1. 기계학습의 개요

A. 기계학습의 개요

1) 데이터 과학

데이터에서 지식을 얻기 위한 절차와 방법론을 연구, 데이터 수집, 클리닝, 분석, 시각화, 배포 등의 반복적인 절차를 연구하는 학문 분야

2) 데이터마이닝(Data Mining)

대용량의 데이터로부터 유용한 정보를 캐내는(mining) 작업

대용량 데이터에 존재하는 데이터 간의 관계, 패턴, 규칙 등을 찾아내고 모형화해서 기업의 경쟁력 확보를 위한 의사결정을 돕는 일련의 과정

3) 기계학습(Machine Learning)

① 데이터 과학에서 도출된 개념과 방법론의 분석과 모델링에 사용되는 보편적인 알고리즘과 기술을 연구

② 컴퓨터가 스스로 학습한다는 개념이 아님

③ 지도 학습

학습용 데이터로 작업, 입력과 출력(결과물, 해답)이 주어진다.

④ 비지도 학습

답을 미리 알려주지 않고 알고리즘에 의하여 데이터의 숨겨진 패턴을 찾아내는 방법

4) 기계학습의 활용

스팸탐지, 음성인식, 주가예측, 헬스케어 등

5) 기계학습을 위한 프로그래밍 언어

Python, R, Java

B. 기계학습 계획하기

1) 기계학습 순환 주기

문제 정의 => 데이터 수집 => 데이터 전처리 => 기계 학습 수행 => 결과 제시

① 데이터와 문제 정의

해결하려는 문제는 무엇인가? 그것이 왜 중요한가? 그에 대한 답은 무엇인가?

② 데이터 수집

③ 데이터 전처리(데이터 클리닝)

누락된 데이터 보충(결측값), 노이즈가 섞인 데이터를 부드럽게 처리, 특이한 데이터 제거(이상치)

④ 데이터 분석과 모델링

⑤ 평가

학습 모델을 올바르게 평가하고 새로운 데이터에서도 만족스런 결과를 낼 수 있는지 확인

2) 데이터와 문제의 정의

① 데이터 : 숫자, 단어, 측정값, 관찰 결과, 사물에 대한 묘사, 이미지 등으로 구성된 값의 모음

② 범주형 데이터(Categorical Data)

사전에 정해진 특정 유형으로 분류되는 데이터, 평균이 의미가 없음

명목형 : 값들 간의 크기 비교가 불가능, 상호 배타적이며 순서와 무관한 데이터

ex) 정치적 성향(좌파, 우파), 성별, 혈액형, 축구선수 등번호, 출신학교 코드, 출신국가, 직업구분, 주택보유 여부 등

순서형 : 대, 중, 소와 같이 값에 순서를 매길 수 있는 경우

ex) 성적 등급, 방의 크기, 학력 등

③ 연속형 데이터(Continuous Data)

정량적 데이터, 평균이 의미가 있음

등간척도 : 온도, 시간 등

비율척도 : 키, 몸무게, 점수, 투표율 등

3) 데이터 수집

① 데이터의 발견과 관찰

② 데이터 수집

설문 조사

스크레이핑

다양한 데이터 소스

공공데이터 포털(http://www.data.go.kr) 공공데이터 25,000여개 제공

공공 데이터 활용 창업경진대회(http://www.startupidea.kr)

카글 등 외국 사이트

실험 및 시뮬레이션

실제로 실행하기 어려운 실험을 컴퓨터를 이용하여 간단히 행하는 모의실험

물리적 시뮬레이션 : 모델하우스, 댐 건설을 위한 모형 제작

컴퓨터 시뮬레이션 : 비행 시뮬레이션, 게임 시뮬레이션

③ 데이터 샘플링

4) 데이터 전처리(데이터 클리닝, Data Cleaning)

본격적인 기계학습 절차가 진행되기 전에 정확하지 않거나 불충분한, 관련성이 떨어지는 데이터 제거

① 누락된 값 채우기

레코드 제거

필드 제거 - 대부분의 값이 누락된 필드라면 해당 필드를 삭제한다.

