PYTHON을 이용한 데이터 분석

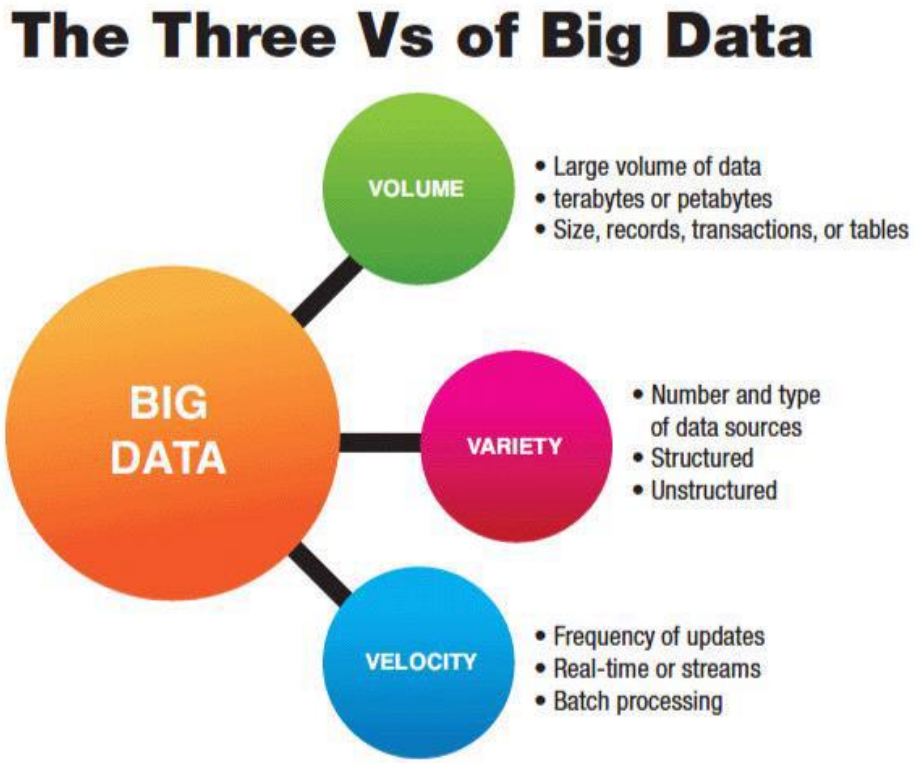
Park.yk

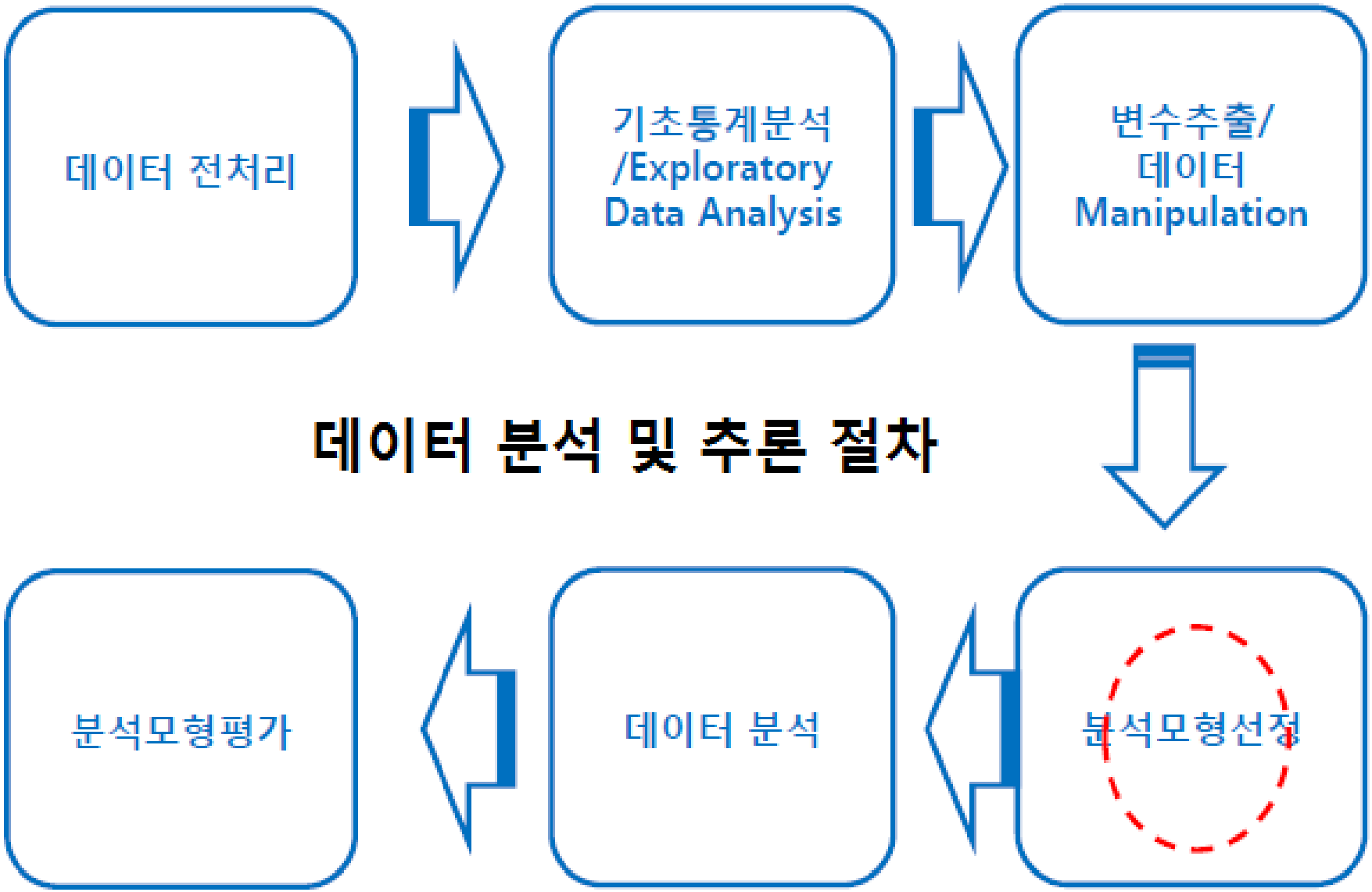
Big Data 정의

- 더 나은 의사결정, 시사점 발견 및 프로세스 최적화를 위해 사용되는 새로운 형태의 정보처리가 필요한 대용량, 초고속 및 다양성의 특성을 가진 정보자산(Gartner)

- 일반적인 데이터베이스 소프트웨어 도구가 수집, 저장, 관리, 분석하기 어려운 대규모의 Data.

- 3V를 가진 Data(Velocity, Volume, Variety)




데이터 전처리
기초통계분석
/Exploratory
Data Analysis
변수추출/
데이터
Manipulation
분석모형선정
데이터 분석
분석모형평가

데이터 분석 및 추론 절차

통계학의 분류	
기술 통계학 (Descriptive statistics)	자료를 변수 별로 따로따로 또는 관계되는 변수끼리 묶어서 요약. (평균, 분산, 표준편차, 사분위수, 중위수 등) 기술 통계는 추론 통계의 기초작업 수행을 위한 과정이라 할 수 있다.
추론 통계학 (Inferential statistics)	정리된 자료에 담긴 의미를 해석하여 미지의 세계에 대해 추론. (상관분석, 회귀분석, 분류, 인공신경망, 딥러닝 등)

* 데이터 분석을 위한 라이브러리 모음

참조 사이트 : pydata (http://pydata.org)

- numpy - 고속 연산
- scipy - 과학 분석 알고리즘
- pandas -데이터 표현 및 처리
- matplotlib - 시각화 도구
- scikit-learn - Machine Learning, 추론 통계처리
- Tensorflow - 기계학습, Deep Learning

*** NUMPY ***

- 데이터는 수 많은 숫자들로 이루어져 있다. 이 많은 숫자들을 효율적으로 계산하기 위해서는 관련된 데이터를 모두 하나의 변수에 넣고 처리해야 한다. 하나의 변수에 여러 개의 데이터를 넣는 방법으로 파이썬의 리스트를 사용할 수도 있지만 리스트는 속도가 느리고 메모리 소모가 크다. 더 적은 메모리로 더 빠르게 데이터를 처리하려면 배열을 사용하는 것이 좋다. 배열 사용을 위한 표준 패키지로 numpy가 있다.
- numpy는 수치 해석용 파이썬 패키지다. 다차원의 배열자료 구조인 ndarray 클래스를 지원하며, 벡터와 행렬을 사용하는 선형대수 계산에 주로 사용한다.

- Numpy는 파이썬에서 과학적 계산을 위한 핵심 라이브러리다.
- 고성능 다차원 배열 객체와 이들 배열과 함께 작동하는 도구들을 제공한다.
- 배열(ndarray)을 만들거나 다루기 위한 많은 함수를 제공한다.
- ndarray는 모두 같은 타입을 가진 값들의 grid이며, 양의 정수 튜플로 인덱스 되어 있다.
- ndarray는 다차원 배열객체(numpy.ndarray)로 파이썬의 리스트형 자료를 처리하는 빠르고 유연한 자료구조다.
- 차원의 수는 배열의 rank이고, 배열의 shape은 각 차원의 크기를 알려주는 정수의 튜플이다.

NUMPY

- **import numpy as np**

  numpy 라이브러리를 np라는 이름으로 import 시킨다.
- np.array([1, 2, 3]) : 파이썬의 리스트 자료를 통해 rank가 1인 배열(1차원)을 생성한다
- 슬라이싱(Slicing) : numpy로 만든 배열도 슬라이싱이 가능하다.
- 모든 numpy 배열은 같은 타입을 갖는 요소의 grid이다.
- numpy는 배열을 구성하는데 사용할 수 있는 숫자 자료타입의 큰 집합을 제공한다.
- numpy는 배열을 만들 때, 자료타입을 추측한다. 그러나 배열을 구성할 때는 명시적으로 자료타입을 선택적 인자로 포함하여 만들 수도 있다.
- Python의 list 대비 numpy의 ndarray는?

  C언어의 배열 처럼 연속적인 메모리 배치를 가지기 때문(원소접근과 속도 빠름)에 모든 원소가 같은 자료형이어야 한다.

- **배열 연산(Array math)**

  - 기본적인 수학 함수는 배열에 요소별(elementwise)로 적용되고,
    그냥 연산자(+, -, *, /)나 혹은 numpy 모듈의 함수(add,subtract,multiply,divide)로 사용할 수 있다.
  - 벡터화 연산을 하므로 for문을 사용하지 않고 바로 배열에 대한 연산이 가능하다.

NUMPY

- **Transpose(전치)**

배열에서 모양을 바꿀 필요가 있다. 이를 전치라 하며 데이터 모양이 바뀐 뷰를 반환. 이러한 연산의 가장 간단한 예는 행렬을 전치하는 것이며, 배열 객체의 T 속성을 사용해 처리한다.

- **브로드캐스팅 : Broadcasting은 크기가 다른 배열 간의 연산을 말한다.** 작은 배열과 큰 배열이 있을 때 작은 배열을 여러 번 반복해 큰 배열에 연산을 수행할 수 있다.

두 배열의 브로드캐스팅은 다음 규칙을 따른다.

1. 두 배열의 rank가 같지 않다면, 모양이 같은 길이를 가질 때까지 적은 rank 배열의 모양을 붙인다.

2. 두 배열이 그 차원에서 같은 크기를 갖거나, 차원에서 배열들 중 하나의 크기가 1이라면 그 두 배열은 차원에서 호환 가능하다고 말한다.

3. 만약 그들이 모든 차원에서 호환 가능하다면 그 배열들은 함께 브로드캐스트 할 수 있다.

4. 브로드캐스팅 후에, 각 배열은 두 입력 배열들의 모양이 동일한 것처럼 행동한다.

5. 어느 차원에서 한 배열의 크기가 1이고 다른 배열의 크기가 1보다 크다면, 그 첫 배열은 복사된 것처럼 행동한다.

*** PANDAS ***

- 고수준의 자료구조와 빠르고 쉬운 데이터 분석용 자료구조 및 함수를 제공한다. 이는 NumPy의 고성능 배열 계산 기능과 스프레드시트, SQL과 같은 관계형 데이터베이스의 유연한 데이터 조작기능을 조합한 것이다. 아울러 세련된 인덱싱 기능으로 쉽게 데이터를 재배치하여 집계 등의 처리를 편리하게 한다.

행렬은 수의 사각형 배열이다!

⇧

Scalar : 행렬이나 벡터의 각원소(실수)

Series : vector - 1차원 배열

DataFrame : marix - 2차원 배열

) 생성용 클래스

*** PANDAS ***

- **Series**는 일련의 객체를 담을 수 있는 1차원 배열과 같은 자료구조로 색인을 갖는다.

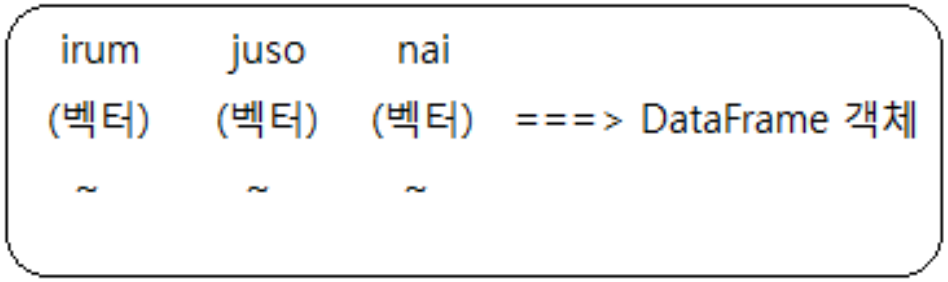
```
obj = Series([3, 7, -5, 4])
# list, tuple type 가능. TypeError:'set' type is unordered

obj2 = Series([3, 7, -5, 4], index=['a', 'b', 'c', 'd'])
# 생성 시 색인을 지정
```

- 파이썬 dict type의 자료로 Series 객체를 생성")

```
names = {'mouse':12000, 'keyboard':25000, 'mornitor':'450000'}
```

*** PANDAS ***

DataFrame : 표 모양(2차원 형태 자료)의 자료구조로 여러 개의 칼럼을 갖는다. 각 칼럼은 서로 다른 종류의 값을 기억할 수 있다. 같은 길이의 리스트에 담긴 사전을 이용해 DataFrame 객체 생성. numpy의 array의 차이점이라면 각 칼럼 마다 type이 다를 수 있다.



- 작성 예)

```
from pandas import Series, DataFrame
data = {
  'irum':['홍길동', '한국인', '신기해', '공기밥', '한가해'],
  'juso':('역삼동', '신당동', '역삼동', '역삼동', '신사동'),
  'nai':[23, 25, 33, 30, 35],
}
```

#list, tuple type만 가능. data의 마지막 콤마는 있어도 상관없다.

*** PANDAS ***

- dict type의 자료로 DataFrame() 생성자를 호출하면 DataFrame 객체가 생성됨.
- DataFrame의 칼럼은 사전형식이나 속성형식으로 접근이 가능
- 순서를 변경할 수 있다.
- 칼럼에 값 대입으로 수정 가능
- Series를 대입하면 DataFrame의 색인에 따라 값이 대입되고, 없는 색인은 NaN이 된다.
- DataFrame의 행 또는 열 삭제. 속성으로 axis=0 행, axis=1 열“
- index명이나 열이름으로 정렬하기. axis=0 행, axis=1 열“
- DataFrame의 특정 열 값을 문자열 자르기
- 재색인할 때 값을 보간하거나 채워 넣기
- 산술연산
- 기술적 통계와 관련된 메소드
- 자료 합치기
- group by
- 파일 읽기/저장

웹 문서 읽기

* 웹에서 get 방식 요청

```
params = {'param1': 'value1', 'param2': 'value'}
res = requests.get(URL, params=params)
```

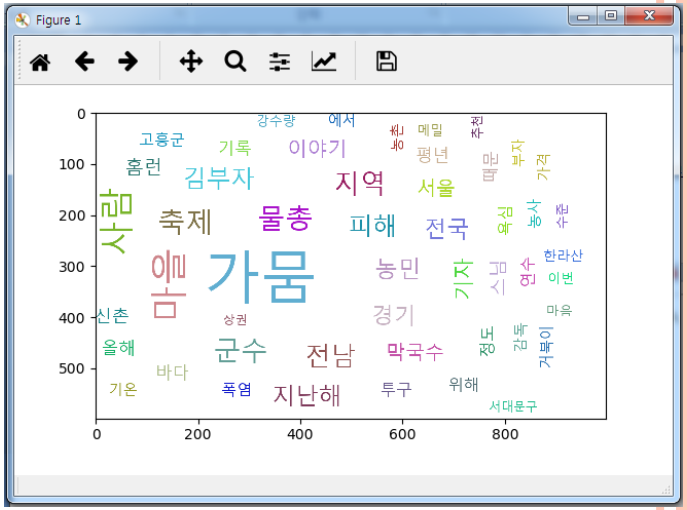
* 웹에서 post 방식 요청은 get 메서드를 post로 변경

- lxml.html.parse(io.StringIO(문자열)) : 웹에서 가져온 문자열을 트리 형식으로 변환
- 루트 객체 가져오기: 트리객체.getRoot()
- 태그에 해당하는 데이터를 Element의 list로 가져오기:

  루트객체.findall('.//태그명')

- 루트객체.find("from")은 루트객체 태그 하위에 from과 일치하는 첫 번째 태그를 찾아서 리턴하고, 없으면 None을 리턴하며 루트객체.findall(" from")은 루트객체 태그 하위에 from과 일치하는 모든 태그를 리스트로 리턴하고, 루트객체.findtext("from")은 루트객체 태그 하위에 from과 일치하는 첫번째 태그의 텍스트 값을 리턴한다.
- Element의 get('속성'): 속성 값 가져오기
- Element의 text_content() : 태그 안의 내용 가져오기

웹 문서 읽기



*** BeautifulSoup으로 워드클라우드 그래프 그리기**

아래 모듈을 먼저 설치한다.

