 만들고, 속성 중요도(feature importances)가 가장 낮은 특성을 제거하는 방식 (후진 제거법)
    - ✓ 처리 절차
      - ① 모든 특성을 대상으로 분석 모델을 만든다
      - ② 속성 중요도가 가장 낮은 속성을 제거한다
      - ③ 나머지 속성들로 분석 모델을 만든다
      - ④ 지정한 속성 개수만 남을 때 ②~③ 과정을 반복한다
    - ✓ 부분집합 속성 개수의 결정
      - 모든 특성을 대상으로 분석한 결과의 정확도와 최대한 근접하는 속성 수

4.2 속성 부분집합 선택

다음은 붓꽃(iris)에 대한 꽃받침 길이/너비, 꽃잎 길이/너비, 품종에 관한 데이터의 일부이다. 품종 판별에 중요한 영향을 미치는 속성의 부분집합을 의사결정트리 귀납법을 적용하여 선택하시오.

- 데이터 파일: ch4-2(붓꽃데이터).csv
- 원본 투플 수: 150개

일련번호	꽃받침길이	꽃받침너비	꽃잎길이	꽃잎너비	품종
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa

☞ **해답은 ch4-2.ipynb 참고**

4.3 차원 축소

- 차원 축소 (dimensionality reduction)
  - 원천 데이터의 축소판(압축판)을 얻기 위하여 데이터 부호화나 변환을 적용하는 것
  - 무손실(loseless) vs. 손실(lossy)
  - 대표적 차원축소 기법 : 웨이브렛 변환(wavelet transform)과 주성분분석(principal components analysis)

4.3.1 웨이블릿 변환

- 이산 웨이블릿 변환(DWT; discrete wavelet transform)
  - 데이터 벡터 X를 다른 수치적 벡터(Numerically Vector)인 X'으로 변환하는 것을 의미 (두 벡터 X와 X'의 길이(속성수)는 같음)
  - 같은 길이에도 데이터 축소로 볼 수 있는 것은 변환 데이터가 압축되어 보여지기 때문 (웨이블릿 계수 중에서 가장 유력한 것들의 일부만을 저장함으로써 데이터의 근사치를 유지) → 데이터 희소성(data sparsity)은 데이터 연산의 복잡도를 크게 감소시킬 수 있음
  - 데이터의 주요 특징들을 보존하면서도 잡음을 제거하는 역할을 하기도 하므로, 데이터 정제를 위해서도 효과적
- DWT 절차
  - 각 반복 때마다 데이터를 반으로 나눠서 계산 속도를 향상시키는 피라미드 알고리즘(pyramid algorithm)을 사용
  - 웨이블릿 원형함수(모 웨이블릿) : 데이터 변환 과정에서 적용
    - ✓ 고주파 버전 : 시계열 분석에 사용
    - ✓ 저주파 버전 : 빈도 분석에 사용

4.3.1 웨이블릿 변환

- 피라미드 알고리즘

1) 입력 벡터의 길이 L(L>=n)을 2의 정수 제곱으로 만들기 위해 필요한 만큼 0을 패딩으로 채운다.
2) 각 변환에 2개의 함수를 적용한다. 첫 번째는 합이나 가중평균을 적용한 데이터 평활화이고, 두 번째는 데이터의 세부적인 특성을 두드러지게 하는 가중차(weighted difference)의 계산이다.
3) 데이터 벡터 X의 두 데이터 포인트 쌍인 (x2i, x2i+1)에 2개의 함수가 적용되고, 그 결과로 길이가 L/2인 두 데이터 셋을 만든다. 하나는 빈도 분석에 사용되는 저주파 버전, 다른 하나는 시계열 분석에 사용되는 고주파 버전이다.
4) 앞선 과정을 길이 L이 2가 될 때까지 반복한다.
5) 위의 반복에서 구해진 데이터 집합으로부터 선택된 값은 변형된 데이터의 웨이블릿 계수로 지정된다.

4.3.1 웨이블릿 변환 - 실무예제

실무 예제 4-3

다음은 2014년도 서울시 광진구의 시간대별 대기오염 측정 결과의 일부이다. SO_2, NO_2, CO, O_3, PM10을 측정치 대상으로 웨이블릿 변환하여 데이터를 축소시키시오.

⊙ 데이터 파일 : ch4-3(대기오염도측정).csv

⊙ 원본 투플 수 : 7,648개

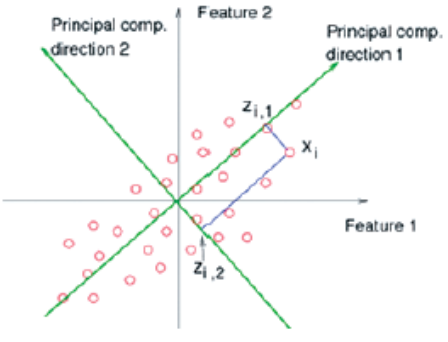
년	월	일	시	SO2	NO2	CO	O3	PM10
2014	1	1	1	0.01	0.037	1	0.005	179
2014	1	1	2	0.009	0.037	1	0.005	179
2014	1	1	3	0.008	0.037	1.1	0.004	181
2014	1	1	4	0.006	0.038	1.2	0.004	167
2014	1	1	5	0.006	0.039	1.2	0.004	149

출처: 서울시 열린데이터광장(data.seoul.go.kr)

☞ **해답은 ch4-3.ipynb 참고**

4.3.2 주성분 분석

- 주성분 분석(PCA; Principal Components Analysis)
  - n개의 속성을 가진 투플(n차원의 데이터 벡터)에 대하여, 데이터를 표현하는데 최적으로 사용되어질 수 있는 k(k≤n)개의 직교벡터(orthogonal vector)들을 찾음
  - 원본 데이터가 훨씬 작은 차원의 공간으로 투영되어지므로 데이터의 차원축소가 이루어짐
  - 속성 부분집합 선택은 속성의 초기 집합의 부분집합을 유지하며 속성 집합의 크기를 줄이는 반면, PCA는 필수적인 속성들의 핵심을 결합임
  - 종종 기대하지 않았던 관계를 보여주기도 하여 평범하지 않은 결과 해석을 가능하게 함



출처: PGWiki(http://www.potatogim.net/wiki)

[그림 4.2] PCA 그래프의 예

4.3.2 주성분 분석

- 주성분 분석 절차

1) 입력 데이터를 표준화하여 같은 범위에 속하게 만든다. 표준화하는 이유는 큰 범위를 갖는 속성들이 작은 범위를 갖는 속성들을 압도하지 않도록 하기 위해서이다.
2) 표준화된 입력 데이터를 위한 기저(base)를 제공하는 정직교(or-thonormal) 벡터들을 계산한다. 이들을 주성분(Principal Component)라고 하며, 입력 데이터는 주성분의 선형 조합(linear combination)이다
3) 주성분은 중요도의 내림차순으로 정렬되며, 본질적으로 주성분은 데이터에 대한 새로운 축의 집합으로서의 역할을 한다. 즉, 정렬된 첫 번째 축은 가장 큰 분산을 보여주며, 두 번째 축은 그 다음으로 높은 분산을 보여주는 식이다.
4) 내림차순 정렬되어 있으므로 약한 성분들(낮은 분산의 성분들)을 제거함으로써 데이터 크기를 줄일 수 있다. 즉, 가장 강한 성분들을 사용함으로써 크기가 축소된 원천 데이터의 훌륭한 근사치를 구성할 수 있다.