N/A 값 부여- 데이터가 존재하지 않는다는 의미로 부여

평균 속성값으로 대체

다른 속성을 통해 값을 예상 - 시간의 흐름에 따라 변하는 경우 등

② outlier의 제거

일련의 값의 흐름을 벗어난 값

평균을 지나치게 벗어난 극한값

③ 데이터 변환

머신러닝에 적합한 포맷으로 변환하는 작업

④ 데이터 축소

예측력 저하의 원인이 되는 속성 자체를 제거

고차원 데이터를 저차원으로 변환

5) 분석할 데이터의 종류

① 원시 텍스트 : 데이터가 구조적이지 않기 때문에 분석하기가 어려움

② **csv** : 가장 많이 사용되는 형식, 각각의 필드가 쉼표(,)로 구분되어 있음

③ **json(Javascript Object Notation,** 자바스크립트 객체 표기법)

④ **xml**

⑤ 엑셀파일

⑥ 데이터베이스

⑦ 이미지 : 안면 인식, 패턴 인식 등 다양한 분야에서 활용됨,

2. 기초통계와 검정

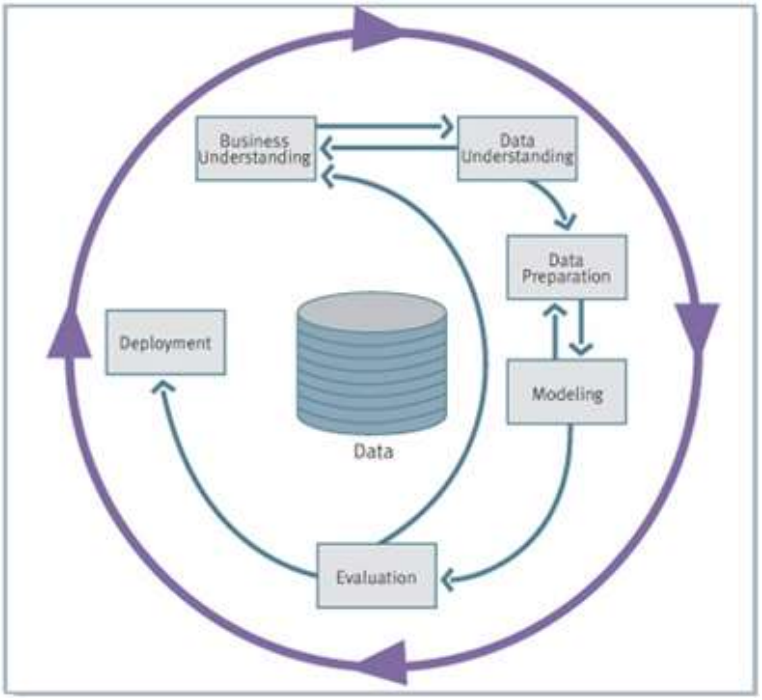
A. 데이터마이닝(Data Mining)

대용량의 데이터로부터 유용한 정보를 캐내는(mining) 작업

대용량 데이터에 존재하는 데이터 간의 관계, 패턴, 규칙 등을 찾아내고 모형화해서 기업의 경쟁력 확보를 위한 의사결정을 돕는 일련의 과정

1) CRISP-DM(CRoss-Industry Standard Processing for Data Mining)

데이터마이닝을 위한 업계 표준 프로세스



① 비즈니스 이해(Business Understanding)

각종 참고 자료와 현업 책임자와의 의사소통을 통해 해당 비즈니스를 이해하는 단계, 반드시 그 분야의 전문가가 함께 참여해야 함.

② 데이터 이해(Data Understanding)

초기 데이터 수집, 데이터에 대한 이해, 데이터 품질 검증

레코드 수, 변수(필드) 종류, 자료의 질 등 현업에서 보유 관리하고 있는 데이터를 이해하는 단계

③ 데이터 준비(Data Preparation)

모델링에 필요한 데이터 변환 및 정제 등의 작업

데이터의 정제, 새로운 데이터 생성, 데이터 업데이트 등 자료를 분석 가능한 상태로 만드는 단계

④ 모델링(Modeling)

데이터 분석을 위한 모델링 기법을 선택하고 모델을 생성하는 단계

⑤ 평가(Evaluation)

결과에 대한 분석 및 평가

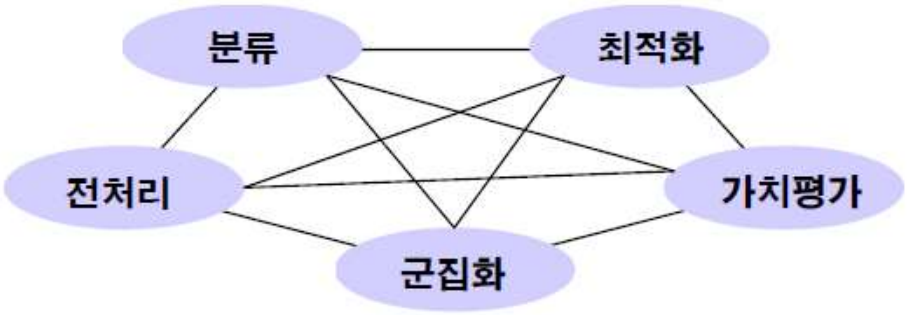
모형의 해석 가능 여부, 새로운 자료에도 잘 적용되는지 검토하는 단계,

필요하다면 모형 재구축

⑥ 배포(Deployment)