~>pip install pytagcloud

~>pip install pygame

~>pip install simplejson

한글 글꼴 등록하기 : 패키지가 설치된 경로로 이동한다. 윈도우의 경우 다음과 같다.

C:\Python34\Lib\site-packages\pytagcloud\fonts

C:\Anaconda3\Lib\site-packages\pytagcloud\fonts

해당 경로로 가보면 ttf 폰트 파일들이 많이 있다. 이 곳에 한글을 지원하는 폰트를 복사한다.

그리고 이 디렉토리에 존재하는 fonts.json 을 텍스트 에디터로 열어 편집을 한다.

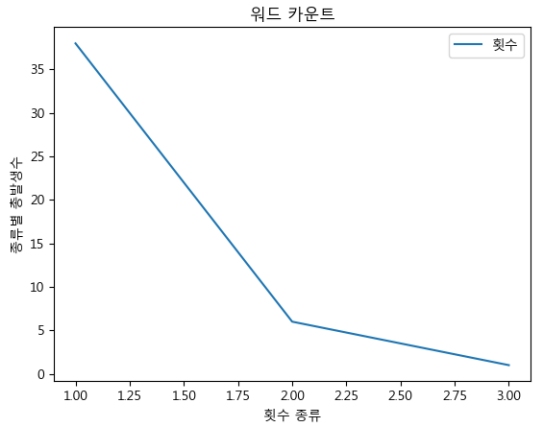
fonts.json ---------

```
{
   "name": "korean",
   "ttf": "malgun.ttf",
   "web": "http://fonts.googleapis.com/css?family=Nobile"
},
```

이 때 malgun.ttf 파일은 C:\Windows\Fonts 에서 "맑은 고딕"을 복사해

c:\...Lib\site-packages\pytagcloud\fonts 폴더에 붙여넣기 해 준다.

- html 자료를 읽어 한글 형태소 분석하기



- Database 연동 후에 자료를 읽어 DataFrame 객체화 하기

*** SCIPY ***

SciPy : NumPy 기반으로 만들어졌다.

NumPy 배열에 작동하는 많은 수의 함수를 제공하며, 과학적이고 공학적인 응용의 다른 타입들에 유용하다.

** MATPLOTLIB **

- Matplotlib는 플로팅 라이브러리이다.

- matplotlib.pyplot 모듈을 도입하여 그래프나 2차원 데이터 시각화 가능.

- 그래프 종류 : line, scatter, contour(등고선), surface, bar, histogram, box, ...

- http://matplotlib.org

- Figure : 모든 그림은 Figure라고 부르는 matplotlib.figure.Figure 클래스 객체 에 포함되어 있다. 내부 plot이 아닌 경우에는 하나의 Figure는 하나의 아이디 숫자와 window를 갖는다. figure()를 명시적으로 적으면 여러 개의 윈도우를 동시에 띄우게 된다.

*** MATPLOTLIB - SEABORN ***

matplotlib 의 기능 보충용 seaborn

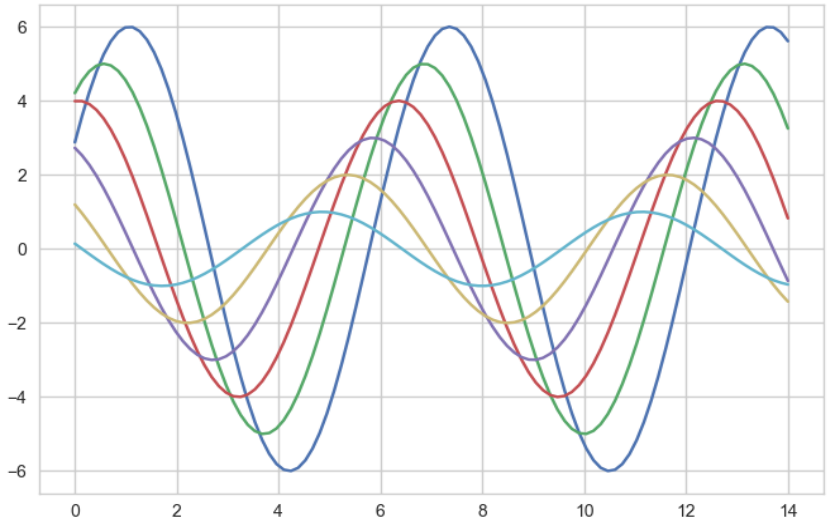
- matplotlib에 seaborn을 사용하면 그래프를 더 멋지게 표현할 수 있다.
- matplotlib 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 패키지다.

- seaborn 설치하기

~>conda install seaborn or
pip install seaborn

예)

```
import seaborn as sns
sns.set_style("whitegrid")
    ...
plt.show()
```



** 데이터 분석 **

- **데이터 분석의 목적**

데이터 분석이란 어떤 입력 데이터가 주어졌을 때 입력 데이터 간의 관계를 파악하거나 파악된 관계를 사용하여 원하는 출력 데이터를 만들어 내는 과정으로 볼 수 있다.

데이터 분석도 분석 목적에 따라 "예측(prediction)", "클러스터링(clustering)", "모사(approximation)" 등 여러 가지 문제가 있다. 여기에서는 널리 사용되는 예측 문제를 살펴보자

- **예측 :** 예측(prediction)은 데이터 분석 작업 중 가장 많이 사용되는 유형 중 하나이다. 예측이란 숫자, 문서, 이미지, 음성, 영상 등의 여러 가지 입력 데이터가 주어지면 데이터 분석의 결과로 다른 데이터를 출력하는 분석 방법이다. 예를 들어 다음과 같은 작업은 예측이라고 할 수 있다.

  - 부동산의 위치, 주거 환경, 건축연도 등이 주어지면 해당 부동산의 가치를 추정한다.
  - 꽃잎의 길이와 너비 등 식물의 외형적 특징이 주어지면 해당하는 식물의 종을 알아낸다.
  - 얼굴 사진이 주어지면 해당하는 사람의 이름을 출력한다.

** 데이터 분석 **

- 데이터 분석에서 말하는 예측이라는 용어는 시간상으로 미래의 의미는 포함하지 않는다. 시계열 분석에서는 시간상으로 미래의 데이터를 예측하는 경우가 있는데 이 경우에는 미래 예측(forecasting) 이라는 용어를 사용한다.

** 데이터 분석 **

- **입력 데이터와 출력 데이터**

예측 문제에서는 데이터의 유형을 입력 데이터(input data)와 출력 데이터(output data)라는 두 가지 유형의 데이터로 분류할 수 있어야 한다.

입력 데이터는 분석의 기반이 되는 데이터로 보통 알파벳 X로 표기한다. 다른 말로 독립변수(independent variable), 특징(feature), 설명변수(explanatory variable) 등의 용어를 쓰기도 한다.

출력 데이터는 추정하거나 예측하고자 하는 목적 데이터를 말한다. 보통 알파벳 Y로 표기하며 다른 말로 종속변수(dependent variable)라고 부른다. 라벨(label) 또는 클래스(class)라고 하기도 한다.

입력 데이터와 출력 데이터를 정확히 파악하는 것은 예측 문제를 구체화하는 첫 번째 단계이다. 특히 예측 성능은 이러한 입출력 데이터의 숫자와 종류에 크게 의존하기 때문에 정확히 어떠한 값을 가지는 입력을 몇 개 사용하겠다는 문제 정의가 예측 문제를 해결하는데 가장 중요한 부분이 될 수도 있다.

** 데이터 분석 **

- **규칙 기반 방법과 학습 기반 방법**

예측 문제는 어떻게 풀 수 있을까? 예측 방법으로는 규칙 기반(rule-based) 방법과 학습 기반(traing-based) 또는 데이터 기반(data-based) 방법이 있다. 규칙 기반 방법은 어떤 입력이 들어오면 어떤 출력이 나오는지를 결정하는 규칙 혹은 알고리즘을 사람이 미리 만들어 놓는 방법이다. 학습 기반 방법 또는 데이터 기반 방법은 이러한 규칙을 사람이 만드는 것이 아니라 대량의 데이터를 컴퓨터에게 보여줌으로써 스스로 규칙을 만들게 하는 방법이다. 여기에서는 규칙 기반 방법은 다루지 않으며 학습 기반 방법만을 다루도록 한다.

예를 들어 개를 찍은 이미지 데이터를 입력하면 "개"라고 출력하고 고양이를 찍은 사진을 입력하면 "고양이"라고 출력하는 예측 시스템을 만든다고 가정해 보자.

규칙 기반 방법을 사용하면 사진에서 눈 모양을 찾아내는 알고리즘을 넣고 눈동자가 세로 방향으로 길면 고양이이고 아니면 개라고 출력하는 규칙을 넣을 수 있다. 이렇게 사람이 세부적인 규칙을 알려주는 방법이 규칙 기반 방법이다.

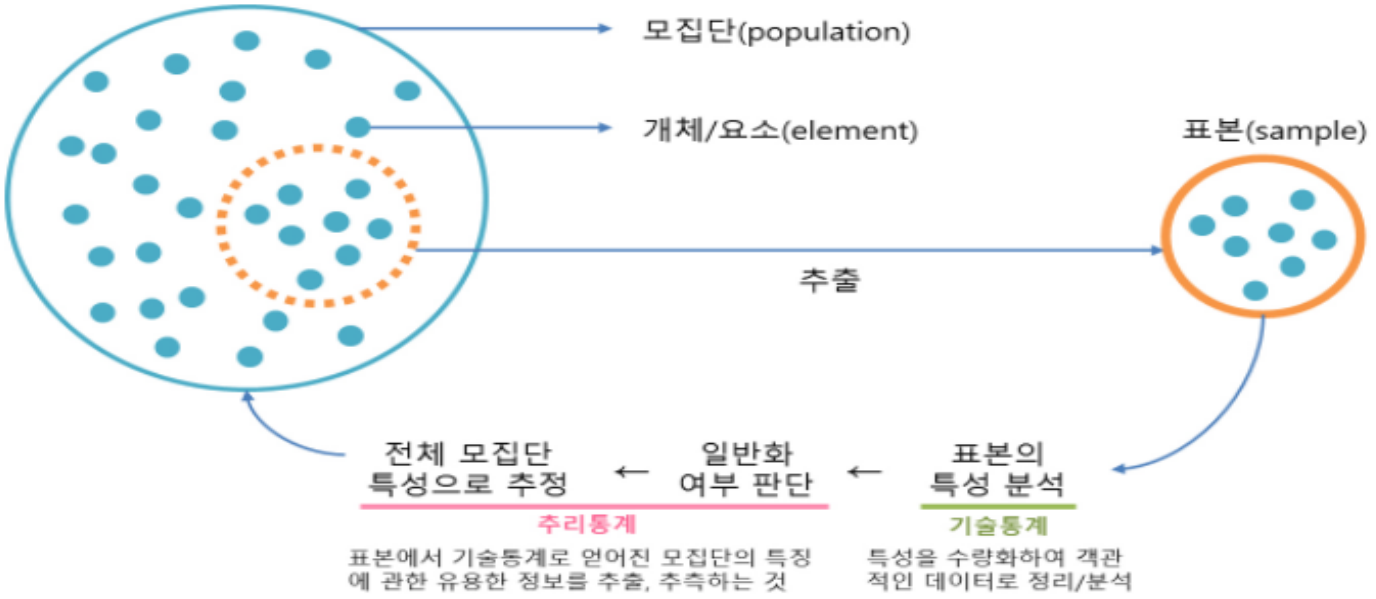
한편 학습 기반 방법은 이러한 규칙을 알려주지 않는 대신 많은 데이터를 주고 스스로 규칙을 찾도록 한다. 앞서 말한 고양이와 개의 구분 문제에서는 개와 고양이를 찍은 사진을 주고 스스로 적합한 규칙을 찾도록 하며 영어를 한국어로 번역하는 문제에서는 수많은 영어 문장과 이에 대응하는 한국어 문장을 주고 스스로 번역 방법을 찾도록 한다.

기술통계 & 추리통계의 개념 정의

통계분석은 크게 **기술통계**(descriptive statistics)와 **추리통계**(inferential statistics)으로 나눌 수 있다.

기술통계	추리통계
수집한 데이터의 주요 특성을 분석 및 기술하는 통계방법 ex) 평균값 (mean),중위수 (median),최빈수 (mode),최대값, 최소값,범위 (range),분산 (variance),표준편차 (standard deviation) 등	수집한 데이터에서 표본(sample)을 추출, 특성을 파악하여 전체 데이터(모집단)의 특성으로 일반화할 수 있는 지 여부를 판단 모집단의 특성을 추정하는 것이 목적 - 간단히 표본을 기초로 향후의 일을 예측하는 것에 초점. ex)선거철.. 후보자의 지지도 조사
사례) H대학교 A학부의 최근 5년 간 4학년 학생들의 과목별 성적을 분석해서 학생들의 성적변화 추세치를 보려고 한다...	사례) B제품의 생산공장에서 라인별 제품의 불량률을 알아보기 위해 일정한 시간 간격으로 제품을 추출하여 분석하려 한다....

기술통계는 추리통계의 기초작업을 수행하기 위한 과정

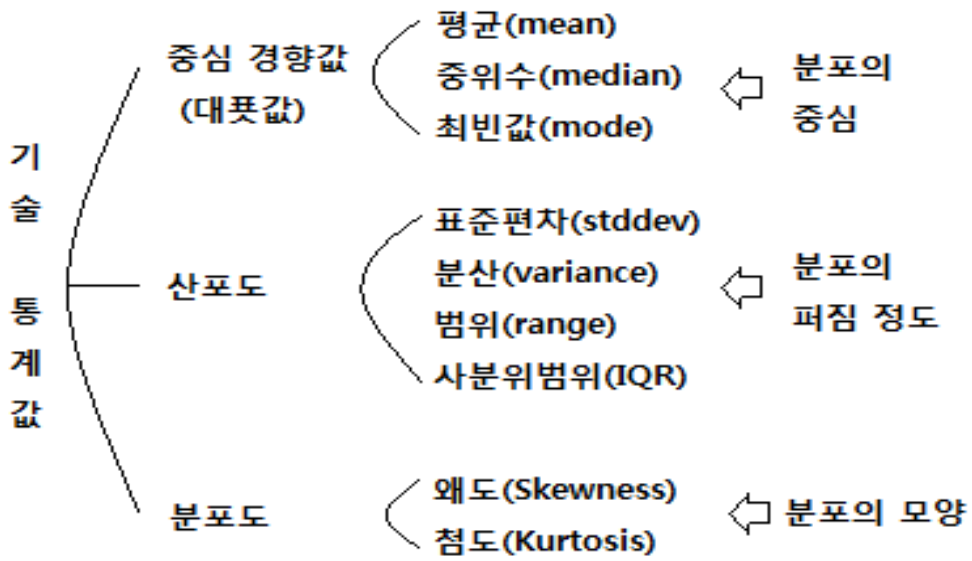


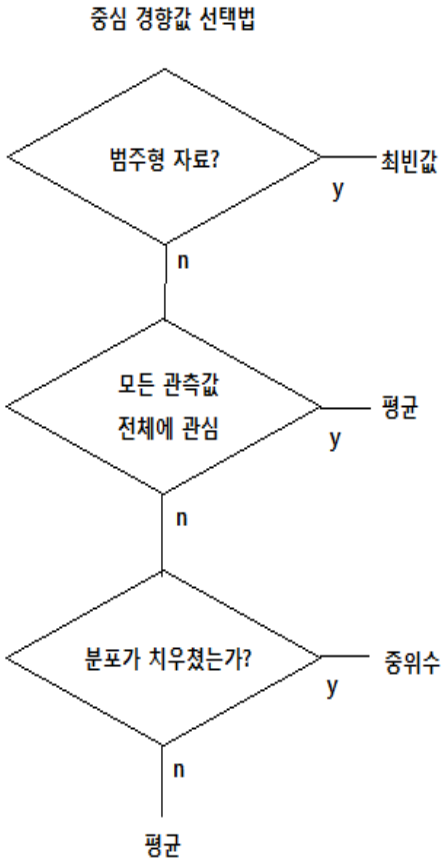
기술통계(Descriptive Statistics)

- 자료를 정리 및 요약하는 기초적인 통계
- 데이터 분석 전에 전체적인 데이터 분포의 이해와 통계적 수치 제공
- 추론통계의 기초자료로 많이 쓰인다.