4.3.2 주성분 분석

다음은 한국신용평가정보에서 나온 국내 증권회사의 주요 재무제표(2007년 3월 31일 기준)의 일부이다. 주성분 분석을 통하여 데이터 차원을 축소시키시오.

⊙ 데이터 파일 : ch4-4(국내증권사제무재표).csv
⊙ 원본 투플 수 : 18개

증권사명	총자본순이익율	자기자본순이익율	자기자본비율	부채비율	자기자본회전율
SK증권	2.43	11.1	18.46	441.67	0.9
교보증권	3.09	9.95	29.46	239.43	0.9
대신증권	2.22	6.86	28.62	249.36	0.69
대우증권	5.76	23.19	23.47	326.09	1.43
동부증권	1.6	5.64	25.64	289.98	1.42

출처 공공데이터포털(www.data.go.kr)

☞ **해답은 ch4-4.ipynb 참고**

4.3.3 회귀와 로그-선형 모형

- 회귀와 로그-선형 모형 (Regression and Log-Linear Model)
  - 회귀와 로그-선형 모형은 주어진 데이터의 근사치를 구하는데 사용
  - 회귀
    - ✓ 선형 회귀(linear regression) : 확률변수(random variable) y를 예측변수(predicator variable)인 x의 선형 함수

      $y = wx + b$ 식 (3-4)
    - ✓ (식 3-4)에서 계수(coefficient) w와 b는 데이터를 분리하는 실제 선과 그 선의 추정치 사이의 오류를 최소화해주는 최소제곱법(method of least square)에 의해 구할 수 있음
    - ✓ 다중 회귀(multiple regression) : 확률변수 y가 2개 이상의 예측변수 x에 의해 모형화되도록 단순선형회귀를 확장한 것
  - 로그-선형모형
    - ✓ n개의 속성으로 표현되는 n차원에서 주어진 n개의 투플 집합을 n차원 공간의 한 점으로서 생각하는 이산 다차원 확률 분포의 근사치를 구함
    - ✓ 차원 조합의 가장 작은 부분집합에 기반하여, 이산화된 속성들의 집합에 대한 다차원 공간 내의 각 점의 확률을 평가하는데 사용
    - ✓ 차원축소와 데이터평활화에 유용
  - 회귀는 고차원으로 갈수록 계산비용이 기하급수적으로 증가하는 반면, 로그-선형 모형은 10차원 정도까지는 확장성(scalability)이 우수함

4.4 수량 축소

- 수량축소(numerosity reduction)
  - 대체할 수 있는 더 작은 형태의 표현을 선택해서 데이터량을 줄이는 기법
  - 모수적 기법(parametric method)
    - ✓ 모수의 특성을 활용하는 기법
    - ✓ 모집단이 정규 분포를 띤다는 가정하에 표본 통계량으로 모집단 통계량을 추정
  - 비모수적 기법(nonparametric method)
    - ✓ 모수의 특성을 활용하지 않는 기법
    - ✓ 표본추출, 군집화, 히스토그램 등

4.4.1 표본추출

- 표본추출(sampling)
  - 큰 데이터 집합을 많은 수의 임의 데이터 샘플(부분집합)으로 표현
  - 표본추출 방법
    - ✓ 비복원 단순무작위표본(SRSWOR: Simple random sample without replacement)
      D로부터 N개의 투플 중에서 임의의 s개를 취하는 방법으로 모든 투플들의 표본으로 추출될 확률은 같다
    - ✓ 복원 단순무작위표본(SRSWR: Simple random sample with replacement)
      각 투플이 D로부터 추출될 때마다 기록된 후 다시 제자리로 복원(replace)된다는 것을 제외하면 SRSWOR와 유사하다. 즉, 각 투플은 추출된 다음에 다시 추출될 수 있도록 D에 되돌려진다
    - ✓ 집략표본(cluster sample)
      D에 있는 투플들이 M개의 상호배반적인 군집(cluster)로 묶여 있는 가운데 s개의 군집을 단순무작위로 추출한다. (s < M)
    - ✓ 층화표본(Stratified sample)
      D가 층(strata)라 불리는 상호배반적인 부분들로 분할되어 있다면, 각 층에서 하나씩 단순무작위로 추출한다 (예, 고객의 나이 그룹 각각에 대하여 하나의 층이 생성되어 있는 고객 데이터로부터 층화표본을 얻음)

4.4.2 히스토그램

- 히스토그램(histogram)
  - 구간화를 사용하여 데이터 분포의 근사치를 구하는 데이터 축소의 전형적 형태
  - 속성값을 버킷(bucket) 혹은 빈(bin)이라 불리는 분리 집합(disjoint subset)으로 분할
  - 각 버킷은 단일한 속성값/빈도의 쌍으로 표현되거나 연속범위(continuous range)로 표현됨
  - 희소 데이터나 밀집 데이터 모두에 효과적이며, 비대칭적 데이터와 균일한 데이터 모두에 효과적
  - 버킷 결정 방법(속성값 분할 방법)
    - ✓ 동등 폭(Equal-width)
      각 버킷의 범위는 균일하다
    - ✓ 동등 빈도(Equal-frequency)
      각 버킷의 빈도가 일정하도록 만든다.(각 버킷이 같은 수의 데이터 표본을 포함함)
    - ✓ V-최적(V-optimal)
      최소분산을 갖는 히스토그램을 의미한다. 히스토그램 분산은 각 버킷이 나타내는 데이터 값들의 가중합(weighted sum)이며, 버킷 가중치는 버킷에 있는 값들의 개수와 같다
    - ✓ 최대차이(MaxDiff)
      인접한 값들의 각 쌍 사이의 차이를 고려한다. 사용자 정의 버킷의 수 β에 대하여, β-1개의 최대차이를 갖는 쌍들에 대하여 각 쌍 사이에 버킷 경계가 정해진다

4.4.3 군집화

- 군집화(clustering)
  - 데이터 투플을 객체로 간주하고, 각 객체들을 군집(cluster)라는 그룹으로 분할
  - 군집 내 객체들과는 유사하면서도 다른 군집 내 객체들과는 유사하지 않도록
  - 유사성 : 공간 내에서 객체들이 어떻게 가까운지의 관점에 따라 거리 함수에 기반하여 정의됨
    - ✓ 클러스터 지름(diameter) : 클러스터의 두 객체 간의 최대 거리
    - ✓ 클러스터 간 중심 거리(centroid distance) : 클러스터 중심 간의 거리
  - 고품질 클러스터 : 클러스터 지름은 짧고, 클러스터 간 중심 거리는 긴 상태

4.4 수량 축소 - 실무예제

다음은 2013년 전국 주요 지점별 유동인구 현황의 일부이다. 샘플링(Sampling), 히스토그램(Histogram), 클러스터링(clustering)을 활용하여 투플 수를 2,000개 이하로 축소시키시오.