최종 보고서 작성 및 배포, 각 관리자에게 전달하여 필요한 조치를 취하는 등 검토가 끝난 모형을 실제 현업에 적용하는 단계

2) 인공지능/데이터마이닝의 다양한 기법들



① 분류 모형(Classification Models)

어떤 기준(정답)에 의해 분석 대상을 2개 혹은 3개 이상의 집단으로 분류하는 예측 모형(부도예측, 기상예측, 채권등급예측 등)

- 다중판별분석(MDA, Multiple Discriminant Analysis)
- 로지스틱 회귀분석(LOGIT, Logistic Regression)
- 인공신경망(ANN, Artificial Neural Networks)
- 사례기반추론(CBR, Case-Based Reasoning)
- 의사결정나무(DT, Decision Trees)
- SVM(Support Vector Machines)

② 최적화 기법(Optimization Methods)

주어진 제약조건 하에서 특정 목적함수를 최대,최소화하는 변수들의 최적값을 도출하는 기법(공장의 생산량 최대화 문제, 비용을 최소화하는 최적 유통경로 등)

- 선형계획법(LP, Linear Programming)
- 유전자 알고리즘(GA, Genetic Algorithms)

③ 가치평가 기법(Valuation Methods)

정성적 측정대상에 대한 가치를 비교,평가하는 기법(차세대 전투기 선정, 신행정수도 선정 등)

- 분석적 계층 프로세스(AHP, Analytic Hierarchy Process)
- 분석적 네트워크 프로세스(ANP, Analytic Network Process)
- 자료포락분석(DEA, Data Envelopment Analysis)

④ 분류/군집화 기법(Clustering Methods)

사전에 정해진 기준없이 서로 동질한 데이터들을 같은 그룹으로 묶어주는 기법(고객 세분화 등)

- K-means 분류기법(K-means clustering)

⑤ 전처리 기법(Preprocessing Methods)

예측모형의 성과를 향상시키기 위해 입력데이터에 대해 사전 처리를 수행하는 기법

- 주성분분석(PCA, Principal Component Analysis)
- 퍼지이론(Fuzzy theory)

B. 데이터 전처리

1) 결측값(Missing Value)의 처리

너무 많은 항목이 비어 있는 변수나 너무 많은 항목이 비어 있는 레코드는 그 자체를 삭제

기타 나머지 항목에 대해서는 일반적으로 다음과 같은 값으로 대체

평균값(Mean) / 중앙값(Median) / 최빈치(Mode)

평균값 : (1+2+3+4+4+5+6+6+6+7+8) / 11 = 4.727

중앙값 : 5

최빈치 : 6

2) 정성적 변수의 정량화

각 속성은 단일변수값(atomic value)을 갖도록 수정

정성적 변수의 경우, 0/1의 binary code로 변환해야 추후 해석이 가능

예) 주소의 변환, 성별의 변환 등

3) 이상치(Outlier)의 제거

상식적으로 말이 안되거나 잘못 입력된 것으로 추정되는 변수값을 조정

일괄적으로 상위10%와 하위 10%에 해당하는 값들을 단일값으로 부여하는 경우도 있음

예) 체중 80Kg이상은 무조건 80Kg로,체중 45Kg 이하는 무조건 45Kg으로

4) 새로운 파생변수 개발

기존의 변수를 조합하여 새로운 변수를 개발

본래는 비율변수인 변수를 의미있는 정보로 구간화하여, 새로운 명목변수로 만듦

5) 정규화(Normalization)

모든 입력변수의 값이 최소 0에서 최대 1사이의 값을 갖도록 조정하거나, 평균 0을 갖는 표준정규분포를 갖도록 값을 조정하는 것

정규화 공식 (Min-Max Normalization)

(x-최소값) / (최대값-최소값)

예를 들어 전체 고객 중 체중이 가장 작은 사람이 40Kg, 가장 큰 사람이 120Kg이라고 하면,

40Kg → 0으로 변환

120Kg → 1로 변환

80Kg → (80-40)/(120-40) = 40/80 = 0.5로 변환

6) 자료의 구분

① 과적합화(Overfitting)의 발생 가능성

다음날의 주가지수를 예측하는 모형 A와 B가 있다.

A는 모형을 구축한 날까지의 주가(과거주가)는 99.99% 맞춘다. 그런데, 그 다음

날부터 주가지수를 예측시켜보니 70%를 맞추었다.