기술통계량 유형 - 대표값, 산포도, 비대칭도 : 왜도, 첨도

기술 통계 분석 - 정보의 손실을 최대로 줄이면서 데이터를 효과적으로 요약할 수 있는 분석방법.





* 통계관련 기본 용어의 이해 *

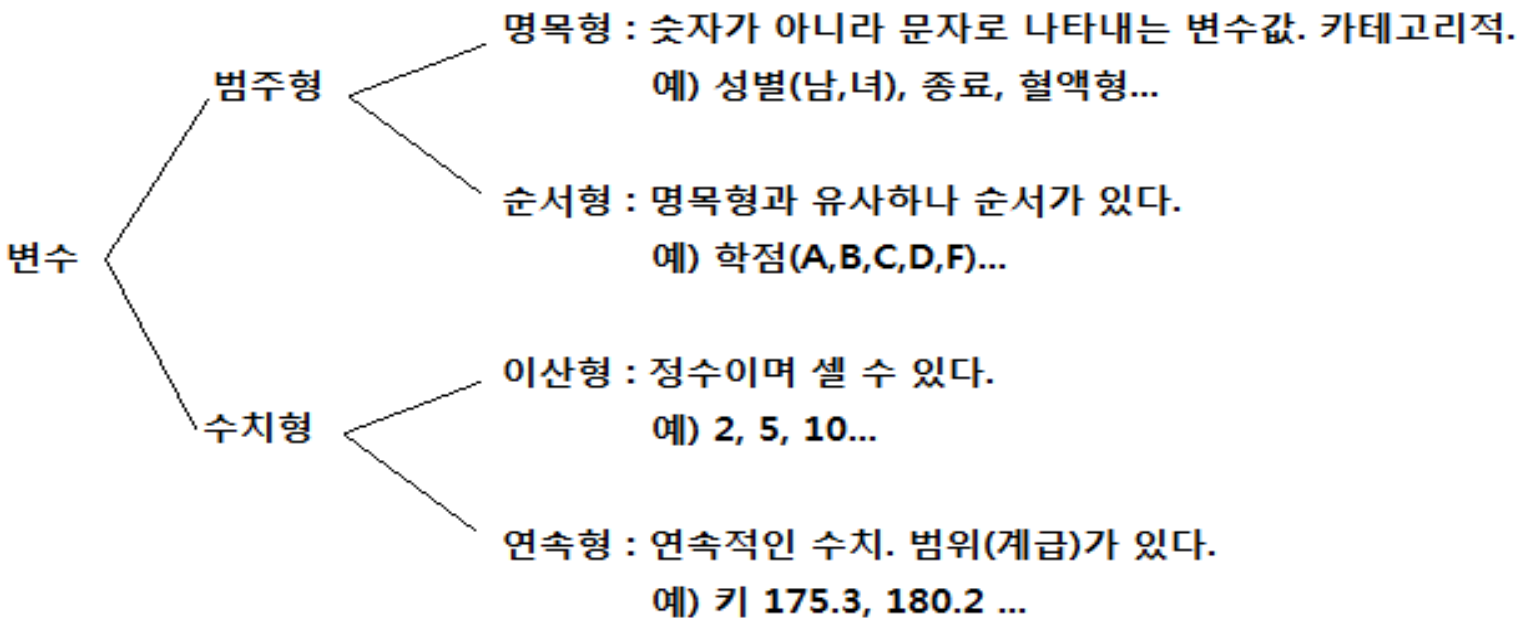
- 모집단의 통계수치를 모수라고 하고, 표본의 통계수치를 통계량이라고 한다.
- 오차란 무엇일까? 간단히 평균으로부터의 치우침 이라고 할 수 있다. 그리고 통계는 이러한 오차를 분석하고 관리하게 되는데 이러한 오차를 표현하는 대표적인 척도가 표준편차와 이의 제곱인 분산이다.
- 한편, 산포도는 변량이 흩어져있는 정도를 말한다. 변량들이 평균에 모여 있으면 산포도가 작다고 하고 변량들이 평균으로부터 떨어져있으면 산포도가 크다고 한다. 산포도를 수치로 나타내는 방법으로는 분산과 표준편차가 주로 쓰인다.
- 대푯값은 자료들을 대표하는 값으로 쓰인다. 자주 쓰는 대푯값은 평균, 최댓값, 최솟값 등이 있다. 분산이나 표준편차 같은 산포도는 왜 필요할까? 대푯값이 자료들 모두를 반영하지 못하기 때문이다. 예를 들면 어떤 친구는 체육을 잘하는데 영어는 형편없는 경우가 있다. 그런데 평균만을 써버리면 그 친구는 그냥 평범한 학생이 되어버린다. 그럼 산포도가 크면 변량들이 제 각각이라는 뜻이겠네? 맞다! 산포도가 크면 그만큼 예측이 어려워진다.
- 표준편차는 동일한 평균값을 갖는 둘 이상의 집단을 비교할 때 유용하게 활용된다. 예를 들어 어느 반 학생들의 기말시험 전체과목의 평균점수는 80점이라고 한다. 이 때에 A학생은 표준편차가 작고, B는 표준편차가 크다고 한다면 A학생의 수학점수는 80점 안팎으로 예측이 가능하지만 B학생의 수학점수는 예측하기가 어렵다. 그래서 통계적으로 결과를 추정할 때에는 모집단의 표준편차가 작을수록 보다 정확한 값을 예측할 수 있게 된다.

기술통계(Descriptive Statistics)

*** 척도(Scale) :**

- 자료가 수집될 때 관찰된 현상에 하나의 값을 할당시키기 위해 사용되는 측정의 수준
- "척도에 따른 분류"
  - 범주형(정성적 : 수량화가 불가 ex) 성별, 지역, 직업 등) - 명목형, 순서형(서열형)
  - 수치형(정량적 : 수량화가 가능 ex) 갯수, 나이, 키 등) - 등간, 비율

*** 통계학에서의 데이터 종류**



데이터에 따른 분석도구

데이터 분석(추론 및 검정) 시 종속변수(반응변수. 결과변수. 어떠한 영향을 받는)와 독립변수(설명변수, 원인변수. 종속변수에 의해 영향을 주는)가 있다. 결국 독립변수와 종속변수는 원인과 결과의 관계를 갖는다.

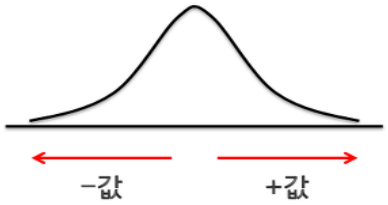
독립변수 (영향을 주는)	종속변수 (영향을 받는)	분석 방법
범주형	범주형	카이제곱 검정
범주형	연속형	T검정(범주형값 2개), 분산분석(ANOVA - 범주형값 3개)
연속형	범주형	로지스틱 회귀분석
연속형	연속형	회귀분석, 구조 방정식

카이제곱분포

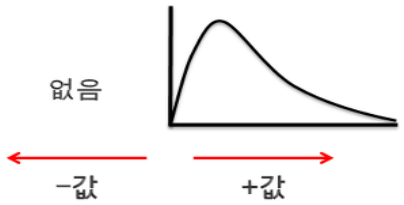
보통 무엇인가를 조사하고 분석할 때, 데이터들의 중심위치를 파악하는 것이 중요한데, 이 중심위치를 표현하는 대표적인 척도가 평균이다. 그리고 평균에서 데이터들이 흩어져 있는 정도, 즉 치우침을 표현하는 대표적인 척도가 분산이다.(표준편차도 있다) 그런데 이 분산이 퍼져있는 모습을 분포로 만든 것이 바로 카이제곱분포다. 분산의 제곱된 값을 다루기 때문에 χ2분포라고 불린다.

카이제곱분포는 데이터나 집단의 분산을 추정하고 검정할 때 많이 사용하는데, 특징 중 하나는, 제곱된 값 분산을 다루기 때문에, -값은 존재하지 않고 +값만 존재한다는 점이다. 그래서 정규분포 그래프와 비교해볼 때, 정규분포는 –값도 다루기 때문에 좌우가 모두 발달하여 좌우대칭인 모양을 갖지만 카이제곱분포는 +값만 다루기 때문에 한쪽만 발달하여 오른쪽 꼬리가 긴 비대칭 모양을 하고 있다.

<<정규분포>>



<<카이제곱분포>>



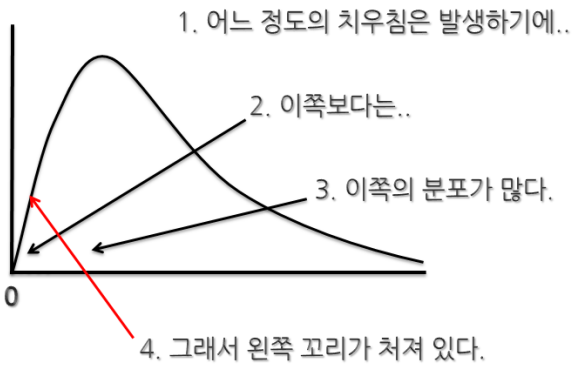
* 카이제곱분포 그래프의 특징
  - 확률변수는 연속확률 변수로서 항상 양(+)의 값을 갖는다.
  - 오른쪽 꼬리를 갖는 비대칭 분포다.
  - 자유도에 따라 모양이 다르다. 자유도(df)가 커질수록 좌우대칭인 정규분포에 가까워진다.

카이제곱분포

카이제곱분포 그래프를 보면 0에 가까울수록 분포가 많고, 0에서 멀어질수록 분포가 감소하는 것을 알 수 있다. 그 이유는 데이터나 집단의 치우침은 어느 정도 크기인 경우가 많지, 치우침이 말도 안 되게 큰 경우는 별로 없기 때문이다. 예를 들어 한국성인 남자의 평균 키가 173cm라는 것은, 174.5cm, 169.0cm, 172.3cm 처럼 평균을 기준으로 치우침이 별로 크지 않은 사람이 많고, 상대적으로 198.4cm, 135.1cm처럼 치우침이 아주 큰 사람은 적다는 말이다. 그러므로 카이제곱분포는 0에 가까울수록(치우침이 작을 경우) 분포가 많고, 0에서 멀어질수록(치우침이 클 경우) 분포가 감소하는 형태를 띠고 있다.

카이제곱 검정의 3가지 목적

1) 독립성 : 두 범주형 변수 간에 관련성이 있는지 여부를 알고자 할 때
2) 적합도 : 두 데이터가 특정한 분포에서 추출된 것인가 알고자 할 때
3) 동질성 : 두 개 이상의 다항분포가 동일한지 여부를 알고자 할 때



카이제곱분포도 t분포와 마찬가지로 연속확률분포이면서 표본분포로, 직접 확률을 구할 때 사용하는 분포가 아니라, 신뢰구간과 가설검정 그밖에 적합도 검정, 동질성 검정, 독립성 검정 등에 사용한다. 그리고 신뢰구간 추정이나 여러 검정을 할 때 χ^2값을 사용하며, 이 χ^2값은 그래프의 x축 좌표에 해당한다.

카이제곱분포

연구문제 : 성별에 따라 선호하는 커피브랜드의 차이가 있는가?

범주형 {성별}에 따라 범주형 {선호하는 커피브랜드}의 차이가 있는가?

범주형 자료에 따른 범주형 자료의 차이를 알아볼 때는? **카이제곱검정 활용**

성별/커피브랜드 모두 범주형 자료이고, 서로 어떤 영향을 미치는가 보기 위해, 카이제곱검정을 활용

▶ **카이제곱검정, 설문지 작성하기** 그렇다면 구체적으로 설문지 작성으로 넘어가 봅시다.

카이제곱을 사용하는 연구문제에서는 설문지를 어떻게 작성해야 할까?

설문지 구성 예시

1, 귀하의 성별은 무엇입니까?

① 남자 ② 여자

2. 선호하는 커피브랜드는 어디입니까?

① A사 ② B사 ③ C사

영향을 주고 받는 변수들 모두 범주형 자료로 구성되어 있는

카이제곱검정을 활용하기 위해서는,

범주형 자료를 얻을 수 있는 설문지 설계를 해야 한다는 것

▶카이제곱검정, 결과 분석하기

위의 설문 구성 예시를 통해 다음의 결과를 얻었다고 생각해 봅시다.

성별에 따른 커피브랜드별 선호도 조사 교차표

	A사	B사	C사
남자	30 %	30 %	40 %
여자	30 %	35 %	35 %

카이제곱분석은 **교차분석**이라고도 말한다. 그림처럼 **교차표를 통해, 결과값을 보기 때문.**

여기까지 보면, 카이제곱검정, 교차분석이 언제, 어떻게 활용되는지 감이 잡힌다.
하지만 여기서 그치는 것이 아니라, 결과값을 해석할 수 있어야 통계분석을 끝냈다고 할 수 있다.
여기서 활용되는 것이 바로 P값, 유의수준이다.

성별에 따른 커피브랜드별 선호도 조사 예시로 좀 더 알아보겠다.

성별에 따른 커피브랜드별 선호도 조사 결과 예시1

	A사	B사	C사
남자	30 %	30 %	40 %
여자	30 %	35 %	35 %

- 남자와 여자의 커피브랜드별 선호 비율이 거의 비슷
➔ **통계적으로 차이가 있다고 하기 애매모호하다.**

성별에 따른 커피브랜드별 선호도 조사 결과 예시2

	A사	B사	C사
남자	30 %	30 %	40 %
여자	60 %	20 %	20 %

- 명확한 차이가 있을 때에는 카이제곱 값 ↑, $p<0.05$
➔ **성별에 따라 선호하는 커피브랜드는 유의미한 차이가 있다.**

교차표를 보면 남자의 결과는 모두 같으니 **여자의 결과를 주목**해 봅시다.
<결과예시1> 보다 <결과예시2> 에서 여자가 선호하는 커피브랜드가 뚜렷하게 드러난다.

<결과예시1>의 경우에는 30%, 35%, 35%로 A사, B사, C사 모두 비슷하지만,
<결과예시2>의 경우에는 A사를 좋아하는 여성의 비율은 60%로, 다른 브랜드보다 선호하는 것으로 나온다.

이처럼 **차이가 존재하고, 유의미한 차이가 있다라고 분석할 수 있는 기준이 유의수준, P값**이다. (P값의 통계적 수치의 기준은 0.05.)

카이제곱검정을 실시하고,

P값이 0.05보다 크게 나온다고 하면, 성별에 따라서 선호하는 커피의 **차이가 없다.**
P값이 0.05보다 작게 나온다고 하면, 성별에 따라서 선호하는 커피의 **차이가 있다.**

이렇게 두 가지 결론으로 도출

통계분석 결과에서 유의수준, P값이라는 용어를 활용하여 통계적으로 표현하고 해석할 줄 알아야 논문 작성의 무리가 없다.

카이제곱분포

▶ 연구문제 예시

그렇다면 또 어떤 연구문제에 카이검증을 적용할 수 있을까?

연구문제 예시

1. **20대와 60대**는 **여당과 야당을 지지하는 사람의 비율**에서 차이가 있을까?
2. **성별**에 따라서 **맥주를 좋아하는 사람과 소주를 좋아하는 사람의 비율**은 다를까?
3. **출신 지역**에 따라서 **특정 야구팀을 선호하는 사람의 비율**이 다를까?
4. **전공**에 따라서 **액션영화와 멜로영화를 좋아하는 사람의 비율**이 다를까?

- 독립변수 : **집단**
- 종속변수 : **어떤 특성의 비율**

위의 연구문제들은 모두 **집단**에 따라서, **어떤 특성의 비율 차이가 있는지**를 알아보는데 관심이 있다.

- 독립변수 : 집단
- 종속변수 : 어떤 특성의 비율

이처럼 **어떤 특성의 비율이 집단에 따라서 다른지에 대한 문제를 검증하고자 할때, 카이검증을 적용**할 수 있다

교차분석(카이제곱(chi2) 가설검정)

교차분석은 명목이나 서열척도와 같은 범주형 자료를 대상으로 교차빈도에 대한 기술통계량을 제공해 줄 뿐만 아니라, 교차빈도에 대한 통계적 유의성을 검증해 주는 추론통계분석 기법이다. 교차 분할표를 통해서 두 개 이상의 범주형(명목, 서열척도) 변인의 관계를 분석하는 방법이다. 예를 들어 성별로 대학진학 여부에 대한 차이가 있는지에 대한 분석이 교차분석의 일종이다.