⊙ 데이터 파일 : ch4-5(유동인구).csv

⊙ 원본 투플 수 : 23,221개

조사일자	시간대	X좌표	Y좌표	행정구역명	남자 10대	남자 20대	남자 30대	남자 40대	남자 50대	여자 10대	여자 20대	여자 30대	여자 40대	여자 50대
2010-06-21	12시~13시까지	343,099.00	417,482.00	대전광역시 서구 월평동	2	24	68	50	31	4	37	64	44	26
2010-06-21	19시~20시까지	343,099.00	417,482.00	대전광역시 서구 월평동	19	44	28	33	21	14	56	49	43	18
2010-06-20	12시~13시까지	343,099.00	417,482.00	대전광역시 서구 월평동	13	33	34	61	55	13	32	29	28	12
2010-06-20	19시~20시까지	343,099.00	417,482.00	대전광역시 서구 월평동	23	33	32	547	129	12	39	13	46	4
2010-06-21	12시~13시까지	343,121.00	417,343.00	대전광역시 서구 월평동	0	9	27	21	6	5	24	20	10	6

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch4-5.ipynb 참고**

연습문제

1. 다음은 서울메트로 역별/시간대별 승하차인원 현황의 일부이다. 데이터큐브 집계를 이용하여 월별/호선별/시간대별 승차/하차 합계로 데이터를 축소시키시오. (단, 시간대는 04~07, 07~09, 09~18, 18~20, 20~02로 구분)

◪ 데이터 원본: (연습4-1)서울메트로역별시간대별승하차인원현황.xlsx ◪ 원본 튜플수 : 60,000개

(5)

역명	구분	00-01	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18
동대문(155)	승차	50				19	517	407	772	1006	893	678	817	952	1072	1144	1148	1227	1232
동대문(155)	하차	162					74	296	515	1445	1173	1090	1192	1396	1267	1202	1172	1090	1018
동대문(155)	승차	56				17	452	443	894	1003	879	736	798	1043	1139	1257	1240	1211	1274
동대문(155)	하차	213					92	278	504	1399	1181	1149	1230	1286	1464	1261	1281	1301	1104
동대문(155)	승차	57				17	482	408	719	1030	837	759	771	1036	1138	1264	1157	1234	1310
동대문(155)	하차	189	2				60	280	537	1433	1196	1194	1237	1413	1316	1145	1308	1125	1090
동대문(155)	승차	35				24	502	398	986	1010	858	714	927	1034	1020	1122	1175	1138	1287
동대문(155)	하차	247					87	337	524	1366	1204	1177	1211	1339	1363	1306	1320	1079	1084
동대문(155)	승차	73	1			24	491	410	884	994	873	713	818	1081	1116	1196	1189	1205	1311

* 출처 : 서울시 열린데이터광장(data.seoul.go.kr)

2. 다음은 2016년도 서울시 일별 대기오염도 정보 중 일부이다. 이산화질소농도, 오존농도, 일산화탄소농도, 아황산가스농도, 미세먼지, 초미세먼지 측정치를 대상으로 웨이블릿 변환과 주성분 분석을 통하여 각각 데이터를 축소시키시오.

◪ 데이터 원본 : (연습4-2)서울시일별평균대기오염도정보.xlsx ◪ 원본 튜플수 : 14,302개

(6)

측정일시	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
20160606	강남구	0.02	0.041	0.2	0.004	23	15
20160606	강동구	0.012	0.04	0.3	0.004	24	16
20160606	강변북로	0.046	0.025	0.3	0.004	28	15
20160606	강북구	0.02	0.044	0.3	0.002	16	12
20160606	강서구	0.023	0.029	0.3	0.005	33	22
20160606	공항로	0.03	0.022	0.5	0.006	37	22
20160606	관악구	0.017	0.052	0.4	0.004	30	18
20160606	광진구	0.018	0.043	0.3	0.004	20	11
20160606	구로구	0.018	0.034	0.4	0.006	32	20

* 출처 : 서울시 열린데이터광장(data.seoul.go.kr)

연습문제

3. 다음은 서울시버스 노선별/정류장별 승하차인원 현황의 일부이다. 샘플링(Sampling), 히스토그램(Histogram), 클러스터링(clustering)을 활용하여 튜플수를 3,000개 이하로 축소시키시오.

◪ 데이터 원본: (연습4-3)서울시버스노선별정류장별승하차인원정보.xlsx ◪ 원본 튜플수 : 38,133개

(7)

사용월	버스노선번호	버스노선명	버스정류장명	승차인원	하차인원	작업일자
201512	9714	9714번(교하운정~서울역)	디지털미디어시티역	1763	209	20160308
201512	9714	9714번(교하운정~서울역)	교하자고지	189	8	20160308
201512	9714	9714번(교하운정~서울역)	교하자고지	0	151	20160308
201512	9714	9714번(교하운정~서울역)	마두역(중)	743	496	20160308
201512	9714	9714번(교하운정~서울역)	중앙공원	1	1504	20160308
201512	9714	9714번(교하운정~서울역)	중앙공원	1115	9	20160308
201512	9714	9714번(교하운정~서울역)	우리은행앞	51	4042	20160308
201512	9714	9714번(교하운정~서울역)	트리플메디컬타운	4359	46	20160308
201512	9714	9714번(교하운정~서울역)	청석마을8단지	5	1182	20160308

* 출처 : 서울시 열린데이터광장(data.seoul.go.kr)

Ch.5 데이터 변환(Data Transformation)

데이터 변환은 데이터 분석에 적절한 형태로 데이터를 바꾸는 전처리 작업을 의미한다. 데이터 변환 방법에는 평활, 집계, 속성구성, 정규화, 이산화, 개념계층 등의 방법이 있는데, 평활이나 집계 기법은 앞 장에서 이미 다루었으므로, 본 장에서는 정규화, 이산화, 개념 계층을 위주로 다루도록 한다.

5.1 정규화

- 정규화(Normalization)
  - 정규화는 속성값은 –1.0 ~ 1.0과 같이 정해진 구간 내에 들도록 하는 기법
  - 신경망을 포함한 분류 알고리즘이나 최단근접 분류와 군집화와 같은 거리측정 등을 위하여 특히 유용
  - 정규화 방법
    - ✓ 최소-최대 정규화 (Min-max normalization)
    - ✓ Z-score 정규화 (Z-score normalization)
    - ✓ 소수 척도화 (Decimal scaling)

5.1.1 최소-최대 정규화

- 최소-최재 정규화(min-max normalization)
  - 원본 데이터에 대하여 선형변환을 수행한다
  - 속성 A에 대한 최소값과 최대값을 각각 minA와 maxA라고 가정하자. 최소-최대 정규화는 다음 계산식에 의해 A의 값 v를 구간 [min'A, max'A]에서의 값 v'으로 사상한다

$$v' = \frac{v - \min_A}{\max_A - \min_A}(\max'_A - \min'_A) + \min'_A \qquad \text{식 (5-1)}$$

  - 원본 데이터 값들 간의 관계를 보존한다
  - 만약 정규화를 위한 입력이 A에 대한 원본 데이터 구간에서 벗어날 경우는 범위초과(out-of-bounds) 오류가 난다

5.1.2 Z-score 정규화

- Z-score 정규화(Z-score normalization)
  - 속성 A에 대한 값을 A의 평균과 표준편차를 기초로 정규화하는 방법이다
  - 와 를 A의 평균과 표준편차라고 가정할 때, 다음 계산식에 의하여 A의 값 v는 v'에 사상된다

$$v' = \frac{v - \bar{A}}{\sigma_A} \qquad \text{식 (5-2)}$$

  - 속성 A의 실제 최소값과 최대값이 알려져 있지 않거나, 최소-최대 정규화에 큰 영향을 주는 이상치가 존재할 때 유용하다

5.1.3 소수 척도화

- 소수 척도화(decimal scaling)
  - 속성 A 값들의 소수점을 이동해서 정규화한다
  - 이동되는 소수점의 수는 A의 최대 절대값에 의존한다. 을 만족하는 가장 작은 정수를 j라고 할 때, A의 값 v는 다음 계산식에 의하여 v′로 사상된다

$$v' = \frac{v}{10^j} \qquad \text{식 (5-3)}$$

5.1 정규화 - 실무예제

다음은 2016년 항만별 선박 입출항 현황의 일부이다. 입항 선박 수, 입항 선박 톤수, 출항 선박 수, 출항 선박 톤수를 최소-최대 정규화, z-점수 정규화, 소수 척도화 방식으로 각각 정규화하시오.