모형 B는 과거주가는 83% 맞추는데, 미래주가는 78% 맞춘다.

A, B중 더 잘 구축된 모형은?

② 과적합화의 예방법 : 모형 구축시, hold-out data의 개념을 도입

Hold-out data (검증) : 모형이 일반성을 갖는지 확인하기 위해 남겨두는 unknown data

통계 모형을 구축할 때, 전체 데이터가 100이라면 학습:검증=8:2 혹은 7:3의 비중으로 자료를 미리 나누어 둠

③ 0/1 예측의 경우 0과 1의 비중이 각 데이터셋마다 1:1의 비중이 되도록 섞어야 함

7) 모형에 들어갈 후보 입력변수 선정

카이제곱 검정(Chi-square Test)

독립표본 t검정 (t-Test) – 이분류 모형의 경우에 사용

분산분석 (ANOVA) – 다분류 모형의 경우에 사용

기법	대상변수A	대상변수B	적용 예
카이제곱검정	이산형	이산형	성별과 구매여부사이에 유의한 관계가 있는가?
독립표본t검정	이산형 (2그룹)	연속형	체중과 구매여부 사이에 유의한 관계가 있는가? (구매자와 비구매자의 평균 체중이 크게 다른가?)
일원배치 분산분석	이산형 (3그룹 이상)	연속형	체중과 고객등급 사이에 유의한 관계가 있는가? (고객등급에 따라 평균 체중이 크게 다른가?)

C. 척도의 4가지 종류

명목척도, 서열척도 => 이산형 변수
등간척도, 비율척도 => 연속형 변수

1) 명목척도 (Nominal Scales)

① 정의 : 대상의 특성을 분류하거나 확인할 목적으로 사용하는 척도

② 예시 : 축구선수 등번호, 입시에 사용되는 출신학교 코드

2) 서열척도 (Ordinal Scales)

① 정의 : 측정대상간의 순서를 밝히기 위해 사용하는 척도(양적 비교 불가)

② 예시 : 수학시험 석차, 선호도 우선순위

3) 등간척도 (Interval Scales)

① 정의 : 부여된 순위 사이의 간격이 동일한 척도

② 예시 : 온도

4) 비율척도 (Ratio Scales)

① 정의 : 순위 사이의 간격이 동일하고, 비율계산이 가능한 척도

② 예시 : 무게, 투표율

D. 통계학의 개요

1) 통계 및 통계 분석

① 통계학(Statistics) 이란?

현황을 가능한 정확히 알아내고 미래에 대한 적절한 예측을 하기 위해 자료를 효율적으로 수집 정리하고 분석하는 학문분야

② 통계분석의 목적

현황을 제대로 파악하기 위해
미래를 예측하기 위해

③ 통계학이 적용되는 대표적인 예들

공학자 : 현재 생산되는 제품의 품질이 어떤 상태에 있는가?
경영자 : 제품에 대한 소비자의 선호도가 어떻게 나타나고 있는가?
사회과학자 : 정치에 대한 국민들의 의견은 어떻게 나타나고 있는가?

2) 모집단과 표본

우리나라 국회의원들이 소유하고 있는 모든 주택의 평균 가격은 얼마인가?

① 모집단 (population)

관심의 대상이 되는 집단 전체
(예) 우리나라 국회의원들이 소유하고 있는 모든 주택

모수 (parameter)
모집단이 가지고 있는 특징을 나타내는 수치
(예) 평균 가격

전수조사 (census)
모집단의 모든 개체를 전부 조사하는 방법
모수의 값을 가장 정확하게 알아내는 방법 : 시간과 비용 문제

② 표본 (Sample)

모수의 값을 알아내기 위해 추출된 모집단의 일부분

일반적인 통계분석에서는 모집단의 일부만 조사하여 모집단의 모수를 추측

100% 정확하지는 않지만, 거의 근접한 값을 제공해 줄 수 있음

대신 시간과 비용을 크게 절약할 수 있는 방법

③ 표본 추출

표본이 모집단을 대표할 수 있도록 추출하는 것이 중요

표본추출이 잘못되어 발생한 해프닝 사례

미국의 1936년 대선 → 민주당 Franklin Roosevelt vs. 공화당 Alf Landon

전화 조사 결과 공화당 승리 예상, 실제로는 민주당 승리

당시에는 전화의 보급률이 매우 낮았음

3) 통계학의 2가지 분야

① 기술통계학

수집된 자료들이 가지고 있는 정보를 손쉽게 파악할 수 있도록 표현하는 방법에 대한 내용을 다루는 학문분야

② 추리통계학

수집된 자료들을 분석해 볼 때 알아내고자 했던 모집단의 모수는 얼마로 추정되는지, 미래의 예측치는 얼마라고 예상되는지 등의 결론을 도출하는 학문분야