- 교차분석은 검정통계량으로 카이제곱을 주로 사용함(교차분석을 카이제곱 검정이라고 함)
- 카이제곱은 변수 간의 백분율을 나타내는 교차표를 작성하고, 두 변수 간의 독립성과 관련성을 분석한다.
- 카이제곱 검증 유형 분류 :
  일원카이제곱 검정(변인 단수 - 적합성),
  이원카이제곱 검정(변인 복수 - 독립성, 동질성)

일원 카이제곱 검정 실습

카이 제곱 검정은 goodness of fit(적합성) 검정이라고도 부른다.
SciPy stats 서브패키지의 chisquare 명령을 사용한다.

* 적합도 검정 실습 : 주사위를 던져서(60회) 관측도수 /기대도수가 아래와 같은 경우 적합한 주사위가 맞는가?'
* 적합성 가설 검정 예
  - 귀무가설 : 기대치와 관찰치는 차이가 없다. 예)주사위는 게임에 적합하다.
  - 대립가설 : 기대치와 관찰치는 차이가 있다. 예)주사위는 게임에 적합하지 않다.

주사위눈금	1	2	3	4	5	6
관측도수	4	6	17	16	8	9
기대도수	10	10	10	10	10	10

참고 – 가설 설정 방법
* 귀무가설 : 같다, 다르지 않다, 차이가 없다, 효과가 없다...
* 대립가설 : 같지 않다, 다르다, 차이가 있다, 효과가 있다...

이원카이제곱 실습

이원카이제곱

: 두 개 이상의 집단 또는 범주의 변인을 대상으로 동질성 or 독립성 검정.

: 유의확률에 의해서 집단 간에 '차이가 있는가? 없는가?' 로 가설을 검정한다.

동질성 검정 - 두 집단의 분포가 동일한가? 다른 분포인가? 를 검증하는 방법으로 두 집단 이상에서 각 범주(집단) 간의 비율이 서로 동일한가를 검정하게 된다. 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법이다.

실습) 교육방법에 따른 만족도 분석 - 동질성 검정

동질성 분석

- 귀무가설 : 교육방법에 따른 만족도에 차이가 없다.
- 대립가설 : 교육방법에 따른 만족도에 차이가 있다.

집단 별 비율검정과 평균차이 검정

*** T검정(범주형 값 2개)**

: 비율 검정 - 빈도수에 대한 비율에 의미가 있다.

: 평균차이 검정 - 표본평균에 의미가 있다.

단일 표본 t-검정 (One-sample t-test)

- 단일 표본 t-검정은 정규 분포의 표본에 대해 기댓값을 조사하는 검정방법이다.
- SciPy의 stats 서브 패키지의 ttest_1samp 명령을 사용한다.
- 모수(평균)를 알고 있는 경우 sample의 평균과 모수(평균)와 여부를 검정

귀무가설 : 모수와 같다.

대립가설 : 모수와 다르다.

T검정, 차이검증이란? T검정은 **두 집단의 평균 점수를 비교**하고자 할 때, 실시하는 분석방법이다.

범주형 자료에 따른 연속형 자료의 차이를 볼 때, T검정과 ANOVA분석 (분산분석)을 사용할 수 있다.

영향을 주는 변수	영향을 받는 변수	통계분석방법
범주형 자료	범주형 자료	카이제곱 검정
	연속형 자료	T검정 분산분석

이 두 분석의 차이는 범주형 자료의 집단이 몇 개인가 이다.

범주형 자료의 집단이 두 개일 경우, **T검정**
범주형 자료의 집단이 세 개이상일 경우, **ANOVA분석 (분산분석)**

연구문제 : 성별에 따라 A사 커피브랜드의 만족도는 차이가 있는가? 통계분석 방법 : T검정

범주형 {성별}에 따라 연속형 {A사 커피 브랜드의 만족도}가 차이가 있는가?

[남자 / 여자] 두 집단

조건 1
범주형 자료에 따라 연속형 자료에 미치는 영향.

조건 2
남자 / 여자라는 범주형 자료의 두 집단.

} T-검정, T-Test

[조건 1] 성별이라는 범주형 자료에 따라, 만족도라는 연속형 자료를 확인하는 검증 과정이다.
[조건 2] 기준이 되는 성별이라는 범주형 자료가 남/녀로 두 집단이다.

▶T검정, 설문지 작성하기

그렇다면, T검정을 사용하는 연구문제에서는 설문지를 어떻게 구성해야 할까?

설문지 구성 예시

1, 귀하의 성별은 무엇입니까?
① 남자 ② 여자

2. A사 커피브랜드에 대해 전반적으로 얼마나 만족하십니까?
① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

이처럼, **범주형 자료와 연속형 자료를 모두 얻을 수 있도록 설문지를 구성**해야 한다.

[범주형 자료] 비교하고자 하는 두 집단을 알아보기 위한 질문
[연속형 자료] 실질적으로 확인하고자 하는 변수를 알아보는 질문

▶ T검정, 결과 분석하기

만족도는 연속형이기 때문에, 만족도의 평균을 구할 수 있겠죠? 데이터 수집에 따라, 아래의 결과를 도출했다.

A사 커피브랜드 성별에 따른 만족도 평균

	평균
남자	3.57
여자	3.14

하지만 여기서 끝이 아니다.
단순히 평균을 비교한다고 하면,
남자의 만족도 평균은 3.57,
여자의 만족도 평균은 3.14로
수치상에서 벌써 차이가 난다.
그렇지만 **이것이 정말 유의미한 차이를 보이는지 확인하기 위해서는 통계적인 검증이 필요합니다.**

T검정을 실시하고, **P값이 0.05보다 크게 나온다**고 하면, 성별에 따라 A사 커피브랜드의 만족도는 **차이가 없다.**
P값이 0.05보다 작게 나온다고 하면, 성별에 따라 A사 커피브랜드의 만족도는 **차이가 있다.**

예를 들면 P값이 0.03이 나왔다고 가정해 보자.
P값의 유의수준 기준이 0.05이므로, 이 경우에는 0.05보다 작기 때문에, 통계적으로 유의미한 차이가 있다고 본다.
즉, 남자의 만족도 평균이 여자의 만족도 평균보다 높다고 결론을 낼 수 있다.

A사 커피브랜드 성별에 따른 만족도 평균

	평균
남자	3.57
여자	3.14

평균의 차이가 눈에 봐도 뚜렷. ➔ t값 ↑, $p<0.05$

결론
성별에 따른 A사 커피브랜드 만족도는 유의미한 차이가 있으며,
남자 (3.57)가 여자 3.14)보다 만족도가 높다.

T 검정

▶연구문제 예시

그렇다면, 또 어떤 연구문제에 T검정을 적용할 수 있을까?

연구문제 예시

1. **서울지역 고등학생과 부산지역 고등학생** 중에서 누구의 **수능점수**가 더 높을까?
2. **남학생과 여학생**은 **지능검사 점수**에서 차이가 있을까?
3. **무용학과 학생과 유아교육학과 학생** 중 어떤 학과 학생들의 **몸무게**가 더 높을까?
4. **중학생과 고등학생**의 **한 달에 받는 용돈**에는 차이가 있을까?

- 독립변수 : 집단
- 종속변수 : 어떤 특성의 평균값

위의 연구문제들은 모두 **두 집단 간에 어떤 특성의 평균 값에서 차이가 있는지**를 알아보는데 관심이 있다.

- 독립변수 : 집단
- 종속변수 : 어떤 특성의 평균값

이처럼 **두 집단 간에 어떤 특성의 평균값에서 차이가 있는지**에 대한 문제를 검증하고자 한다면 T검정, 차이검증을 적용할 수 있다.

분산분석 중 두 집단 평균차이 검정 독립 표본 t-검정(Independent-two-sample t-test)

- 두 개의 독립적인 정규 분포에서 나온 두 개의 데이터 셋을 사용하여 두 정규 분포의 기댓값이 동일한지를 검사한다. SciPy stats 서브패키지의 ttest_ind 명령을 사용한다.
- 독립 표본 t-검정은 두 정규 분포의 분산 값이 같은 경우와 같지 않은 경우에 사용하는 검정 통계량이 다르기 때문에 equal_var 인수를 사용하여 이를 지정해 주어야 한다.

분산분석 중 세 집단 평균차이 검정 (F검정 : ANOVA : Analysis of Variance)

- 선형회귀분석의 결과가 어느 정도의 성능을 가지는지는 단순히 잔차제곱합(RSS : Residuala Sum of Square)으로 평가할 수는 없다. 변수의 스케일이 달라지면 회귀분석과 상관없이 잔차제곱합도 같이 커지기 때문이다. ANOVA는 종속변수의 분산과 독립변수의 분산 간의 관계를 사용하여 선형회귀분석의 성능을 평가하고자 하는 방법이다. 분산분석은 서로 다른 두 개의 선형회귀분석의 성능 비교에 응용할 수 있으며, 독립변수가 카테고리 변수인 경우 각 카테고리 값에 따른 영향을 정량적으로 분석하는데도 사용된다.
- 독립변수가 복수인 경우에는 각 독립변수에 대한 F검정 통계량을 구할 수 있다.

< F검정 통계량으로 가설검정 >

분산분석에서 신뢰수준 95%에서는 -1.96 ~ 1.96의 범위가 귀무가설의 채택역이다.

따라서 F검정 통계량이 채택역에 해당하지 않으면 귀무가설을 기각할 수 있다.

*** 분산분석에서 F검정 통계량과 유의수준 a(알파) 관계표**

F값(절대치)	유의수준a(양측검정 시)
F값(절대치) >= 2.58	a = 0.01 (의.생명 분야)
F값(절대치) >= 1.96	a = 0.05 (사회과학 분야)
F값(절대치) >= 1.645	a = 0.1 (일반 분야)

분산분석(ANOVA분석), 변량분석이란?

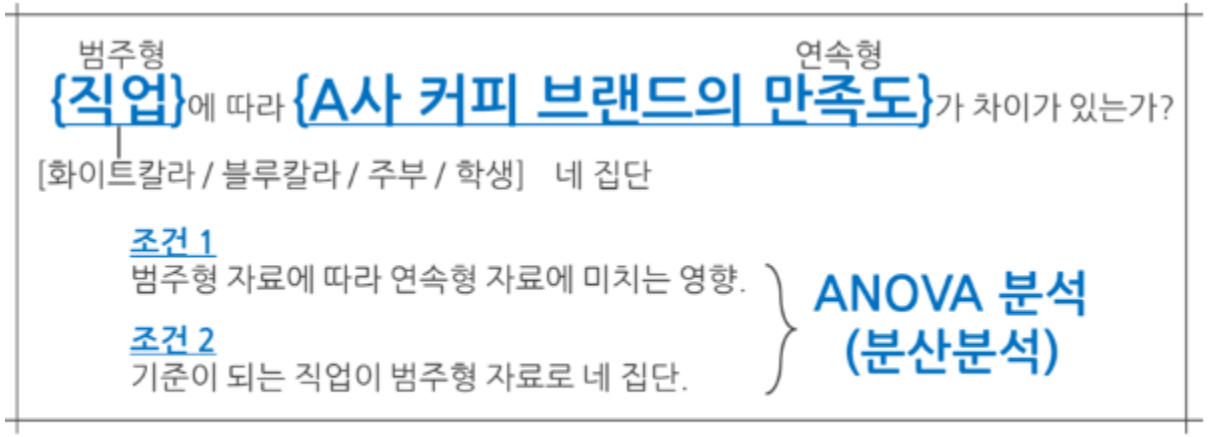
변량분석은 **둘 이상의 집단 간 평균 점수를 비교하고자 할 때 실시하는 분석방법**이다.
차이검증(T검정)으로는 두 집단 끼리만 비교할 수 있지만 변량분석을 이용하면 더 많은 집단끼리도 비교할 수 있다.

범주형 자료에 따른 연속형 자료의 차이를 볼 때, T검정과 ANOVA분석 (분산분석)을 사용할 수 있다.

이 두 분석의 차이는 범주형 자료의 집단이 몇 개인가 이다.

범주형 자료의 집단이 두 개일 경우, **T검정**
범주형 자료의 집단이 세 개 이상일 경우, **ANOVA분석 (분산분석)**

연구문제 : 직업(화이트칼라, 블루칼라, 주부, 학생)에 따라 A사 커피브랜드의 만족도는 차이가 있는가?
통계분석 방법 : ANOVA분석



[조건 1] 화이트칼라, 블루칼라, 주부, 학생이라는 범주형 자료에 따라, 만족도라는 연속형 자료이다.
[조건 2] 기준이 되는 직업이 화이트칼라, 블루칼라, 주부, 학생이라는 범주형 자료로 네 집단이다.

▶ANOVA분석, 설문지 작성하기

그렇다면, ANOVA분석을 사용하는 연구문제에서는 설문지를 어떻게 만들어야 할까?

설문지 구성 예시

1, 귀하의 직업은 무엇입니까?

① 화이트칼라 ② 블루칼라 ③ 주부 ④ 학생

2. A사 커피브랜드에 대해 전반적으로 얼마나 만족하십니까?

① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

[범주형 자료] 비교하고자 하는 세 개 이상의 집단을 알아보는 질문
[연속형 자료] 실질적으로 확인하고자 하는 변수를 알아보는 질문

이처럼, **범주형 자료와 연속형 자료를 모두 얻을 수 있는 설문지로 구성**해야 한다.
T검정과 유사하게 범주형 자료와 연속형 자료를 알아보지만, **ANOVA분석은 세 개 이상의 집단을 알아보는 질문으로 구성** 된다.

▶ANOVA분석, 결과 분석하기

연속형 자료인 만족도의 평균값을 도출한다. 데이터 수집을 바탕으로 얻은 결과값이 다음과 같다고 해 보자.

A사 커피브랜드 직업에 따른 만족도 평균

	평균
화이트칼라	3.14
블루칼라	2.97
주부	2.56
학생	2.47

- 명확한 차이가 있을 때에는 F 값 ↑, $p<0.05$
➔ **직업에 따라서 만족도는 유의미한 차이가 있다.**

통계적 검증을 하고 결과를 해석하는데 있어, **여기서부터 T검정과 조금** 다르다.
그 이유는 바로, **T검정**의 경우에는 **비교집단이 2개**이고 **분산분석**의 경우에는 **비교집단이 3개 이상**이기 때문이다.
비교집단이 2개인 경우에는 단순히 유의미한 차이가 있는지를 확인하면 된다.

하지만 비교집단이 3개 이상인 경우에는 **<1> 집단 간 유의미한 차이가 있는지를 확인**
<2> 각 집단끼리 어떤 차이가 있는지를 확인

집단이 2개이면 1번과 2번을 구분할 필요가 없게 되고, 집단이 3개 이상일 경우에는 1번과 2번 작업이 구분되는 것이다.

<1> 집단과 유의미한 차이가 있는지를 확인

T검정과 같이 유의수준 P값으로 확인하면 된다.

ANOVA분석을 실시하고, **P값이 0.05보다 크게 나온다**고 하면, 직업에 따라 A사 커피브랜드의 만족도는 **차이가 없다.**
P값이 0.05보다 작게 나온다고 하면, 직업에 따라 A사 커피브랜드의 만족도는 **차이가 있다.**

<2> 각 집단끼리 어떤 차이가 있는지를 확인

이 작업은 **사후검정**이라고 불리우는데요. 일반적으로 Scheffe, Duncan 등의 방법을 활용한다.
사후검정 결과값이 다음과 같이 나왔다고 가정해보자.

사후검정 결과 예시

각각의 대소 비교를 위해, 사후검정인 Scheffe, Duncan 등의 방법을 활용

화이트 칼라 > 블루칼라 ($p<0.05$)　　블루 칼라 > 주부 ($p<0.05$)
블루칼라 > 학생 ($p<0.05$)　　주부 = 학생 ($p>0.05$)

화이트 칼라 > 블루 칼라 > 주부 = 학생

➔ **직장에 다니는 사람이 안 다니는 사람보다 만족도가 높으며,**
직장에 다니는 사람 중에서는 화이트칼라가 블루칼라보다 만족도가 높다.

화이트 직종과 블루 직종을 검정하면 P값이 0.05보다 작음을 의미
블루 직종과 주부 직종을 검정하면 P값이 0.05보다 작음을 의미
주부 직종과 학생 직종을 검정하면 P값이 0.05보다 큰 것을 의미

이것을 통해 결론을 내리자면,

화이트 직종과 블루 직종이 주부 직종과 학생 직종보다 만족도가 크다는 것은
직업이 있는 사람이 직업이 없는 사람보다 A사 커피브랜드의 만족도가 높다는 것을 의미한다.
또한 화**이트 직종이 블루 직종보다 A사 커피브랜드의 만족도가 높다**는 것을 알 수 있다.