⊙ 데이터 파일 : ch2-2(선박입출항).csv

⊙ 원본 투플 수 : 30개

항만	입항선박수	입항선박톤수	출항선박수	출항선박톤수
부산	7,301	105,138,280	7,409	103,857,903
인천	2,715	30,716,710	2,716	30,779,186
평택.당진	1,558	23,153,226	1,536	22,778,109
경인항	28	126,236	27	128,344
동해.묵호	552	4,202,603	546	4,039,929

출처 공공데이터포털(www.data.go.kr)

☞ 해답은 ch5-1.ipynb 참고

5.2 수치형 데이터 이산화

- 수치형 데이터 이산화(numeric data discretization) 기법
  - 구간화(binning)
  - 히스토그램(histogram)
  - **엔트로피-기반**
  - **카이제곱-결합**
  - 군집분석
  - **직관적 분할**
- 일반적으로 각 기법들에 있어서 이산화 대상 값들이 오름차순으로 정렬되어 있음을 가정

5.2.1 엔트로피-기반 이산화

- 엔트로피-기반 이산화(entrophy-based discretization)
  - top-down 방식의 이산화 기법으로 분할점들의 결정과 계산을 위해 클래스 분포 정보를 탐색
  - 수치형 속성 A의 값들을 분할하기 위하여 분할점으로 최소 엔트로피를 갖는 A의 값을 선택하고, 계층적 이산화를 달성하기 위하여 그 결과 구간을 분할
  - 이러한 방식을 통하여 속성 A의 개념계층을 생성

5.2.1 엔트로피-기반 이산화

- 엔트로피-기반 이산화 과정

1) A의 각 값은 A의 범위를 분할하는 잠재적인 분할점(split-point)로 간주될 수 있다. 하나의 분할점에 의해 이항형(binary) 이산화를 진행할 수 있다.

2) 이항형 이산화 결과로 분리된 데이터 집합을 D1, D2라 하면, 속성 A의 기대정보 요구량(expected information requirement) InfoA(D)는 다음과 같다.

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2) \quad \text{식 (5-4)}$$

- $|D|$: 데이터집합 D의 투플 개수
- m개의 클래스 $C_1, C_2, ..., C_m$가 주어진 상황에서 D_1의 엔트로피는 $Entropy(D_1) = -\sum_{i=1}^{m} p_i \log^2(p_i)$ where p_i : 데이터 집합 D_1 내에서의 클래스 C_1 확률

3) 1), 2)의 과정을 어떤 정지규칙(stopping criterion)이 만족될 때까지(예를 들어, 모든 분할점 후보에 대한 최소 정보 요구량이 어떤 임계치(threshold) ε보다 작을 때까지, 혹은 구간의 수가 미리 설정한 최대 구간 수 보다 클 때가지) 구해진 각 분할에 재귀적으로 적용된다.

5.2.2 카이제곱-결합 이산화

- 카이제곱(χ^2)-결합 이산화(Interval Merging by χ^2 Analysis)
  - 가장 가까운 이웃 구간을 찾아내고 이들을 결합하여 더 큰 구간을 형성시키는 bottom-up 방식을 적용
  - 지도적(supervised) : 엔트로피-기반 이산화와 같이 클래스 정보를 활용
  - 핵심 개념 : 만약 두 개의 이웃 구간들이 매우 유사한 분포를 가진다면 그 구간들은 결합되고, 그렇지 않으면 분리된 상태로 둠
  - χ^2 -통계량은 3.3.2절의 데이터 통합에서 소개되었으며, 그곳에서 두 개의 범주형 속성 사이의 상관관계를 탐지하는 사용 → 주어진 속성에 대한 '두 개의 이웃 구간들이 독립이다'라는 가설을 검정
    - ✓ χ^2 -결합 구간들은 이산적인 범주로 취급될 수 있기 때문에 (식 3-2)를 적용
    - ✓ (예제 3-1)에서의 방법을 따라 데이터 분할표를 구성
    - ✓ 분할표는 두 개의 이웃 구간을 의미하는 두 개의 열과 m개의 행을 갖는다
    - ✓ (식 3-2)를 적용할 때, 칸 값 o_{ij}는 i번째 구간의 j번째 클래스에 해당하는 투플들의 수
    - ✓ o_{ij} 의 기대빈도수 e_{ij} 는 (구간 i의 튜플수) X (클래스 j의 튜플수) / N
    - ✓ N은 데이터 튜플의 총수
  - 한 구간 쌍에 대해 낮은 χ^2 값은 '구간들이 클래스와 독립'이라는 것을 의미 → 그 구간들을 결합

5.2.2 카이제곱-결합 이산화

- 카이제곱(χ^2)-결합 이산화(Interval Merging by χ^2 Analysis)
  - 정지 규칙(stopping criterion) : 다음 세가지 조건 중 하나에 의해 결정

    1) 모든 인접 구간들에 대한 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)

5.2.3 직관적 분할에 의한 이산화

- 직관적 분할에 의한 이산화
  - 대표적인 기법으로 3-4-5 규칙이 있음
  - 3-4-5 규칙 : 데이터의 주어진 범위를 '가장 큰 유효숫자'(most significant digit)에서의 값의 범위를 토대로, 재귀적(recursively) 그리고 수준별로 3, 4, 5개의 비교적 동일한 폭의 구간들로 분할
  - 3-4-5 규칙의 내용

    1) 어떤 구간이 가장 큰 유효숫자에서 3, 6, 7, 9개의 개별 값들을 포함한다면, 그 범위를 3개의 동일 폭 구간들로 분할한다(단, 7에 대해서는 2-3-2로 묶인 3개의 구간으로 분할)

    2) 어떤 구간이 가장 큰 유효숫자에서 2, 4, 8 개의 개별 값들을 포함한다면, 그 범위를 4개의 동일 폭 구간들로 분할한다

    3) 어떤 구간이 가장 큰 유효숫자에서 1, 5, 10개의 개별 값들을 포함한다면, 그 범위를 5개의 동일 폭 구간들로 분할한다

5.2 수치형 데이터 이산화 - 실무예제

실무 예제 5-2

다음은 2015년 국내 대학 현황이다. 재적 학생 수를 엔트로피-기반, 카이제곱-결합, 직관적 분할 방식을 각각 활용하여 20개 구간 내외로 이산화하시오.