수집된 자료를 통하여 원래 알아보고자 했던 문제에 대한 결론을 도출

통계학의 대부분 내용들이 여기에 해당

4) 통계 분석의 절차

① 문제의 올바른 파악

대상이 되는 모집단에 대해 정확하게 설정

알아내려는 사항이 무엇인지 제대로 파악

② 표본의 추출과 자료의 수집

비표본오차(nonsampling error)가 발생하지 않도록 설계

표본오차 : 모집단을 표본으로 추정하는데서 발생되는 오차 (피할 수 없음)

비표본오차 : 표본추출이 잘못할 경우 발생되는 오차 (피할 수 있음)

자료 수집의 2가지 방법

실험 (experiment) : 자연과학 분야

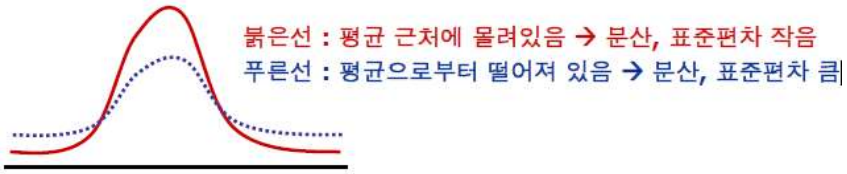
설문 (survey) : 사회과학 분야

③ 수집된 자료의 적절한 정리 (전처리) : 기술통계학

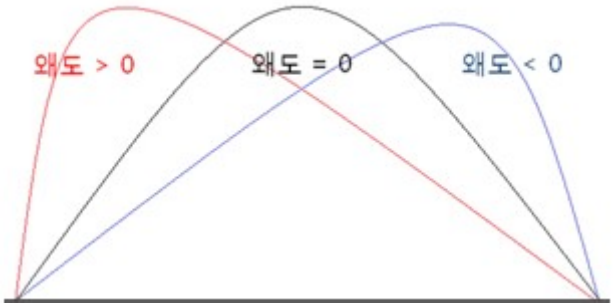
④ 모집단에 대한 추론 : 추리통계학

E. 기술통계량

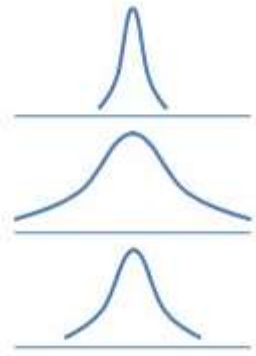
응답자의 평균 소득은 얼마일까?
응답자 나이의 평균은 몇 살인가?
특정 상품 선호도에 대한 관측치들은 평균을 중심으로 얼마나 퍼져있는가?
→ 분산, 표준편차를 통해 확인 가능



왜도(Skewness) : 분포의 편중(치우침)을 나타내는 지표



첨도(Kurtosis) : 분포의 뾰족한 정도를 나타내는 지표



1) 카이제곱 검정(교차분석)

① 이산형 변수로 구성된 2개의 변수가 서로 독립인지 아닌지를 확인하고자 할 때 사용

② 즉, 두 이산형 변수가 서로 상관관계가 있는지, 없는지를 통계적으로 검정할 때 사용되는 기법

③ 이산형(discrete) 변수 : 비계량적인 변수로서, 관측대상을 범주로 나누어 분류한 후, 그에 따라 기호나 숫자를 부여한 변수. 여기서는 평균의 의미가 없다.

(예) 성별, 선호정당, 출신국가, 직업구분, 주택보유 여부 등

④ 연속형(continuous) 변수 : 계량적인 변수로서, 일반적인 수치적 방법에 의해 측정 가능한 변수. 여기서는 평균이 의미가 있다.

(예) 금액, 거리, 무게, 시간 등

⑤ 카이제곱 검정의 해석

점근 유의확률(p-value라고 함)을 토대로 평가한다.

"95% 신뢰수준 하에서" 로 해석하고 싶으면, 유의확률 < 0.05 인지 확인

"90% 신뢰수준 하에서"로 해석하고 싶으면, 유의확률 < 0.1 인지 확인

(즉, 유의확률은 작으면 작을수록 좋다.)

2) 가설 검정의 기초

① 과학 분야에서의 증명 : 반증법에 의거해 증명

"모든 사람은 정직하다"라는 명제가 있을 때 이 명제가 참인지 거짓인지를 확인하는 접근법에는 2가지가 존재

(1) 모든 사람을 일일이 조사해서 정직한지 확인하는 방법

(2) 정직하지 못한 사람(사례)을 하나 찾아내 명제가 거짓임을 입증하는 방법

위 2가지 방법 중 어느 것이 효율적일까? (현실적으로 가능한 방법일까?)