ANOVA

▶연구문제 예시

분산분석에서는 어떠한 연구문제에 적용할 수 있을까?

연구문제 예시

1. **20대, 30대, 40대** 간에 **패스트푸드에 대한 선호도**의 차이가 있을까?
2. **사무직, 기술직, 서비스직** 간에 **연봉**의 차이가 있을까?
3. **초등학생, 중학생, 고등학생, 대학생** 간에 **지능검사 점수**의 차이가 있을까?

- 독립변수 : **둘 이상의 집단**
- 종속변수 : **어떤 특성의 평균값**

위의 연구문제들은 모두 둘 이상의 집단 간에 어떤 특성의 평균값에서 차이가 있는지를 알아보는데 관심이 있다.

- **독립변수 : 둘 이상의 집단**
- **종속변수 : 어떤 특성의 평균값**

이처럼 둘 이상의 집단 간에 어떤 특성의 평균값에서 차이가 있는지에 대한 문제를 검증하고자 한다면, 분산분석, 변량분석을 적용할 수 있다.

상관관계 분석 (Correlation Analysis)

- 회귀분석에서 변수들 간의 인과관계를 분석하기 전에 각 변수들 간에 관련성을 분석하는 선행자료로 이용된다.
- 상관계수의 수치로 관계의 정도를 파악할 수 있다.
- 변수 간 관련성 분석, 관계의 친밀함을 수치로 표현할 수 있다.

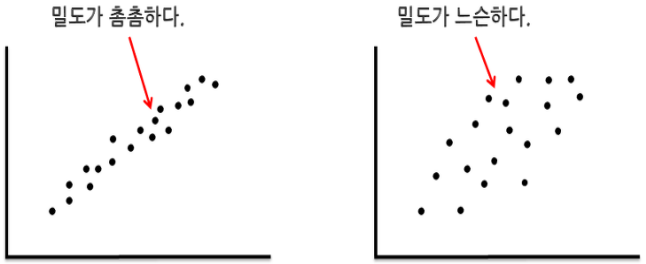
  예) 광고비와 매출액 사이의 관련성 분석, 광고량과 브랜드 인지도의 관련성 분석

상관계수란 ?

상관분석은 두 변수가 서로 어떠한 관계인지를 파악하는 분석이다.

또 점들이 흩어져 있는 모습을 보고 두 변수의 관계를 파악하는데,

기울기에 따라 양의 상관관계와 음의 상관관계로 나눌 수가 있다.



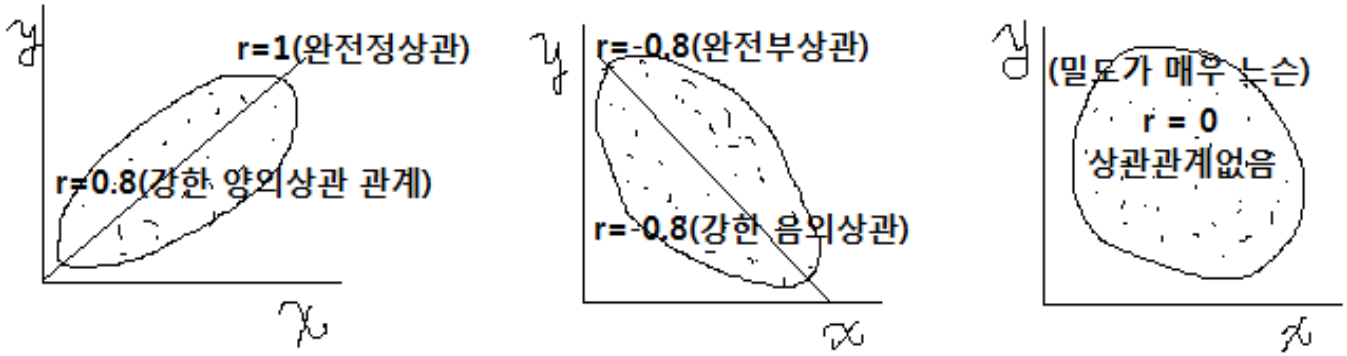
그런데 의문점은, 과연 점들이 모여 있는 "밀도"는 어떻게 표현하는가?이다.

그림을 보면, 둘 다 모두 양(+)의 상관관계이지만, 같다고 하기에는 "밀도"의 차이가 난다. 그래서 먼저 각각 어느 정도의 밀도인지를 알아야 하고, 그로 인해 서로 얼마나 다른지를 파악할 수가 있어야 하는데, 그림으로 이것을 파악하기에는 한계가 있다. 그래서 통계에서는 "숫자"를 사용해서 밀도를 표현하는데, 이 밀도를 표현한 숫자를 보통 상관계수라고 부른다. (기호는 r을 사용한다.)

상관관계 분석

*** 상관계수 r과 상관관계 정도**

- 상관계수는 밀도를 숫자로 표현한다. 밀도를 가지고 상관관계를 정확하게 표현하기 힘들다. 그래서 숫자화 해야 한다. 이 것을 정도에 따라 구분한 것 중 하나가 피어슨 상관계수다.
- 상관계수r은 -1 ~ 1 사이의 값을 갖는다. (1 : 완전상관(밀도 촘촘), 0 : 상관관계 없다)



x가 커지면 y도 커진다(정비례) x가 커질수록 y는 작아진다(반비례)

상관분석은 기본적으로 변수가 2개 이상이므로 평균에서 치우침이 두 변수에 의해 발생하게 되기 때문에 분산 외에 공분산 값을 알아야 한다.

공분산은 두 개 이상의 확률변수에 대한 관계를 보여 주는 값이다. 즉, 확률변수 x와 y에 대해 x가 변할 때 y가 변하는 정도를 나타내는 값을 말한다. 관련이 없으면 0, 관련이 많을 수록 1에 가까워진다.

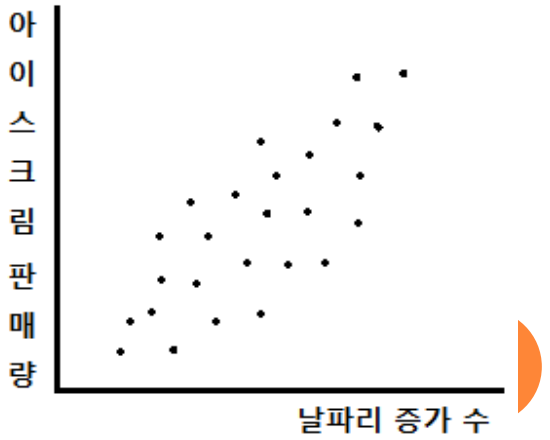
상관관계 분석

상관분석은 변수 간에 어떠한 관계가 있는지 상관관계를 파악할 수는 있지만, 서로가 직접적인 영향을 주고 받는지에 대한 인과관계는 정확하게 파악할 수 없다. 이 것은 회귀분석을 이용하면 가능해 진다.

- 상관 분석 : 변수간의 관련성 분석
- 회귀 분석 : 변수간의 인과관계 분석으로 사용범위가 넓은 분석방법

예) 날씨가 더워질수록 아이스크림이 잘 팔린다.
날씨가 더워질수록 날파리가 늘어난다.

"아이스크림 판매가 늘어나니 날파리도 증가한다"
라고 판단할 수는 없다.
또 다른 변수인 "날씨"가 더워진 관계로 발생한
상관관계일 뿐이지 서로(아이스크림 / 날파리)가
직접적인 영향(인과관계)을 준 것은 아니다.
그러므로 임의의 변수 간에 관계를 파악하고 설명할
때는 신중을 기하는 것이 중요하다라고 생각된다.



상관관계 분석

▶연구문제 예시

그렇다면, 상관분석은 어떤 연구문제에서 적용할 수 있을까?

연구문제 예시 - 변수 : **두 가지 연속 변수**

1. **부모의 수입**과 **성적**의 관련성
2. **키**와 **몸무게**의 관련성
3. **나이**와 **스마트폰 사용시간**의 관련성

위의 연구문제들은 **두 가지 연속 변수 간에 관련성이 있는지**를 알아보는데 관심이 있다.
두 가지 연속변수 간에 관련성이 있는지에 대한 문제를 검증하고자 한다면 상관분석을 적용할 수 있다.

** Machine Learning(기계학습) **

- 사람과 기계와의 소통이 가능한 이유는 기존에 수많은 알고리즘을 통해서 기계에게 학습을 시킨 후 새로운 데이터가 입력되면 기계 스스로가 해석할 수 있는 기계학습이 가능하기 때문이다.

- 알고리즘을 통해 기계에게 학습을 시킨 후에 새로운 자료가 입력된 경우 해당 자료의 결과를 분석하여 예측결과를 제공해 준다. 예를 들어 검색어 자동 완성, 악성코드 감지, 자료 인식 등의 다양한 예측을 필요로 하는 분야에서 사용될 수 있다.

- Machine Learning?
  - 일상에서 접하는 Machine Learning
    - 상품의 추천/ 스팸 메일 분류/ 쿠폰발급/ 대출심사 등
  - 기업에서 적용하는 Machine Learning
    - Business decision / Productivity 증대/ Disease detection / Anomaly detection / Forecasting weather

** Machine Learning(기계학습) **

지도학습(Supervised Learning)은 훈련 데이터로 부터 예측/추정, 분류 함수를 만들어 내는 기계학습 방법이다. 이는 사전에 입출력에 대한 사전정보를 어느 정도 갖고 있는 상태에서 입력된 자료에 대한 모델을 만들고, 이를 통해 추정 및 예측을 할 수 있다. 독립변수들과 종속변수를 가지며, 사람의 지도 하에 독립변수와 종속변수 간에 관계를 일반화 하는 학습을 하게 된다. Y값을 가지고 있으니 학습이 끝나고 나면 결과가 나온다. 과거의 데이터를 가지고 학습을 시키고 나면, 학습 모델에 새로운 데이터에 적용해서 예측/추정, 분류 등의 작업을 수행한다.

: Regression, Support Vector Machine ...

비지도학습(Unsupervised Learning)은 관찰한 데이터로 부터 숨겨진 패턴/규칙을 탐색, 찾아내는 기계학습 방법이다. 이는 최종적인 정보가 없는 상태에서 기계 스스로가 정해진 패턴에 의해 분석결과를 만들어 내게 된다. 비지도학습에서는 종속변수(y) 값이 없고, 사람의 개입도 없다. 그냥 Input 변수 (x1, x2, ...) 만 주고 컴퓨터에게 수많은 데이터 속에 숨겨져 있는 패턴을 찾도록 한다. 연관/순차규칙 분석, 군집화, 차원 축소, 네트워크 분석 등이 비지도 학습에 속한다. 그런데 비지도 학습은 Y값이 없어, 학습 종료 후에 도출결과를 객관적으로 평가하기가 다소 애매하며 분석가의 주관이 들어갈 수 있다.

: Classifier, Clustering ...

** Machine Learning(기계학습) **

*** 지도학습 / 비지도학습 비교 ***

분류	지도학습	비지도학습
주관	사람이 개입	컴퓨터 자체
기법	확률과 통계 기반 추론 통계	패턴분석 기반, 데이터 마이닝
유형	회귀분석, 신경망 (y변수 O)	군집분석, 연관분석 (y변수 X)

* 지도학습은 독립변수와 종속변수가 있지만, 비지도학습은 종속변수가 없다.

기계학습 방법	가능 분야	분석방법의 종류 및 알고리즘
지도학습 (Supervised Learning)	예측, 추정 (Prediction, Estimation)	✧ Linear Regression ✧ Regression Tree, Model Tree ✧ SVM(Support Vector Machine) ✧ Neural Network, Deep Learning ✧ ARIMA, Exponential Smoothing
	분류 (Classification)	✧ Decision Tree ✧ Logistic Regression, Discriminant Analysis ✧ k-NN(k-Nearest Neighbor), CBR(Case-Based Reasoning) ✧ Naïve Bayes Classification ✧ SVM, Neural Network ✧ Ensemble (Bagging, Boosting, Random Forest)
비지도학습 (Unsupervised Learning)	패턴/구조 발견 (Pattern/Rule)	✧ Association Rule Analysis, Sequence Analysis ✧ Network Analysis, Link Analysis, Graph theory ✧ Structural Equation Modeling, Path Analysis
	그룹화 (Grouping)	✧ k-Means Clustering, Hierarchical Clustering, Density-based Clustering, Fuzzy Clustering ✧ SOM(Self-Organizing Map)
	차원 축소 (Dimension Reduction)	✧ PCA(Principal Component Analysis), Factor Analysis, SVD(Singular Value Decomposition)
	영상,이미지,문자 (Video, Image, Text, Signal processing)	✧ Wavelet/Fast Fourier Transformation, DTW(Dynamic Time Warping), SAX(Symbolic Aggregate Approximation), Line/Circular Hough Transformation ✧ Text mining, Sentiment Analysis

회귀분석(Regression Analysis)

변수 간의 인과관계를 밝히기란 매우 어려운 문제다. 수학적 방법 이외에 다양한 외적 조건도 따져봐야 한다. 회귀분석은 이런 과정 중에 하나에 불과하다.

- 특정변수(독립변수)가 다른 변수(종속변수)에 어떤 영향을 미치는가를 분석.
  즉, 인과관계를 분석.
- 독립, 종속변수는 등간 또는 비율척도(연속형 데이터)로 구성되어야 한다.
  한 변수의 변화에 따른 다른 변수의 값을 파악.
- 독립변수가 종속변수에 영향을 미치는 변수를 규명하고, 이 들 변수들에 의해서 회귀방정식을 도출하여 회귀선을 추정한다. $Y = a + Bx$
- 회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 예측에 이용될 수 있다.
- 회귀분석의 기본 가정 충족 조건
  : 선형성, 잔차 정규성, 잔차 등분산성, 잔차 독립성, 다중 공선성 등

회귀분석(Regression Analysis)

~ **상관관계 분석 :** 변수 간의 **관련성** 분석

~ **회귀분석 :** 변수 간의 **인과관계** 분석, 사용범위가 넓은 분석방법이다

1) **단순회귀분석** : 독립변수와 종속변수가 각각 1개인 경우에 독립변수가 종속변수에 어떠한 영향을 미치는지 인과관계를 분석.

* 연구모델 : 제품적절성(독립변수) -> 제품 만족도(종속변수)

- 귀무가설 : 제품의 품질(당도)과 가격수준을 결정하는 제품 적절성(독립변수)은 제품 만족도(종속변수)에 영향을 미치지 않는다. (영향이 없다.)

- 연구가설 : 제품의 품질(당도)과 가격수준을 결정하는 제품 적절성은 제품 만족도에 정(正)의 영향을 미친다.

2) **다중회귀분석**: 여러 개의 독립변수로 1개의 종속변수에 미치는 영향 분석

-연구모델 : 제품 적절성, 제품 친밀도 -> 제품 만족도

-귀무가설 : 음료수 제품의 적절성과 친밀도는 제품 만족도에 정(正)의 영향을 미치지 않는다.

-연구가설 : 음료수 제품의 적절성과 친밀도는 제품 만족도에 정(正)의 영향을 미친다.

회귀분석이란?

회귀분석은 **독립변인이 종속변인에 영향을 미치는지 알아보고자 할 때 실시하는 분석방법**이다.

연속형 자료에 따른 연속형 자료의 영향력을 검증하고자 할 때, **회귀분석**을 사용한다.

연속형 변수끼리 미치는 영향력이라고 하면 조금 헷갈릴 수 있다. 간단한 예를 통해 알아보겠다.

영향을 주는 변수	영향을 받는 변수	통계분석방법
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱 회귀분석

연속형 변수로, 커피 맛, 가게 인테리어, 직원 친절도가 있다고 생각해보자.

위의 변수는 1점에서 5점 척도로 측정한다. 물론 7점 척도를 사용해도 된다.

커피 숍의 입장에서 위의 변수 중에 무엇이 만족도에 영향을 주는지 확인하고자 하는 것이다.

그리고 만족도를 높이려면, 무엇을 개선해야 하는지를 파악할 때, 회귀분석을 활용할 수 있다.

연구 문제 : A커피 숍의 커피의 맛, 가게 인테리어, 직원 친절도가 고객 만족도에 미치는 영향 통계분석 방법 : 회귀분석

A커피숍의 {커피의 맛, 가게 인테리어, 직원 친절도}(연속형)가 {고객 만족도}(연속형)에 미치는 영향

[독립변수] 연속형 자료 - 커피의 맛, 가게 인테리어, 직원 친절도

[종속변수] 연속형 자료 - 만족도

이처럼 두 연속형 자료가 미치는 영향에 대해 알아보고자 할 때 회귀분석을 사용한다.