- 데이터 파일 : ch5-2(국내대학현황).csv
- 원본 투플 수 : 1,720개

학제	학교명	지역	설립	재적학생수
전문대학	한국철도대학	경기	국립	180
전문대학	강원도립대학	강원	공립	1658
전문대학	경남도립거창대학	경남	공립	1614

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch5-2.ipynb 참고**

5.3 범주형 데이터를 위한 개념 계층

- 범주형 데이터를 위한 개념 계층(conceptual hierarchy)
  - 범주형 데이터는 기본적으로 이산 데이터로서 값들 사이에 순서가 없는 유한 개의 구별 값을 갖음
  - 개념 계층 생성을 위한 몇가지 방법들

1) 스키마 단계에서의 명시적 생성

스키마 수준에서 속성들의 개념 계층을 정의한다.

예) 주소의 경우, 스키마 수준에서 '동 < 기초시군구 < 광역시도 < 국가'를 정함

2) 명시적 데이터 그룹화에 의한 계층 일부 생성

개념 계층의 일부를 수동으로 정의한다.

예) {서울특별시, 경기도, 인천광역시} ⊂ 수도권, {전라북도, 전라남도, 광주광역시} ⊂ 호남권 등

3) 속성들의 개념 계층을 자동 생성

개념 계층은 일반적으로 상위 개념수준이 몇 개의 하위 개념수준을 포함하므로, 상위수준의 개념 수가 하위수준의 개념 수보다 적다. 이러한 점에 착안하여 각 수준에서의 구별 값들의 개수를 근거로 상위와 하위의 순서를 자동으로 정하여 개념 계층을 생성한다.

4) 속성들의 부분 집합만을 생성

개념 계층의 생성 시에 관련 속성들의 일부만 사용하는 방식이다.

예) 주소에 대해서 동과 기초시군구 만을 사용하여 개념 계층 생성

연습문제

1. 다음은 서울시버스 노선별/정류장별 승하차인원 현황의 일부이다. 승차인원과 하차인원을 최소-최대 정규화, z-점수 정규화, 소수 첨도화 방식으로 각각 정규화하시오.

◪ 데이터 원본: (연습5-1)서울시버스노선별정류장별승하차인원정보.csv ◪ 원본 튜플수 : 38,133개

사용월	버스노선번호	버스노선명	버스정류장명	승차인원	하차인원	작업일자
201512	9714	9714번(교하운정~서울역)	디지털미디어시티역	1763	209	20160308
201512	9714	9714번(교하운정~서울역)	교하차고지	189	8	20160308
201512	9714	9714번(교하운정~서울역)	교하차고지	0	151	20160308
201512	9714	9714번(교하운정~서울역)	마두역(중)	743	496	20160308
201512	9714	9714번(교하운정~서울역)	중앙공원	1	1504	20160308
201512	9714	9714번(교하운정~서울역)	중앙공원	1115	9	20160308
201512	9714	9714번(교하운정~서울역)	우리은행앞	51	4042	20160308
201512	9714	9714번(교하운정~서울역)	트리플메디컬타운	4359	46	20160308
201512	9714	9714번(교하운정~서울역)	청석마을8단지	5	1182	20160308

* 출처 : 서울시 열린데이터광장(data.seoul.go.kr)

(8)

2. 다음은 부산광역시 어린이집 현황의 일부이다. 아동현원을 엔트로피-기반, 카이제곱-결합, 직관적분할 방식을 각각 활용하여 10개 구간 내외로 이산화하시오. (단, 카이제곱-결합 방식에서의 독립성 여부 판단 기준 속성은 어린이집유형으로 한다)

◪ 데이터 원본 : (연습3-2)부산시어린이집현황.csv ◪ 원본 튜플수 : 206개

No	어린이집유형	어린이집	어린이집특성	평가인증여부	정원	아동현원	보육교직원현원
1	민간	21세기어린이집	일반	Y	70	70	10
2	국공립	ymca어린이집	장애아통합	Y	60	60	11
3	가정	가람어린이집	일반	N	14	11	3
4	민간	가빈어린이집	일반	N	59	50	14
5	가정	거목어린이집	일반	Y	20	20	5
6	민간	건강한영아전담어린이집	영아전담	Y	32	32	10
7	민간	과학나라어린이집	일반	N	39	21	6
8	민간	구름나무어린이집	일반	N	32	29	7
9	법인·단체등	구포원광어린이집	일반	Y	49	45	7

* 출처 : 공공데이터포털(www.data.go.kr)

(9)

Ch.6 Case Study

본 장에서는 지금까지 다룬 데이터전처리 기법들이 실제적인 문제에서 어떻게 적용되는지 case study를 통해서 알아본다. 분석 주제는 '취업률 관점에서의 대학정보공시 정량지표 분석'이다. 대학들은 대학정보공시 지표항목에 대해서 의무적으로 공개해야 한다. 취업률 또한 대학정보공시 지표항목의 하나로서 각종 대학평가에서 중요한 요소로 활용되고 있다. 취업률 이외에도 대학의 경쟁력을 보여주는 다양한 대학정보공시 정량지표들이 존재하는데, 이들이 취업률과 어떠한 상관관계가 있는지를 분석함으로써 대학들이 취업률 상향을 위해서 어떠한 정책들을 펼쳐나가야 되는지에 대한 시사점 제공에 분석 목적이 있다. 본 저서의 목정 상 분석 결과 자체에 초점을 두기 보다는 이러한 분석을 잘 수행하기 위해서 어떤 데이터 전처리 기법들이 어떻게 활용되는지에 대한 사례를 소개하는데 초점을 두고자 한다.

6.1 분석 모형

- 분석 절차

① 데이터셋 준비

원천 데이터의 확보 대학알리미 사이트(http://www.academyinfo.go.kr/) 에서 대학정보공시 관련 데이터를 확보할 수 있음

② 각 변수 값의 데이터 정규화(Z-score)

정량지표들 사이의 값의 단위(scale)이 천차만별이기 때문에 데이터 정규화를 통해 동일한 값의 단위로 정량지표 값을 변환

③ 취업률 상관도 분석 및 변수(지표) 선택 : 상관분석, 선형회귀분석

정량지표 간의 상관관계 분석을 통해 분석 차원을 결정함. 상관분석(correlation analysis)과 선형회귀분석(linear-regression analysis) 기법 적용

④ 취업률 기준 두 개의 집단으로 분류

취업률 상위 집단과 하위 집단 분류

⑤ 각 집단 별 군집(cluster) 분석

취업률 상·하위 집단에서의 정량지표 분포 분석 및 비교를 통하여 취업률 상향을 위해서 중점을 두고 관리해야할 지표 도출. 분포 분석을 위해 클러스터링 기법 적용

⑥ 취업률 상향 전략 도출

분석 결과를 토대로 취업률 향상을 위한 정책 방향 시사점 도출

6.2 분석대상 데이터셋

- 분석대상 데이터셋
  - 2014년 전문대학 정보공시 데이터
    - ✓ 데이터 원본 : (원본)전문대학정보공시데이터.xlsx
    - ✓ 원본 투플수 : 137개
  - 분석대상 정량지표

학생충원률, 중도탈락학생비율, 학생1인당장학금, 장학금수혜율, 학자금대출이용학생비율, 학생1인당교육비, 전임교원1인당학생수, 전임교원강의담당비율, 업체당실습학생수평균, 현장실습이수율, 캡스톤디자인이수비율, 캡스톤디자인학생당지원금액, 취업률 (이상 13개 지표)