1000명이 정직함을 확인했을 때, 1001명째 사람이 정직하다고 확언할 수 있을까?

② 귀무가설(H0)과 대립가설(Ha)

우리가 알고 싶은 명제(주로 A와 B는 다르다, A와 B는 차이가 있다 등)는 대립가설로 두고, 그 반대인 귀무가설(주로 A와 B는 같다, A와 B는 차이가 없다 등)이 기각됨을 보임으로서 내가 입증해 보이고 싶은 명제를 증명

3) 독립표본 t-검정

서로 독립된 두 집단간의 평균의 차이가 통계적으로 유의미한지 비교하고자 할 때 사용

즉, 서로 독립된 두 집단에 대해 각 집단별 특정 연속형 변수 평균값이 서로 차이가 있는지 없는지를 통계적으로 검정할 때 사용되는 기법

예) 전체 응답자 중 남자와 여자 사이의 연령은 차이가 있는가?

4) 대응표본 t-검정(Paired-samples t-test)

서로 동일한 모집단에서 추출된 두 표본에 대해 특정 연속형 변수 평균값이 서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용되는 기법

예) 한 회사에서 자사가 개발한 한 달간의 식이요법 프로그램이 효과가 있는지 여부를 분석하고자 함

5) 일원배치 분산분석(One-way ANOVA)

세 개 이상의 집단간의 평균의 차이가 통계적으로 유의미한지 비교하고자 할 때 사용

예) 학력수준에 따라 직무만족도의 수준은 차이가 있는가?

6) 가설검정의 정리

분석의 대상이 되는 두 변수의 특징에 따라, 다음과 같이 다른 검정기법을 적용해야 함

기법	대상변수A	대상변수B	적용 예
카이제곱 검정	이산형	이산형	성별과 구매여부 사이에 유의한 관계가 있는가? 성별과 결혼유무 사이에 유의한 관계가 있는가?
독립표본 t검정	이산형 (2그룹)	연속형	체중과 구매여부 사이에 유의한 관계가 있는가? (구매자와 비구매자의 평균 체중이 크게 다른가?) 성별에 따른 평균 취업률의 차이가 있는가?
대응표본 t검정	이산형 (2그룹/ Pair)	연속형	보충수업 후 성적의 향상이 있는가?
일원배치 분산분석	이산형 (3그룹 이상)	연속형	체중과 고객등급 사이에 유의한 관계가 있는가? (고객등급에 따라 평균 체중이 크게 다른가?) 거주지역에 따른 평균소득액의 차이가 있는가?

F. 실습예제

1) 기초통계량

```
import numpy as np
from scipy import stats

data=np.array([4,5,1,2,7,2,6,9,3])
#평균
a=np.mean(data)
print(a)
```

```
4.333333333333333
```

```
#중위수
b=np.median(data)
print(b)
```

```
4.0
```

```
#최빈값
c=stats.mode(data)
print(c[0][0])
```

```
2
```

```
from statistics import variance,stdev

points=np.array([20,80,90,95,87,89,95,99,97,100,60,70,77,
88,89,89,90])

#분산
a=variance(points)
print(a)
```

```
374
```

```
#표준편차
b=stdev(points)
print(b)
```

```
19.339079605813716
```

```
#범위
c=np.max(points) - np.min(points)
print(c)
```

```
80
```

```
#최대,최소값
a=np.max(points)
b=np.min(points)
c=a-b
print(a)
print(b)
print(c)
```

```
100
20
80
```

```
# 백분위
for val in [20,80,100]:
    d=np.percentile(points,val)
    print(str(val)+"%", d)
```

```
20% 77.6
80% 95.0
100% 100.0
```

```
#사분위수
a,b,c=np.percentile(points,[25,50,75])
print(a) #1사분위수
print(b) #2사분위수(중위수)
print(c) #3사분위수
print(c-a) #IQR(InterQuartile Range)
```

```
80.0
89.0
95.0
15.0
```