▶ 회귀분석의 종류

회귀분석 중에서도,

영향을 주는 변수가 1개이면 단순회귀분석이고, **영향을 주는 변수가 2개 이상 이면 다중회귀분석**이다.

분석을 할 때에는 SPSS에서 독립변수만 같이 집어 넣으면 되기 때문에, SPSS의 분석방법에서는 크게 달라지지 않는다.

회귀분석의 종류

독립변수 (영향을 주는 변수) 1개 : 단순회귀분석 — **단순회귀분석**

독립변수 (영향을 주는 변수) 2개 이상 : 다중회귀분석 — **다중회귀분석**

▶회귀분석, 설문지 작성하기

예시를 통해, 알아보도록 하자. 회귀분석을 사용하는 연구문제에서는 설문지를 어떻게 작성해야 할까?

설문지 구성 예시

1. A사 커피브랜드의 다음 항목에 대한 만족도를 체크하십시오.

[커피의 맛]	① 매우 불만족	② 불만족	③ 보통	④ 만족	⑤ 매우 만족
[인테리어]	① 매우 불만족	② 불만족	③ 보통	④ 만족	⑤ 매우 만족
[직원 친절도]	① 매우 불만족	② 불만족	③ 보통	④ 만족	⑤ 매우 만족

2. A사 커피브랜드에 대해 전반적으로 얼마나 만족하십니까?

① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

회귀분석은 영향을 주고 받는 변수들 모두 연속형 자료로 구성되어 있기 때문에,
이처럼, **연속형 자료를 얻을 수 있는 설문지를 구성**해야 한다.

▶회귀분석, 결과 분석하기

회귀분석의 경우에는 **통계수치 값을 약간은 이해해야 한다.**

R제곱과 F값이라는 것이 있는데, **R제곱은 식의 설명력이라고 보면 된다.**

독립변수 3개가 만족도를 얼마나 설명하느냐를 판단하고, 대략 30#정도 나오면 높은 수치라 할 수 있다.

F값은 모형 적합도를 나타낸다.

P값이 0.05보다 작으면 이 모형이 적합하다고 할 수 있다.
P값이 0.05보다 크면 이 모형은 부적합하다고 할 수 있다.

R제곱과 F값

R제곱 : 독립변수가 종속변수의 몇 퍼센트인가를 설명하는 수치

F값 : 회귀식의 적합도 ($p<0.05$보다 작아야 회귀식이 유의미함)

그 다음은 회귀식에 대해 분석하는 것이다.
B는 표준화되지 않은 영향력을 판단할 때 사용하고, β는 이 영향력의 상대적인 차이를 비교할 때 사용한다.

B와 β

B :비표준화 계수
절대적인 영향력의 크기

β :표준화 회귀계수
상대적인 영향력의 크기,
종속변수에 가장 큰 영향을 미치는 변수가 무엇인가를 판단할 때 활용

회귀식이 아래 예시처럼 나왔다고 생각해 보자.

회귀식 예시

만족도(y)=0.5 + 0.8*커피맛 + 0.7*인테리어 + 0.6*직원

➔ **커피맛이 1만큼 증가하면, 만족도는 0.8정도 증가한다고 예측할 수 있음**

회귀식을 통해, 이 값들이 얼마나 만족도에 영향을 미치는지 알 수 있으며,
커피 맛이 1이 증가하면, 만족도는 0.8정도가 증가한다고 볼 수 있다.

표준화, β 값 같은 경우는 커피 맛, 인테리어, 직원 친절도의 표준편차가 다르기 때문에,
자료가 얼마나 넓게 퍼지느냐 좁게 퍼지느냐에 따라 이것이 영향력이 큰지 작은지 판단할 수 있다.

여기에서는 커피 맛의 숫자가 가장 큰 것으로 볼 때, 커피 맛이 만족도에 제일 큰 영향을 줬다고 해석하면 된다.

▶연구문제 예시

회귀분석은 어떤 연구문제에서 적용할 수 있을까?

연구문제 예시

1. **부모의 수입** 이 **성적**에 미치는 영향
2. **키**가 **몸무게**에 미치는 영향
3. **나이**가 **스마트폰 사용시간**에 미치는 영향

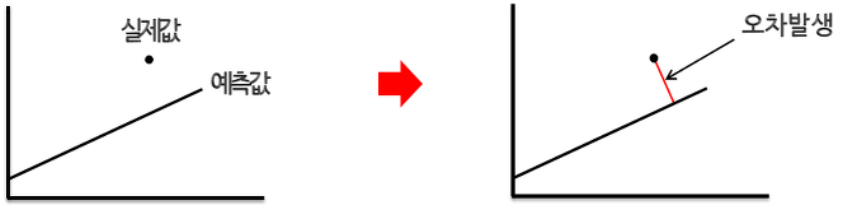
- 독립변수 : 연속변수
- 종속변수 : 연속변수

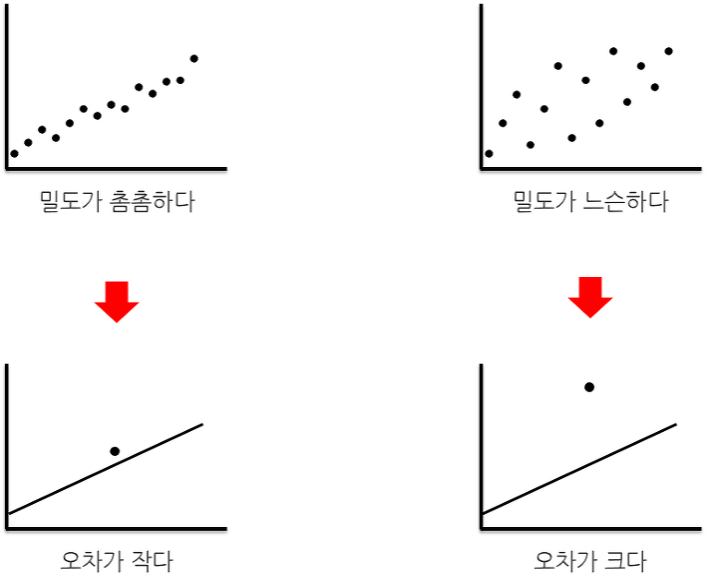
위의 연구 문제들은 **연속변수인 독립변인이 연속변수인 종속변인에 영향을 미치는지**를 알아 본다.

대부분의 연구는 독립변인이 종속변인에 영향을 미치는지를 검증하기 때문에,
학위논문에서 가장 많이 활용되는 것 중 하나가 바로 회귀분석 이다.

결정계수

결정계수는 상관분석의 상관계수와 유사하다. 일단 회귀분석은 회귀식을 활용해서 무엇인가를 예측하는 분석이다. 그래서 무엇인가를 예측할 때, 회귀분석을 사용하면 눈대중으로 막 잡은 수치보다는 훨씬 신뢰할 수가 있다. 하지만 회귀분석으로 예측을 해도, 정답인 실제 값은 안 나온다. 다만 틀릴 확률이 존재하는 예측 값이 나오면서, 항상 오차가 발생한다.





〈회귀식의 정확도가 높다〉 〈회귀식의 정확도가 낮다〉

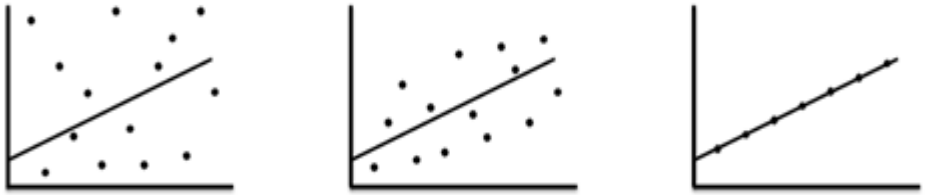
그런데 점들이 모여 있는 밀도에 따라서, 오차의 크기가 다르다. 예를 들어 점들이 모여 있는 밀도가 촘촘할 경우에는, 예측값과 실제값이 얼마 차이 나지 않는다.(오차가 작다) 하지만 점들이 모여 있는 밀도가 느슨할 경우에는, 예측값과 실제값이 많이 차이 난다.(오차가 크다) 그래서 똑같은 회귀분석이라도, 점들이 모여 있는 밀도에 따라 오차의 크기가 다르고, 그로 인해 회귀식의 정확도가 달라진다.

결정계수

이렇게 점들이 모여 있는 밀도에 따라 회귀식의 정확도가 결정되는데, 문제는 정확도가 구체적으로 어느 정도인지, 즉 "얼마나?" 정확한지를 판단할 수 있어야 한다. 그런데 얼마나 정확한지는, 그림으로 파악할 수가 없다. 그래서 통계에서는 "숫자"를 활용하는데, 회귀식이 얼마나 정확한지를 나타내는 숫자가 결정계수다.

기호는 R2을 사용한다.

그래서 결정계수(R2)를 사용하면 회귀식이 얼마나 정확한지를 나타낼 수 있는데, 보통 숫자 0부터 1까지만(0≤R2≤1) 사용한다. 그래서 결정계수가 0에 가까울수록 "회귀식의 정확도는 매우 낮다"고 할 수 있고, 결정계수가 1에 가까울수록 "회귀식의 정확도는 매우 높다"고 할 수 있다. 그래서 결정계수가 낮을수록 예측 값은 믿을 게 못되고 반대로 결정계수가 높을수록 예측 값은 믿을 만하다.



$R^2 = 0$ 〈믿을 게 못 된다〉

$R^2 = 0.5$ 〈어느 정도 믿을 만 하다〉

$R^2 = 1$ 〈믿을 만 하다〉

결정계수를 구하는 방법은 크게 2가지가 있는데, 하나는 상관계수를 제곱해서 구하는 방법이고, 나머지 하나는 분산분석의 데이터를 가지고 구하는 방법이다.
(회귀변동/총변동으로 구한다)

로지스틱 회귀분석 (Logistic Linear Regression)

종속변수와 독립변수 간의 관계를 나타내어 예측모델을 생성한다는 점에서 선형회귀분석 방법과 유사하다. 하지만 독립변수(x)에 의해서 종속변수(y)의 범주로 분류한다는 측면은 분류분석 방법으로 간주된다.

특징

1) 분석 목적 : 종속변수(y 범주형)와 독립변수(x 연속형) 간의 관계를 통해서 예측모델을 생성하는데 있다.
2) 회귀분석과 차이점 : 종속변수는 반드시 범주형이어야 한다.
   (이항형 : Yes/No 또는 다항형 : iris의 Spices 칼럼)
3) 정규성 : 정규분포 대신에 이항분포를 따른다.
4) 로짓변환 : 종속변수의 출력범위를 0과 1로 조정하는 과정을 의미한다.
   예)혈액형 A인 경우 => [1,0,0,0]
5) 활용분야 : 의료, 통신, 날씨 등 다양한 분야에서 활용

로지스틱회귀분석은 선형회귀분석과 달리 종속변수(y 결과변수)가 범주형 데이터인 경우에 사용되는 기법이다. 선형회귀분석 모델에서는 독립변수(x 설명변수)를 입력하면 수치형 결과를 얻게 된다.

그러므로 신장, 시험성적, 연간소득 따위를 예측하고 싶은 경우에는 선형분석을 사용하면 된다. 이와 달리 예측하고자 하는 것이 수치화 하기 힘든 변수, 예를 들어 어떤 고객이 부도를 낼 것인지의 여부, 타이타닉호에서 살아남을 것인지 여부, 어떤 사람의 최종학력 알기 등의 경우에는 로지스틱회귀분석을 사용하면 된다.
회귀분석의 일종이므로 당연히 지도학습에 해당된다.

로지스틱 회귀분석

연속형 자료에 따른 범주형 자료의 영향력을 파악하기 위해, 로지스틱 회귀분석을 사용한다.

영향을 주는 변수	영향을 받는 변수	통계분석방법
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱 회귀분석

로지스틱 회귀분석이라 하면, 통계를 좀 아는 분들도 어려워하지만, 사실 회귀분석과 큰 차이가 없다.

그럼 좀 더 구체적인 연구 문제 예시를 통해 알아보도록 하겠다.

연구문제 : 정치관심도, 여당선호도, 야당선호도가 선거참여에 미치는 영향
통계분석방법 : 로지스틱 회귀분석

연속형 자료 {정치관심도, 여당선호도, 야당선호도}가 범주형 자료 {선거참여}에 미치는 영향

[연속형 자료] 정치관심도, 여당선호도, 야당선호도

[범주형 자료] 선거참여 ⇒ 선거 참여를 했다. ⇒ 1
선거 참여를 하지 않았다. ⇒ 0

위의 그림처럼, **선거 참여 유무를 0과 1로 표현한다면 범주형 자료로 구분**된다.

영향을 주는 변수 (독립변수X)	**영향을 받는 변수** (종속변수Y)
정치 관심도 (5점 척도) 여당 선호도 (5점 척도) 야당 선호도 (5점 척도)	선거 참여 여부 (1: 참여, 0:비참여)

회귀분석 결과에서는 β 값이 나오지만, **로지스틱 회귀분석에서는 Exp(B) 값과 P 값을** 본다.

▶로지스틱 회귀분석, 결과 분석하기

로지스틱 회귀분석 결과 값이 다음과 같이 나왔다고 생각해 보자.

결과 예시	Exp(B)	P
정치 관심도	1.324	0.01
여당 선호도	0.800	0.01
야당 선호도	1.010	0.90

정치관심도에 대한 Exp(B) 값이 1.324 P 값이 0.001이다.

P 값이 0.05보다 작기 때문에 유의미하다는 결과가 나온다.

β 값은 0을 기준으로 0에 가까우면 유의미하지 않게 나오는데, **로지스틱 회귀분석은 1이 기준**이 된다.

해석할 때, 정치관심도가 1점 증가할수록, 선거에 참여할 확률이 1.324배 정도 높아진다는 결론이 나온다.

1배를 기준으로 해서 '몇 배 늘어난다'고 해석해야 한다.

만약에 **여당선호도에 대한 Exp(B)가 0.8, P 값이 0.01이 나온다**고 하면, 유의미한 것이다.

여당선호도가 1점 증가할수록 선거참여 할 확률은 0.8배로 떨어진다는 것이다.

* 0.8로 표시된 것은 0.8배 증가한 것이 아니라, 1배를 기준으로 20%정도 줄었다는 것이다.

결과 해석

정치 관심도가 1점 높아질수록, 선거참여 가능성은 1.324배 정도 높아진다.
여당 선호도가 1점 높아질수록, 선거참여 가능성은 0.800배 정도 낮아진다.
야당 선호도는 선거참여에 유의미한 영향을 미치지 못한다.

한마디로 여당을 좋아하는 사람들은 선거를 하지 않는다는 것이다.

로지스틱 회귀분석은 **연속형 변수들을 독립변수**로 놓고, **범주형 자료를 Yes or No로 나오는 범주형 변수를 종속변수로 활용**할 때, 진행할 수 있다.

Support Vector Machine(SVM)

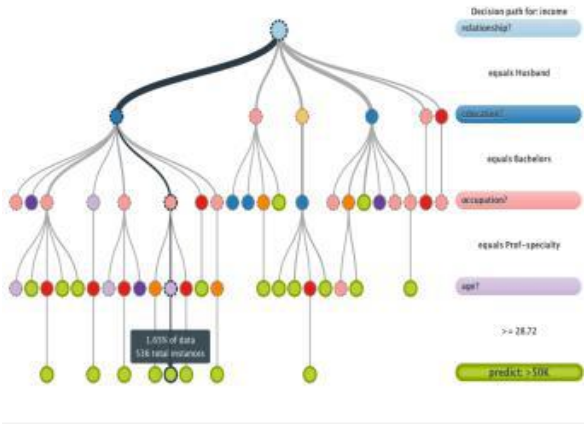
분류와 회귀분석을 위해 주로 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. SVM은 선형 분류와 더불어 비선형 분류에서도 사용될 수 있다

SVM은 다음과 같은 현실 세계의 문제들을 해결하는데 사용된다:

- SVM은 텍스트와 하이퍼텍스트를 분류하는데 있어서, 학습 데이터를 상당히 줄일 수 있게 해준다.
- 이미지를 분류하는 작업에서 SVM을 사용할 수 있다. SVM이 기존의 쿼리 개량 구조보다 상당히 높은 검색 정확도를 보인 것에 대한 실험 결과가 있다.
- SVM은 분류된 화합물에서 단백질을 90%까지 구분하는 의학 분야에 유용하게 사용된다.
- SVM을 통해서 손글씨의 특징을 인지할 수 있다.