  - 예제 데이터

순번	학생충원율	중도탈락학생비율	1인당장학금	장학금수혜율	학자금대출이용학생비율	학생1인당교육비	전임교원1인당 학생수(재학생기준)	전임교원강의담당비율	업체당 실습학생수평균	현장실습이수율	캡스톤디자인이수비율	캡스톤디자인학생당지원금액	취업률
1	0.957	0.097	2,912.20	51	18.2	8,829.70	32.8	51.6	2.7	16.1%			0.682
2	0.972	0.057	2,367.70	39.7	23.1	8,402.70	35.8	39.9	3.1	50.9%			0.585
3	1.015	0.037	2,879.00	49.2	22.9	8,418.40	33.2	52.2	2.7	17.2%	1.000	67	0.593
4	1.012	0.100	2,410.40	49.2	21.5	7,435.50	36.7	40.8	3.1	34.8%			0.613
5	0.945	0.061	2,231.90	89.1	9.9	9,234.30	32.4	36.3	2.8	32.6%			0.492

*** 출처 : 대학알리미 사이트(http://www.academyinfo.go.kr/)**

6.2 분석대상 데이터셋

- 분석대상 데이터셋
  - 각 대학명은 순번으로 무기명 처리하였으며, 지표명은 변수명으로 쓰기에는 다소 길어서 영문 약어로 대치

지표명	변수명
순번	seq
학생충원율	sc1
중도탈락학생비율	sc2
1인당장학금	sc3
장학금수혜율	sc4
학자금대출 이용학생비율	sc5
학생1인당교육비	sc6
전임교원1인당 학생수(재학생기준)	sc7
전임교원강의담당비율	sc8
업체당 실습학생수평균	sc9
현장실습이수율	sc10
캡스톤디자인이수비율	sc11
캡스톤디자인학생당지원금액	sc12
취업률	tc

[표 6.1] 지표명과 변수명 매핑표

6.3.1 결측치 처리 및 데이터 정규화

- 결측치 처리
  - 결측치를 해당 속성의 평균값으로 대체하는 방법 사용 (2.1절 참조)

- 데이터 정규화
  - 지표값의 단위와 크기는 가지각색
    - ✓ 충원율, 중도탈락률과 같이 비율로 표시되는 지표값는 0~1 사이인 반면, 학생1인당교육비와 같이 금액으로 표시되는 지표값은 천단위가 넘어감
  - 정규화된 값들을 가지고 분석할 경우, 분석결과에 대한 해석 또한 훨씬 더 직관적임
  - 대학정보공시 지표들의 값 분포는 대부분 정규분포에 가까우므로 Z-score 정규화 적용

☞ **ch6.ipynb 참고**

6.3.2 상관분석

- 상관분석(correlation analysis)
  - 속성(변수) 간의 상관도 분석을 통하여 분석 차원을 단순화함으로써 분석의 복잡도를 감소시키고자 할 때 유용한 전처리 방식
  - 두 변수 간 상관도는 상관계수(correlation coefficient)로 표현
    - ✓ -1.0과 -0.7 사이이면, 강한 음적 선형관계,
    - ✓ -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계,
    - ✓ -0.3과 -0.1 사이이면, 약한 음적 선형관계,
    - ✓ -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계,
    - ✓ +0.1과 +0.3 사이이면, 약한 양적 선형관계,
    - ✓ +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계,
    - ✓ +0.7과 +1.0 사이이면, 강한 양적 선형관계 선택

☞ **ch6.ipynb 참고**

6.3.2 상관분석

순위	정보공시 지표	취업률 상관계수	상관관계 해석	절대값
1	학생충원율	0.353	뚜렷한 양적 선형관계	0.353
2	현장실습이수율	0.351	뚜렷한 양적 선형관계	0.351
3	학자금대출 이용학생비율	-0.264	약한 음적 선형관계	0.264
4	학생1인당교육비	0.217	약한 양적 선형관계	0.217
5	장학금수혜율	0.138	약한 양적 선형관계	0.138
6	전임교원강의담당비율	0.131	약한 양적 선형관계	0.131
7	중도탈락학생비율	-0.126	약한 음적 선형관계	0.126
8	전임교원1인당 학생수	0.093	거의 무시될 수 있는 선형관계	0.093
9	캡스톤디자인학생지원금액	0.073	거의 무시될 수 있는 선형관계	0.073
10	캡스톤디자인이수비율	-0.043	거의 무시될 수 있는 선형관계	0.043
11	1인당장학금	0.020	거의 무시될 수 있는 선형관계	0.020
12	업체당 실습학생수 평균	0.012	거의 무시될 수 있는 선형관계	0.012

[도표 6-3] 정보공시 지표와 취업률 간의 상관계수 (순위기준)

6.3.2 상관분석

- 지표 그룹핑(grouping)
  - 12개의 지표로 구성된 분석 차원을 효과적으로 줄이기 위해서 의미적으로 유사한 지표들끼리 그룹핑하고, 각 그룹 내에서 취업률과의 상관도가 높은 대표 지표들을 최종 분석 대상 속성 후보로 결정
  - 지표 그룹은 지표가 가지는 의미에 따라 크게 학생충원, 학생재정, 전임교원, 현장실습, 캡스톤디자인 등 5개의 그룹으로 구분
  - 각 그룹에 속한 지표와 그룹 관점에서의 취업률과의 상관관계 분석 결과

□ 학생충원 그룹 (학생충원율, 중도탈락학생비율)

학생충원율이 중도탈락학생비율 보다 취업률과의 상관도가 강함. 중도탈락학생비율의 경우, 학생관리 수준 측면과 학생의 질적 측면이라는 상반된 양면성이 긍정과 부정의 효과를 상쇄하고 있는 것으로 보임

□ 학생재정 그룹 (학생1인당 교육비, 학자금 이용학생비율, 장학금수혜율, 1인당장학금)

장학금은 양의 상관관계를 학자금대출은 음의 상관관계를 보이고 있으나, 장학금 보다는 학자금대출비율이 취업률에 더 큰 영향을 미치는 것으로 나타남. 학생1인당 교육비는 대학의 학생에 대한 투자정도가 취업률에 영향을 미치고 있음을 나타냄

6.3.2 상관분석

□ 현장실습 그룹 (현장실습이수율, 업체당 실습학생수평균)

현장실습이수율이 취업률과 높은 양의 상관관계를 보이고 있음. 이는 대학의 적극성, 교육의 현장성이 취업률에 매우 긍정적인 영향을 미치는 것으로 분석될 수 있음. 업체당 몇 명의 학생이 현장실습을 수행하는가는 취업률에 큰 영향이 없는 것으로 나타남

□ 전임교원 그룹 (전임교원강의담당비율, 전임교원1인당학생수)

전임교원1인당 학생수보다는 전임교원강의담당비율이 상대적으로 취업률에 끼치는 영향이 큰 것으로 나타남. 이는 강의의 질적 측면보다는 학생관리적 측면에서 전임교원과 학생의 접촉면이 넓어지는 것이 취업률에 긍정적인 영향을 미치는 것으로 해석 가능함

□ 캡스톤디자인 그룹 (캡스톤디자인이수비율, 캡스톤디자인학생당지원금액)

학생당지원금액이 이수비율보다 훨씬 더 좋은 취업률과의 양의 상관관계를 가지는 것으로 나타남으로써 캡스톤디자인에 있어서 양적인 면보다는 질적인 면이 취업률에 상대적으로 긍정적 영향을 미치는 것으로 분석됨