2) 카이제곱 검정

```
#data1과 data2가 같은지 다른지 알고 싶다.
#귀무가설 : 두 데이터는 차이가 없다.
#대립가설 : 두 데이터는 차이가 있다.
from scipy import stats

data1 = [4,6,17,16,8,9]        # 관측치
data2 = [10,10,10,10,10,10]    # 기대치
chis = stats.chisquare(data1, data2)
#검정 통계량과 p-value
chis
# pvalue가 0.05보다 작으므로 귀무가설 기각, 대립가설 채택
# 두 데이터는 차이가 있다
```

```
Power_divergenceResult(statistic=14.200000000000001,
pvalue=0.014387678176921308)
```

```
import pandas as pd
from scipy import stats

survey=pd.read_csv("d:/data/smoke/survey.csv")
#survey
#Smoke와 Exer 필드를 기준으로 집계
data=pd.crosstab(survey.Smoke, survey.Exer)
print(data)
#카이제곱검정 수행
result=stats.chi2_contingency(observed=data)
print(result[0]) #검정통계량
print(result[1]) #p-value
#p-value가 0.48로 0.05보다 크므로 흡연습관과 운동횟수에는 상관관계가 없다.
```

```
Exer   Freq  None  Some
Smoke
Heavy     7     1     3
Never    87    18    84
Occas    12     3     4
Regul     9     1     7
5.488545890584232
0.48284216946545633
```

3) 단일표본 t검정

```
#전체 학생들 중 20명의 학생들을 선택하여 전체 학생들의 평균키가 175cm인지 아닌지 알고 싶다.
#귀무가설 : 학생들의 평균키는 175cm이다.
#대립가설 : 학생들의 평균키는 175cm가 아니다.
import numpy as np
from scipy import stats
 #랜덤 시드 설정(같은 결과가 나옴)
np.random.seed(1)
 # np.random.normal(0, 5) : 평균 0, 표준편차 5인 난수
heights = [180 + np.random.normal(0, 5) for a in range(20)]
#print(heights)
result = stats.ttest_1samp(heights, 175)
print("검정통계량 : %.3f , p-value : %.3f" % result)
# p-value가 0.05보다 작으므로 95% 신뢰수준 하에서
# 학생들의 평균키는 통계적으로 유의하게 차이가 난다고 할 수 있음
# 따라서 귀무가설을 기각하고 대립가설을 채택한다. 즉, 학생들의 평균키는 175cm가 아니다.
```

검정통계량 : 3.435 , p-value : 0.003

4) 독립표본 t검정

```
#그룹1과 그룹2에서 각각 20명의 학생들을 선택하여 평균키가 같은지 다른지 알고 싶다.
#귀무가설 : 두 그룹 학생들의 평균키는 같다.
#대립가설 : 두 그룹 학생들의 평균키는 같지 않다.
import numpy as np
from scipy import stats

#랜덤 시드 설정(같은 결과가 나옴)
#np.random.seed(1)
 #그룹1 : 평균 170, 표준편차 5
group1 = [170 + np.random.normal(0, 5) for a in range(20)]
#그룹2 : 평균 175, 표준편차 10
group2 = [175 + np.random.normal(0, 10) for a in range(20)]
print(group1)
print(group2)
print("group1의 평균:",np.mean(group1))
print("group2의 평균:",np.mean(group2))
result1 = stats.ttest_ind(group1, group2)
print("검정통계량 : %.3f , p-value : %.3f." % result1)
# p-value가 0.05보다 작으므로 95% 신뢰수준 하에서
# 두 그룹 학생들의 평균키는 통계적으로 유의하게 차이가 난다고 할 수 있음
# 따라서 귀무가설을 기각하고 대립가설을 채택한다. 즉, 그룹1과 그룹2 학생들의 평균키는 같지 않다.
```

```
검정통계량 : -2.329 , p-value : 0.025.
```

5) 대응표본 t검정

```
#다이어트 약을 복용한 사람들 중 20명을 선택하여 복용 전후의 체중 차이가 유의미한지 알고 싶다.
#귀무 가설: 복용 전후의 체중 차이가 없다.
#대립 가설: 복용 전후의 체중 차이가 있다.
import numpy as np
from scipy import stats
 #랜덤 시드 설정(같은 결과가 나옴)
np.random.seed(1)
 #복용전 : 평균 60, 표준편차 5
before = [60 + np.random.normal(0, 5) for _ in range(20)]
print(before)
#복용후 : 평균 : 복용전체중 x 0.99, 표준편차 0.02
after = [w * np.random.normal(0.99, 0.02) for w in before]
print(after)
#대응표본 t검정
result = stats.ttest_rel(before, after)
print("검정통계량: %.3f , p-value : %.3f" % result)
# p-value가 0.05보다 작으므로 95% 신뢰수준 하에서
# 다이어트약 복용전후의 체중은 통계적으로 유의하게 차이가 난다고 할 수 있음
# 따라서 귀무가설을 기각하고 대립가설을 채택한다. 즉, 다이어트약 복용전후의 체중 차이가 있다.
```

```
[68.1217268183162, 56.94121793174962, 57.359141238682724,
54.63515688921915, 64.3270381466234, 48.492306515598585,
68.7240588210824, 56.19396549552449, 61.59519548028549,
58.75314812261295, 67.31053968522487, 49.69929645251173,
58.38791397993246, 58.07972822665792, 65.66884721167719,
54.500543663429845, 59.13785896224782, 55.61070791039314,
60.21106873357797, 62.914076068579114]
```