분류분석 - Decision Tree -

- 종속변수(y변수) 존재
- 종속변수 : 예측에 Focus을 두는 변수
- 해석이 쉽다.
- 상호작용 효과 해석
- 비모수 검정 : 선형성, 정규성, 등분산성 가정 필요 없음



- Decision Tree는 여러 가지 규칙을 순차적으로 적용하면서 독립 변수 공간을 분할하는 분류 모형이다. 분류(classification)와 회귀 분석(regression)에 모두 사용될 수 있다.
- 단점 : 유의수준 판단 기준 없음(추론 기능 없음)

  비연속성/ 선형성 또는 주효과 결여/ 비안정성(분석용 자료에만 의존하므로, 새로운 자료의 예측에서는 불안정할 수 있음.
- 규칙(Rule)을 기반으로 의사결정트리 생성

  규칙(Rule)의 예 : 키가 크면 인기가 많음.

  규칙이 적용되어 tree 생성

RandomForest 분류모형

기계학습에서의 랜덤 포레스트(random forest)는 분류, 회귀분석 등에 사용되는 앙상블 학습방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치(회귀 분석)를 출력함으로써 동작한다.

–Classification and regression tree.

–CART: greedy, top-down binary, recursive partitioning

- 데이터를 몇 개의 영역으로 구분
- 구분되는 영역은 Y변수 값의 분류에 따라 결정

- 월등히 높은 정확성
- 간편하고 빠른 학습 및 테스트 알고리즘
- 변수소거 없이 수천 개의 입력 변수들을 다루는 것이 가능
- 임의화를 통한 좋은 일반화 성능
- 다중 클래스 알고리즘 특성

K-NN

지도 학습(Supervised Learning)의 한 종류로 레이블이 있는 데이터를 사용하여 분류 작업을 하는 알고리즘이다. 알고리즘의 이름에서 볼 수 있듯이 데이터로부터 거리가 가까운 k개의 다른 데이터의 레이블을 참조하여 분류하는 알고리즘이다. 주로 거리를 측정할 때 유클리디안 거리 계산법을 사용하여 거리를 측정하는데, 벡터의 크기가 커지면 계산이 복잡해진다.

- **kNN의 장점**

  - 알고리즘이 간단하여 구현하기 쉽다
  - 수치 기반 데이터 분류 작업에서 성능이 좋다

- **kNN의 단점**

  - 학습 데이터의 양이 많으면 분류 속도가 느려진다 (사실 사전 계산을 할 수 없기 때문에 학습 과정이 따로 없기 때문에 분류 속도가 느리다)
  - 차원(벡터)의 크기가 크면 계산량이 많아진다

Neural Network

인간이 뇌를 통해 문제를 처리하는 방법과 비슷한 방법으로 문제를 해결하기 위해 컴퓨터에서 채택하고 있는 구조. 인간은 뇌의 기본 구조 조직인 뉴런(neuron)과 뉴런이 연결되어 일을 처리하는 것처럼, 수학적 모델로서의 뉴런이 상호 연결되어 네트워크를 형성할 때 이를 신경망이라 한다.

이를 생물학적인 신경망과 구별하여 특히 인공 신경망(artificial neural network)이라고도 한다. 신경망은 각 뉴런이 독립적으로 동작하는 처리기의 역할을 하기 때문에 병렬성(parallellism)이 뛰어나고, 많은 연결선에 정보가 분산되어 있기 때문에 몇몇 뉴런에 문제가 발생하더라도 전체 시스템에 큰 영향을 주지 않으므로 결함 허용(fault tolerance) 능력이 있으며, 주어진 환경에 대한 학습 능력이 있다. 이와 같은 특성 때문에 인공지능 분야의 문제해결에 이용 되고 있으며, 문자 인식, 화상 처리, 자연 언어 처리, 음성 인식 등 여러 분야에서 이용되고 있다.

NEURAL NETWORK

신경망(neural network) 기법은 반복적인 학습 과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내고 이를 일반화함으로써 대용량 데이터로부터 의사결정에 필요한 유용한 정보를 찾아내는 블랙박스 기법이다. 여러 분야의 다양한 문제에 적용될 수 있고, 독립변수와 종속변수의 관계를 살피기가 어려운 복잡한 데이터에 대해서도 좋은 결과를 주는 것으로 알려져 있다. 하지만, 분류나 예측 결과만을 제공할 뿐 결과가 어떻게 나왔는가에 대한 이유를 설명하지 못한다. 따라서 신경망은 데이터의 여러 변수들간의 관계를 밝혀내는 데는 적합하지 않으며, 예측이 필요한 경우에 이용될 수 있다.

신경망 분석의 장단점

장점	적용 가능한 문제의 영역이 넓음	o 입력,출력마디에 이산형,연속형 변수 모두 사용가능하며 기법을 적용할 수 있는 문제의 영역이 Decision Tree나 통계에 비해 넓음
	제품이 많음	o 상용화된 데이터마이닝 제품이 많으며, 제품 선택의 폭이 넓음
단점	과정에 대한 설명부족	o 분류나 예측결과만 제공할 뿐 결과에 대한 근거를 설명하지 못함
	모형 구축의 어려움	o 복잡한 학습과정을 거치기 때문에 모형 구축시 많은 시간이 소요. 따라서 입력변수의 수가 너무 많으면 통계나 Decision Tree를 이용, 변수 선별 후 구축하는 대안을 고려 할 수 있음
	전문가 필요	o 다양한 Parameter값을 설정하는 작업이 전문성을 필요로 하기 때문에 비전문가들이 사용하기 어려움

NEURAL NETWORK

신경망 분석의 활용

가. 신경망 활용 사례

- 인공지능 자동차 에어컨 (설정온도, 실내온도, 외부온도, 날씨, 시간)
- 주택가격결정(위치,층수,평형,주변시세,대중교통,브랜드,유형)
- 프로선수연봉측정(방어율, 투구수, 총실점, 던진이닝,연차,선수유형)
- 얼굴인식(눈크기, 코비율, 눈사이비율, 눈썹형태, 입술모양, 피부색,상태)

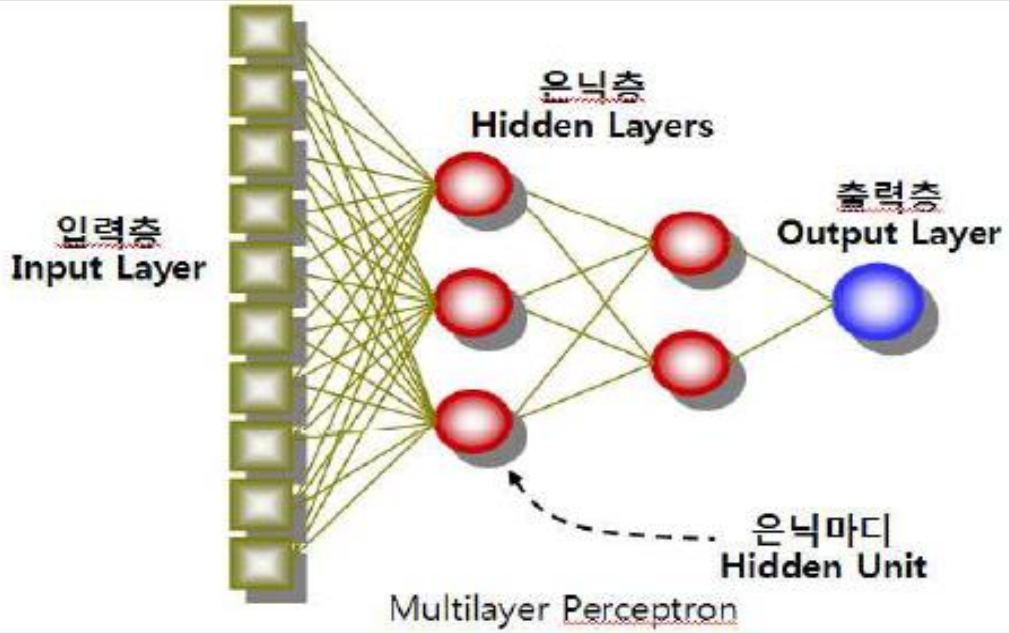
나. 신경회로망 응용가능분야

- 항공기 : 항공기 자동항법 및 제어시스템, 부품 고장 진단
- 자동차 : 자동차 자동항법시스템, 브레이크 진단 시스템
- 은행 : 신용평가
- 국방 : 목표 추적, 얼굴 인식, 센서, 레이더, 이미지 신호처리, 물체 구분
- 전자 : 직접회로 칩 레이아웃, 공정 제어, 칩 고장 분석, 비전, 음성인식,
- 문자인식, 홍채인식, 비선형 모델링
- 의학 : 암세포 분석, EEG & ECG 분석, 이식 시간 최적화
- 재정 : 대출상담, 신용카드 사용 분석, 통화 분석 및 예측
- 로봇 : 로봇 암 제어, 기계 비전, 궤적 제어, 이동로봇 경로 제어
- 증권 : 시장 분석, 채권 분석, 주가 예측 및 상담 시스템

NEURAL NETWORK

신경망의 구조 및 학습방법

가. 신경망의 구조



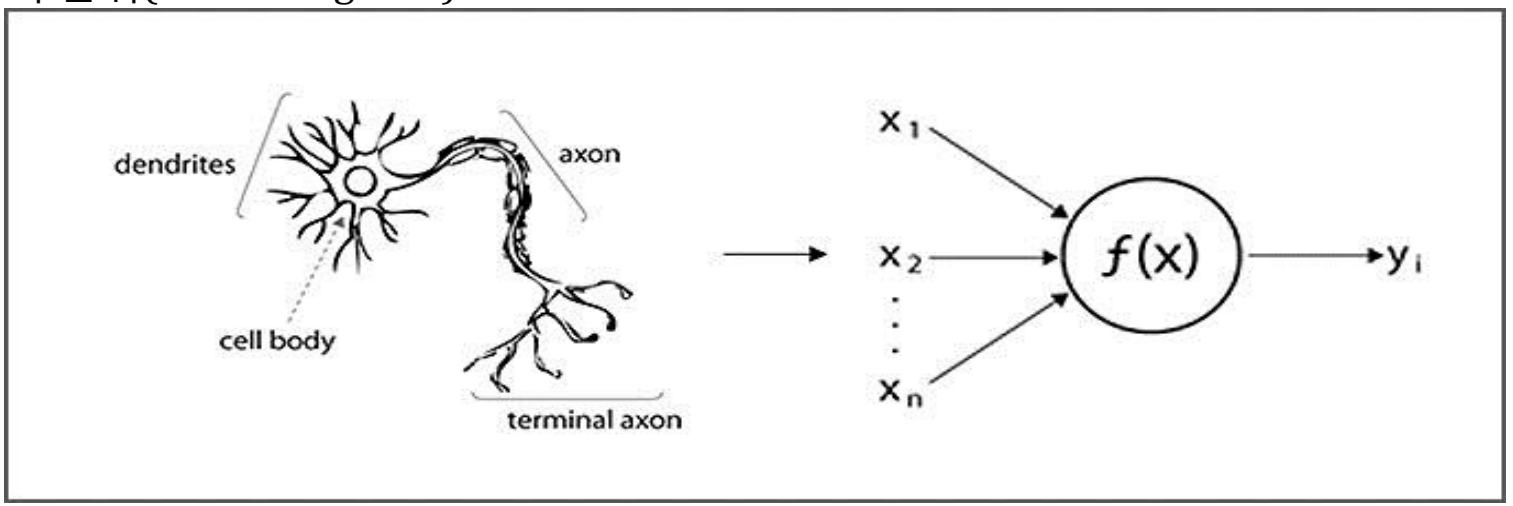
입력층	o 입력변수에 대응
은닉층	o 가중치와 활성화 함수에 대응
출력층	o 출력결과에 따른 표현 또는 실행
가중치	o 입력값에 대한 노드의 연결강도
활성화 함수	o 임계치에 따라 다음노드로 출력값 결정

Neural Network

생물학적 신경망을 모방하여 인공신경망을 모델링한 내용을 보면 처리 단위 측면에서는 생물적인 뉴런(neurons)이 노드(nodes)로, 연결성(Connections)은 시냅스(Synapse)가 가중치로 모델링 되었다.

생물학적 신경망	인공신경망
세포체	노드(Node)
수상돌기(dendrite)	입력(Input)
축삭(Axon)	출력(Output)
시냅스	가중치(Weight)

- 처리 단위(Processing Unit) : Neuron vs. node



Neural Network - 실생활에서 적용

인공 신경망은 다음과 같은 몇 가지 종류로 사용될 수 있다.

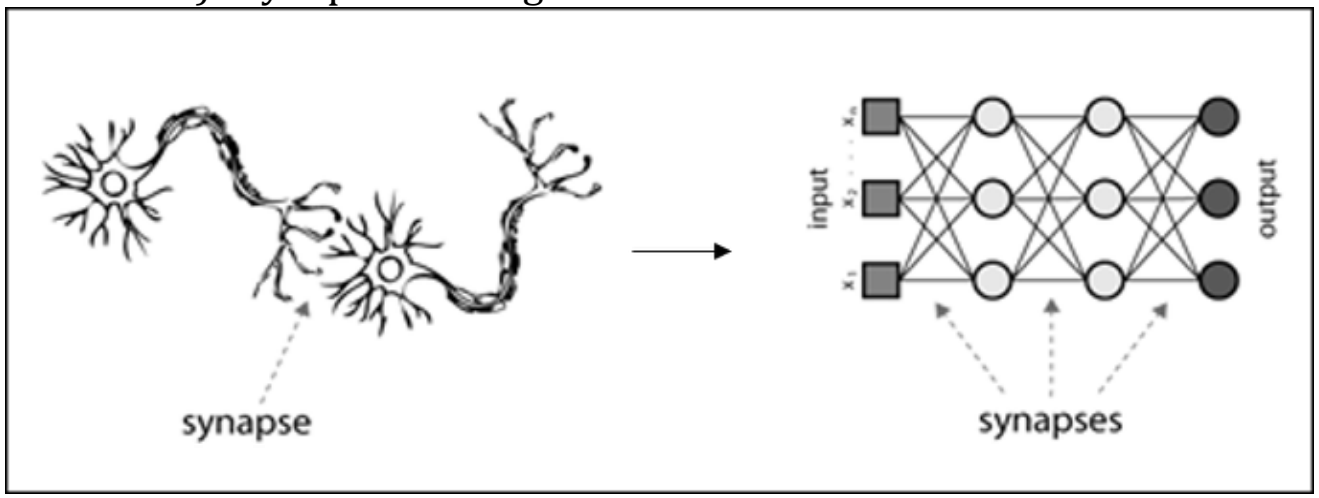
- 함수 추론, 회귀 분석, 시계열 예측, 근사 모델링
- 패턴 인식 및 순서 인식 그리고 순차 결정 같은 분류 알고리즘
- 필터링, 클러스터링, 압축 등의 데이터 처리
- 인공 기관의 움직임 조정 같은 로봇 제어
- 컴퓨터 수치 제어

또한 인공신경망은 여러 가지 암 진단에도 사용되었다. HLDN 이라는 인공 신경망 기반 폐암 검출 시스템은 암 진단의 정확성과 속도 향상을 이루었고 전립선 암에도 사용되었다. 이 시스템은 많은 환자의 데이터로부터 특정한 모델을 만들어서 모델과 환자 한명과 비교를 통해서 진단한다.모델은 다른 변수의 상관관계나 가정에 의존하지 않는다. 인공 신경망 모델은 임상 실험 방법보다 더 정확하게 동작 하였고 한 기관에서 훈련된 모델이 다른 기관에서도 결과를 예측 할 수 있었다.

Neural Network

신경세포(뉴런)의 입력은 다수이고 출력은 하나이며, 여러 신경세포로부터 전달되어 온 신호들은 합산되어 출력된다. 합산된 값이 설정값 이상이면 출력 신호가 생기고 이하이면 출력 신호가 없다.