6.3.2 상관분석

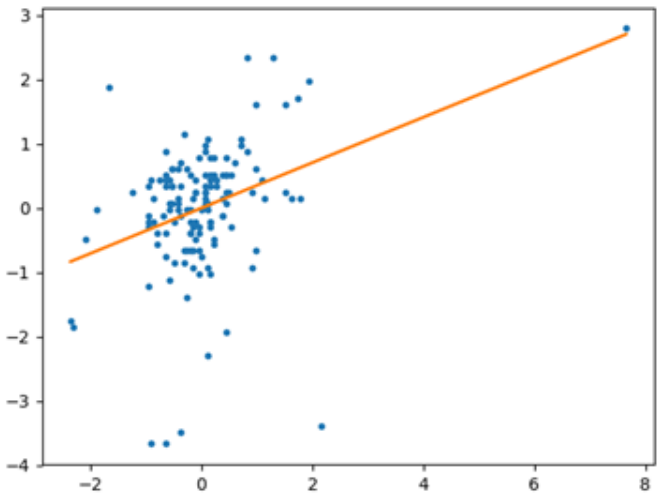
- 속성(지표) 선택
  - 일반적으로 양의 관계든 음의 관계든 상관도가 높은(절대값이 큰) 속성이 선택되는 것이 효과적
  - 그러나, 단순히 상관도가 높은 순서대로 선택하는 것보다는 다양한 관점에서의 분석이 가능하도록 여러 지표그룹에서 골고루 선택하는 것 또한 간과해서는 안될 것임

지표그룹	정보공시 지표	취업률 상관계수	상관관계 해석	후보변수 선택여부
학생충원	학생충원율	0.353	뚜렷한 양적 선형관계	선택
	중도탈락학생비율	-0.126	약한 음적 선형관계	배제
학생재정	학생1인당교육비	0.217	약한 양적 선형관계	선택
	학자금대출 이용학생비율	-0.264	약한 음적 선형관계	선택
	장학금수혜율	0.138	약한 양적 선형관계	배제
	1인당장학금	0.020	거의 무시될 수 있는 선형관계	배제
전임교원	전임교원강의담당비율	0.131	약한 양적 선형관계	선택
	전임교원1인당 학생수	0.093	거의 무시될 수 있는 선형관계	배제
현장실습	현장실습이수율	0.351	뚜렷한 양적 선형관계	선택
	업체당 실습학생수 평균	0.012	거의 무시될 수 있는 선형관계	배제
캡스톤 디자인	캡스톤디자인학생당지원금액	0.073	거의 무시될 수 있는 선형관계	선택
	캡스톤디자인이수비율	-0.043	거의 무시될 수 있는 선형관계	배제

[도표 6-4] 정보공시 지표와 취업률 간의 상관계수 (지표그룹기준)

6.3.3 선형회귀분석

- 각 후보 지표(속성)에 대한 검증 결과
  - 학생 충원율
    - ✓ 학생충원율과 취업률 간의 선형은 기울기가 급한 강한 양적 관계를 보이고 있으며, 몇몇 예외적인 데이터(outlier)가 있으나 그 수가 많지 않으므로 학생충원율을 최종변수(분석대상 속성)로 선택함

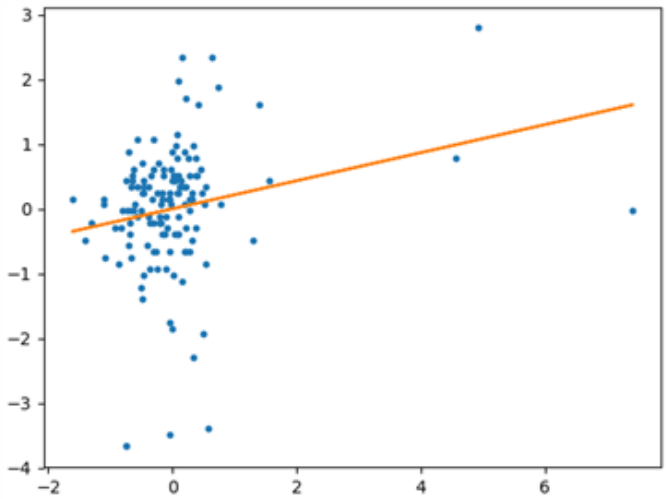


☞ **ch6.ipynb 참고**

[도표 6-5] 학생충원율과 취업률 간 선형회귀분석 결과

6.3.3 선형회귀분석

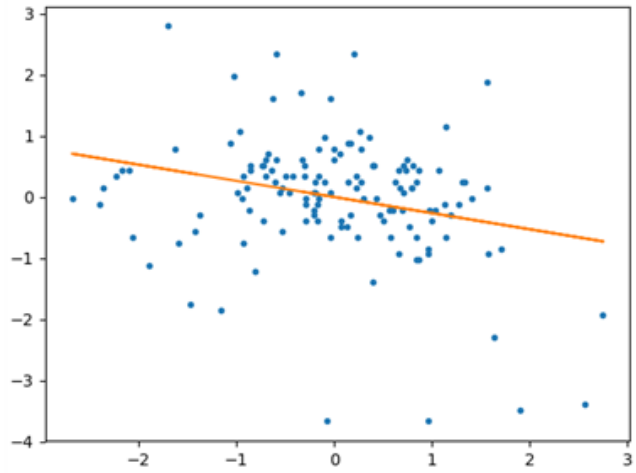
- 각 후보 지표(속성)에 대한 검증 결과
  - 학생1인당 교육비
    - ✓ 학생1인당교육비과 취업률 간의 선형은 기울기가 있는 양적 관계를 보이고 있으며, 몇몇 예외적인 데이터(outlier)가 있으나 그 수가 많지 않으므로 학생1인당교육비를 최종변수로 선택함



[도표 6-6] 학생1인당교육비와 취업률 간 선형회귀분석 결과

6.3.3 선형회귀분석

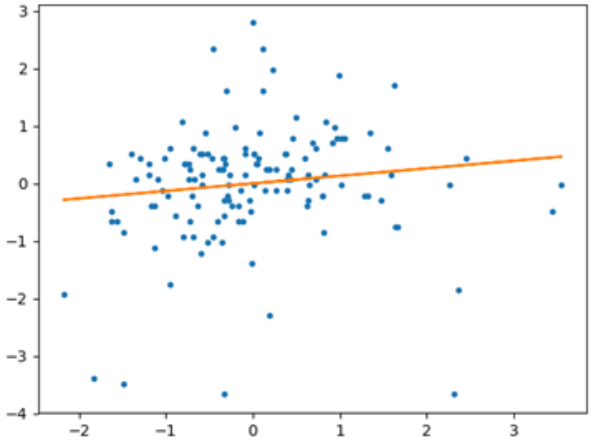
- 각 후보 지표(속성)에 대한 검증 결과
  - 학자금대출이용학생비율
    - ✓ 학자금대출이용학생비율과 취업률 간의 선형은 비교적 기울기가 급한 음적 관계를 보이고 있으며, 학자금대출이용학생비율이 비교적 넓게 분포되어 있고 비율이 높을수록 취업률이 낮은 경향성을 비교적 뚜렷하게 보이고 있으므로 학자금대출이용학생비율을 최종변수로 선택