[65.94098797171107, 57.67544499710449, 57.819839215935254,
54.63788246116392, 64.84275566541605, 47.34427263203588,
67.86790793112998, 54.58033393455883, 60.64923115285458,
58.78881770728494, 65.70631311881824, 48.80793606499072,
57.001583230054315, 56.517144665313666, 64.13055954741337,
53.94173367626836, 57.224973536332286, 55.315321289307555,
61.60772730564428, 63.218635763189475]

검정통계량: 2.915 , p-value : 0.009

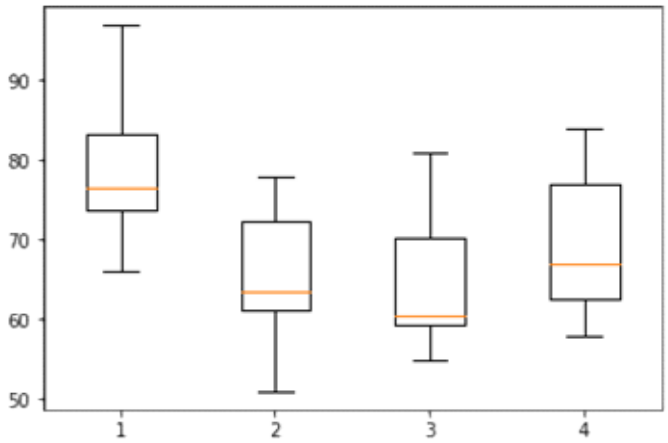
6) 아노바 분석

```
# 김부장은 4개의 각기 다른 신입사원 교육훈련 기법의 효과성을 평가하고자 한다.
# 새로 입사한 32명의 신입사원에게 4가지 기법을 임의로 적용시켜 교육을 시켰다.
# 한 달간의 훈련기간이 끝난 후 표준 시험을 쳤는데 그 점수는 아래와 같다.
# 4개의 교육훈련 기법간 차이가 있는가? 만약 있다면 어떻게 다른가?
#3개 이상의 대응표본을 비교해야 하므로 일원배치 분산분석 기법을 사용해야 한다.
#귀무가설 : 4개의 교육훈련 기법간의 차이가 없다.
#대립가설 : 4개의 교육훈련 기법간의 차이가 있다.
import scipy.stats as stats
import matplotlib.pyplot as plt
%matplotlib inline
a = [66,74,82,75,73,97,87,78]
b = [72,51,59,62,74,64,78,63]
c = [61,60,57,60,81,55,70,71]
d = [63,61,76,84,58,65,69,80]

print("a 평균:",np.mean(a))
print("b 평균:",np.mean(b))
print("c 평균:",np.mean(c))
print("d 평균:",np.mean(d))

# matplotlib plotting
plot_data = [a,b,c,d]
plt.boxplot(plot_data)
plt.show()
```

```
f, p = stats.f_oneway(a, b, c, d)
print(f, p)
# 결과 분석 : p-value가 0.05보다 작으므로
# 95% 신뢰수준 하에서 두 집단간 평균은 통계적으로 유의하게 차이가 있다고 할 수 있다.
# 따라서, 통계적으로 볼 때 4개의 교육훈련 기법의 효과에 차이가 있다고 잠정적으로 결론지을 수 있다.
# 상자수염그림(boxplot)
```



```
import pandas as pd
from scipy import stats

#귀무가설: 세가지 비료의 수확량은 차이가 없다.
#대립가설: 세가지 비료의 수확량은 차이가 있다.

data = pd.read_csv("d:/data/anova/fertilizers.csv")
#비료1,비료2,비료3
print(data)
result = stats.f_oneway(data["fertilizer1"],
data["fertilizer2"], data["fertilizer3"])
print(result)
print(result[0]) #통계량
print(result[1]) #p-value
#p-value가 0.048이므로 귀무가설을 기각하고 대립가설을 채택한다.
#따라서 세가지 비료의 수확량은 차이가 있다.
```

```
   fertilizer1  fertilizer2  fertilizer3
0           72           54           48
1           62           56           62
2           90           58           92
3           42           36           96
4           84           72           92
5           64           34           80
F_onewayResult(statistic=3.7551268418654105,
pvalue=0.04762461989261837)
3.7551268418654105
0.04762461989261837
```