[도표 6-7] 학자금대출이용학생비율과 취업률 간 선형회귀분석 결과

6.3.3 선형회귀분석

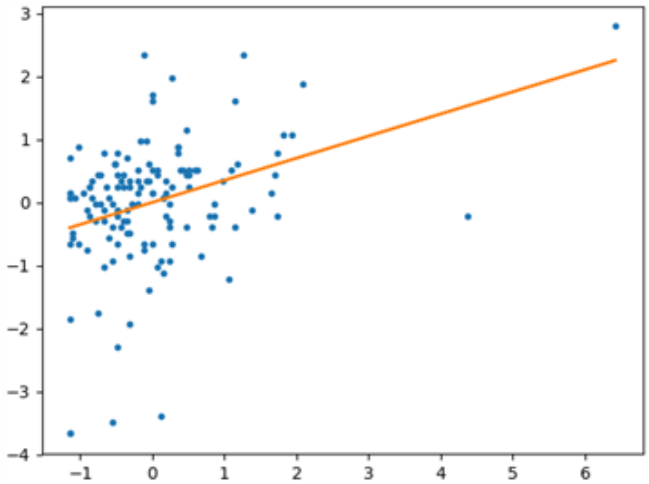
- 각 후보 지표(속성)에 대한 검증 결과
  - 전임교원강의담당비율
    - ✓ 전임교원강의담당비율과 취업률 간의 선형은 기울기가 비교적 있는 양적 관계를 보이고 있고, 전임교원강의담당비율이 비교적 넓게 분포되어 있으며, 몇몇 예외적인 데이터(outlier)가 있으나 비율이 높을수록 취업률이 높은 경향성을 비교적 뚜렷하게 보이고 있으므로 전임교원강의담당비율을 최종변수로 선택함



[도표 6-8] 전임교원강의담당비율과 취업률 간 선형회귀분석 결과

6.3.3 선형회귀분석

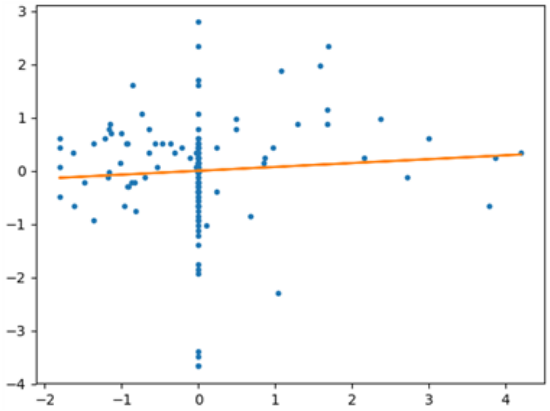
- 각 후보 지표(속성)에 대한 검증 결과
  - 현장실습이수율
    - ✓ 현장실습이수율과 취업률 간의 선형은 기울기가 급한 강한 양적 관계를 보이고 있으며, 이수율이 높을수록 취업률이 높은 경향성을 매우 뚜렷하게 보이고 있으므로 현장실습이수율을 최종변수로 선택함



[도표 6-9] 현장실습이수율과 취업률 간 선형회귀분석 결과

6.3.3 선형회귀분석

- 각 후보 지표(속성)에 대한 검증 결과
  - 캡스톤디자인학생당지원금액
    - ✓ 캡스톤디자인학생당지원금액과 취업률 간의 선형은 기울기가 비교적 완만한 양적 관계를 보이고 있으며, 지원금액의 분포가 비교적 넓은 가운데 몇몇 예외적인 데이터(outlier)가 있으나 양의 분포를 확인할 수 있으므로 최종변수로서로 삼을 수도 있겠으나, 결측치가 상당수 존재(137개 레코드 중 75개)하여 결측치를 추정하여 채운다 하더라도 분석 결과를 왜곡할 가능성이 존재하므로 배제함



[도표 6-10] 캡스톤디자인학생당지원금액과 취업률 간 선형회귀분석 결과

6.3.3 선형회귀분석

- 최종 분석 대상 지표(속성)
  - 6개의 후보 속성 중 캡스톤디자인학생당지원금액을 제외한 나머지 5개 속성이 최종 분석 대상 속성으로 채택
  - 종합적으로, 12개의 분석 차원을 5개의 차원으로 차원축소함

지표그룹	정보공시 지표	취업률 상관계수	상관관계 해석	후보변수 선택여부	최종변수 포함여부
학생충원	학생충원율	0.353	뚜렷한 양적 선형관계	선택	포함
	중도탈락학생비율	-0.126	약한 음적 선형관계	배제	불포함
학생재정	학생1인당교육비	0.217	약한 양적 선형관계	선택	포함
	학자금대출 이용학생비율	-0.264	약한 음적 선형관계	선택	포함
	장학금수혜율	0.138	약한 양적 선형관계	배제	불포함
	1인당장학금	0.020	거의 무시될 수 있는 선형관계	배제	불포함
전임교원	전임교원강의담당비율	0.131	약한 양적 선형관계	선택	포함
	전임교원1인당 학생수	0.093	거의 무시될 수 있는 선형관계	배제	불포함
현장실습	현장실습이수율	0.351	뚜렷한 양적 선형관계	선택	포함
	업체당 실습학생수 평균	0.012	거의 무시될 수 있는 선형관계	배제	불포함
캡스톤 디자인	캡스톤디자인학생당지원금액	0.073	거의 무시될 수 있는 선형관계	선택	불포함
	캡스톤디자인이수비율	-0.043	거의 무시될 수 있는 선형관계	배제	불포함

[도표 6-11] 상관분석과 선형회귀분석을 통한 분석 차원 축소 과정 정리

연습문제

1. 6장의 사례를 대상으로 분석 차원을 축소하는 방법으로 상관분석 & 선형회귀분석 대신에 주성분분석을 사용하여 수행하고 결과를 비교해 보시오.

◪ 데이터 원본 : (연습6-1)전문대학정보공시데이터.csv ◪ 원본 투플수 : 137개

(10)

순번	학생충원율	중도탈락학생비율	1인당장학금	장학금수혜율	학자금대출이용학생비율	학생1인당교육비	전임교원1인당 학생수(재학생기준)	전임교원강의담당비율	업체당 실습학생수평균	현장실습이수율	캡스톤디자인이수비율	캡스톤디자인학생당지원금액	취업률
1	0.957	0.097	2,912.20	51	18.2	8,829.70	32.8	51.6	2.7	16.1%			0.682
2	0.972	0.057	2,367.70	39.7	23.1	8,402.70	35.8	39.9	3.1	50.9%			0.585
3	1.015	0.037	2,879.00	49.2	22.9	8,418.40	33.2	52.2	2.7	17.2%	1.000	67	0.593
4	1.012	0.100	2,410.40	49.2	21.5	7,435.50	36.7	40.8	3.1	34.8%			0.613
5	0.945	0.061	2,231.90	89.1	9.9	9,234.30	32.4	36.3	2.8	32.6%			0.492
6	1.300	0.076	1,992.10	32.2	16.8	9,234.30	37.9	41.8	3.0	61.2%	0.634	249	0.867
7	1.236	0.079	2,350.40	35.8	19.9	9,911.80	39.5	37.8	2.5	58.7%	0.698	42	0.682
8	0.957	0.135	2,070.50	69.1	8.1	9,351.00	34.3	39.1	2.4	53.5%	1.000	107	0.646
9	1.080	0.064	1,784.50	62.9	14.8	7,566.60	39.6	38.9	2.4	78.1%	0.673	76	0.725
10	1.193	0.079	3,105.90	49.8	21.8	8,991.00	39.1	43.9	2.9	26.9%	0.878	298	0.720

* 출처 : 대학알리미 사이트(http://www.academyinfo.go.kr/)