- 연결(connection) : Synapse vs. weight



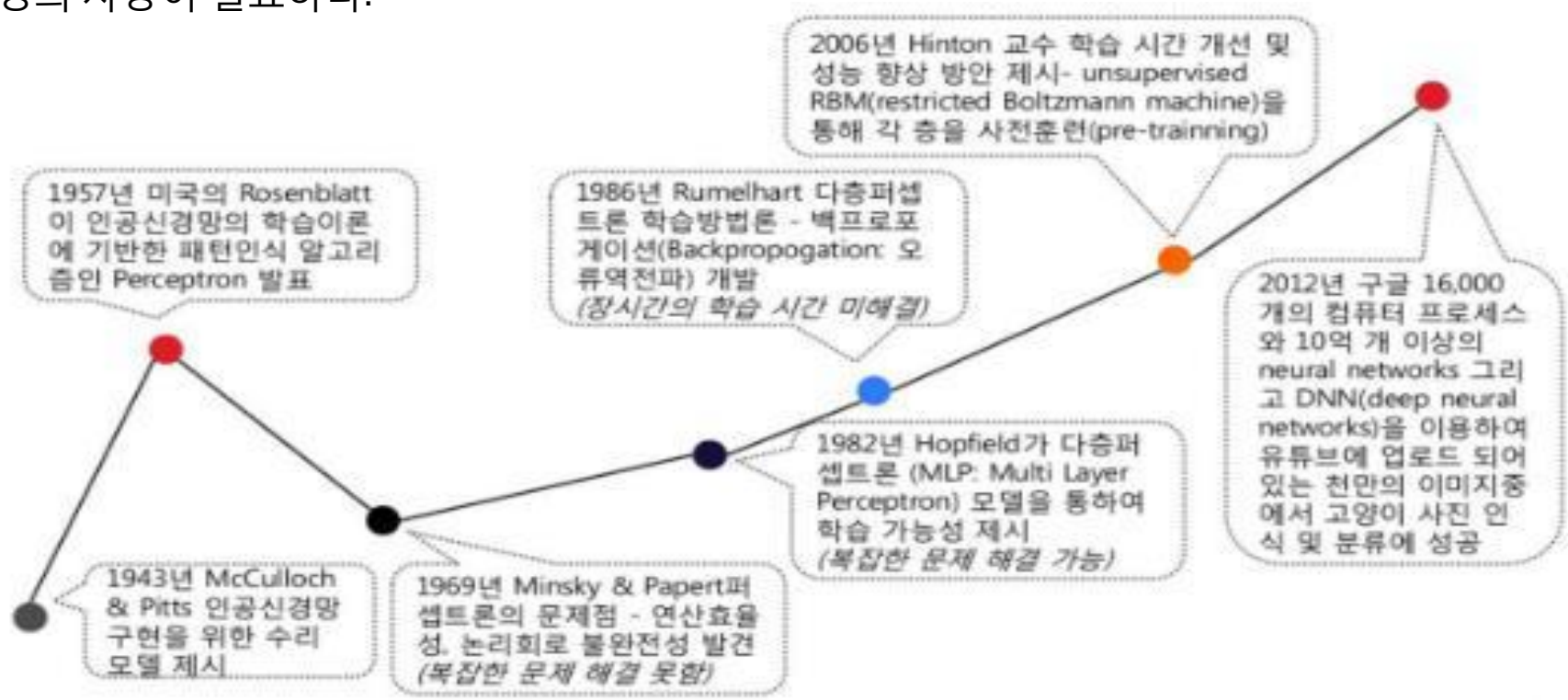
인간의 생물학적 신경세포가 하나가 아닌 다수가 연결되어 의미 있는 작업을 하듯, 인공신경망의 경우도 개별 뉴런들을 서로 시냅스를 통해 서로 연결시켜서 복수개의 계층(layer)이 서로 연결되어 각 층간의 연결 강도는 가중치로 수정(update) 가능하다. 이와 같이 다층 구조와 연결강도로 학습과 인지를 위한 분야에 활용된다.

DEEP LEARNING

다층 신경망의 성능문제의 해결방법이 2006년Hinton의 "A fast learning algorithm for deep belief nets"를 통해 제시되면서, 딥러닝으로 주목 받음

딥러닝은 은닉층을 계산에 활용하는 방식에 따라Convolutional Neural Network, Deep Belief Network, Recurrent Neural Network 등으로 나뉨

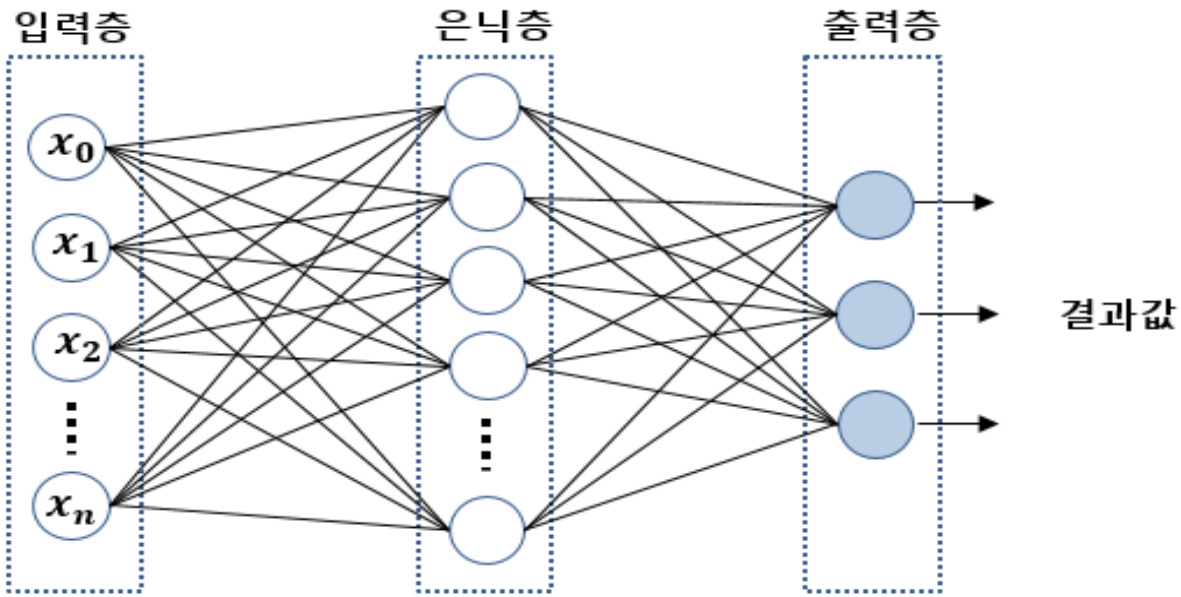
충분한 성능을 얻기 위해서는 대용량 계산의 처리가 가능해야 하며, GPU 또는 클라우드 컴퓨팅 환경의 사용이 필요하다.



다층 신경망 (MLP)

인공신경망인 단층 퍼셉트론은 그 한계가 있는데, 비선형적으로 분리되는 데이터에 대해서는 제대로 된 학습이 불가능하다는 것입니다. 예를 들면 단층 퍼셉트론으로 AND연산에 대해서는 학습이 가능하지만, XOR에 대해서는 학습이 불가능하다는 것이 증명되었습니다.

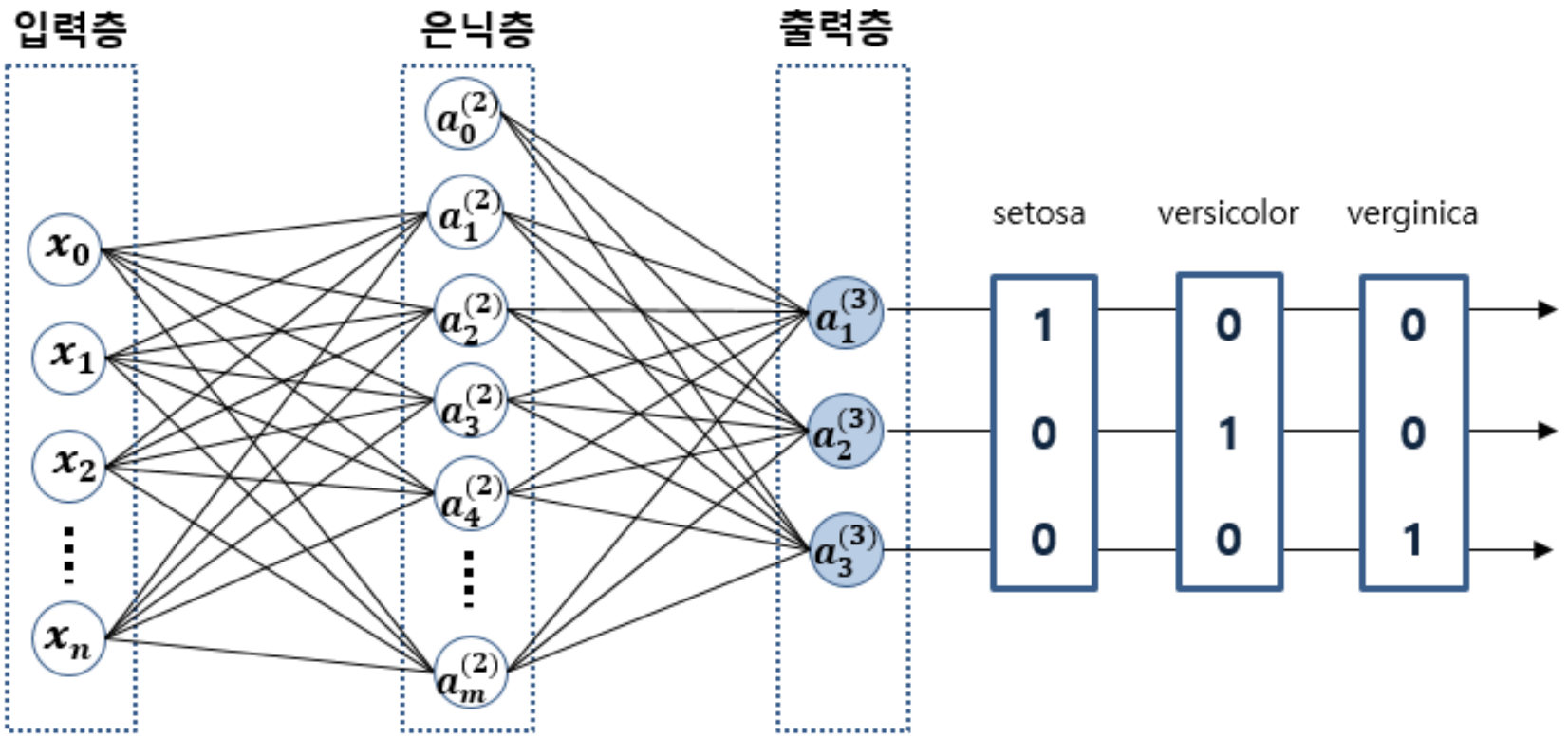
이를 극복하기 위한 방안으로 입력층과 출력층 사이에 하나 이상의 중간층을 두어 비선형적으로 분리되는 데이터에 대해서도 학습이 가능하도록 다층 퍼셉트론(줄여서 MLP)이 고안되었습니다. 아래 그림은 다층 퍼셉트론의 구조의 한 예를 보인 것입니다.



입력층과 출력층 사이에 존재하는 중간층을 숨어 있는 층이라 해서 은닉층이라 부릅니다. 입력층과 출력층 사이에 여러개의 은닉층이 있는 인공 신경망을 **심층 신경망(deep neural network)**이라 부르며, 심층 신경망을 학습하기 위해 고안된 특별한 알고리즘들을 **딥러닝(deep learning)**이라 부릅니다.

IRIS DATA로 MLPCLASSIFIER

단층 퍼셉트론과는 달리 다층 퍼셉트론의 출력층의 출력 노드는 여러개가 될 수 있습니다. 만약 3개 품종을 지닌 아이리스를 분류하는 다층 퍼셉트론은 출력층의 출력노드를 3개로 구성하고 그 결과값에 따라 품종 분류는 아래와 그림과 같은 형식으로 하면 될 것입니다.



출력층을 3개의 노드로 구성하고, 아이리스의 3개 품종 setosa, versicolor, verginica에 대한 실제값을 (1, 0, 0), (0, 1, 0), (0, 0, 1)로 정의합니다.

Visualization of MLP weights on MNIST

Sometimes looking at the learned coefficients of a neural network can provide insight into the learning behavior. For example if weights look unstructured, maybe some were not used at all, or if very large coefficients exist, maybe regularization was too low or the learning rate too high.

This example shows how to plot some of the first layer weights in a MLPClassifier trained on the MNIST dataset.

The input data consists of 28x28 pixel handwritten digits, leading to 784 features in the dataset. Therefore the first layer weight matrix have the shape (784, hidden_layer_sizes[0]). We can therefore visualize a single column of the weight matrix as a 28x28 pixel image.

To make the example run faster, we use very few hidden units, and train only for a very short time. Training longer would result in weights with a much smoother spatial appearance.

군집분석(CLUSTERING)

군집분석이란 개인 또는 여러 개체를 유사한 속성을 지닌 대상들끼리 그룹핑하는 탐색적 다변량 분석기법이다. 군집분석은 거리값(Distance Measure)을 이용해 가까운 거리에 있는 것들끼리 묶어 분류한다. 거리의 종류는 다양하지만 그 중 가장 흔히 사용하는 것은 유클리디안 거리를 사용한다. 군집분석은 크게 계층적 군집분석과 비계층적 군집으로 분류할 수 있다.

계층적 군집분석은 개별대상 간의 거리에 의하여 가장 가까이 있는 대상들로 부터 시작하여 결합해 감으로써 나무모양의 계층적 구조를 형성해 나가는 방법으로 이 과정에서 군집의 수가 감소한다. 계층적 군집분석은 군집이 형성되는 과정을 정확하게 파악할 수 있다는 장점이 있으나 자료의 크기가 크면 분석하기 어렵다는 단점이 있다.

방법 : 단일결합법, 완전결합법, 평균결합버, 중심결합기준법, Ward법

비계층적 군집분석은 군집의 수를 정한 상태에서 설정된 군집의 중심에서 가장 가까운 개체를 하나씩 포함해 나가는 방법으로 많은 자료를 빠르고 쉽게 분류할 수 있지만 군집의 수를 미리 정해 줘야 하고 군집을 형성하기 윈한 초기값에 따라 군집의 결과가 달라진다는 어려움이 있기 때문에 계층적 군집분석을 통해 대략적인 군집의 수를 파악하고 이를 초기 군집 수로 설정한다.

방법: k-means clustering

군집분석(CLUSTERING)

Clustering?

–Cluster의 개수나 구조에 관한 특별한 사전가정 없이, 개체들 사이의 유사성/거리에 근거해 cluster를 찾고 다음 단계의 분석을 하게 하는 기법

–유사한 개체들을 cluster로 그룹화하여 각 집단의 성격을 파악

Clustering 장점

–데이터에 대해 탐색적 기법으로, 데이터 내부구조 등이 주어지지 않아도 자료구조를 탐색

–추가적인 분석을 위해 사용할 수 있음

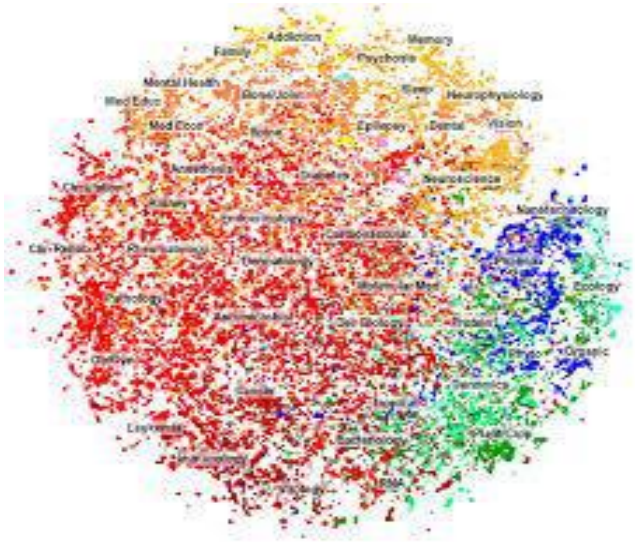
–유사성, 비유사성 만 계산할 수 있다면 여러 형태 데이터 적용 가능

Clustering 단점

–자료유형이 혼합된 경우, 거리정의 등이 어려울 수 있음

–초기 군집 수 설정이 중요

–결과 해석에 주의



군집분석(CLUSTERING)

군집분석이란?

데이터 간의 유사도를 정의하고 그 유사도에 가까운 것부터 순서대로 합쳐 가는 방법으로,
유사도의 정의에는 거리나 상관계수 등 여러가지가 있다.

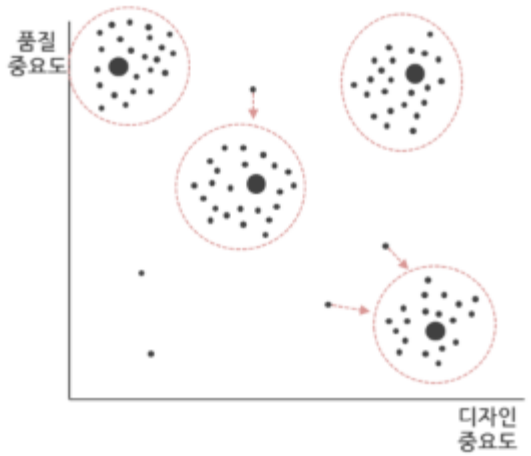
군집분석은 익숙하지 않아서 많이 어려워하지만, 개념만 제대로 이해하면 그리 어려운 통계분석 방법은 아니다.

군집분석은 본 분석을 들어가기 앞서, 사전작업이라고 볼 수 있다.
비슷한 특성을 가진 개체를 그룹으로 만들고, 그 그룹 간 서로 비교하는 것이 일반적인 통계분석의 흐름이다.

그럼 조금 더 구체적인 예시와 함께 알아보도록 하겠다.
예를 들어, 소비자의 특성을 군집(그룹)을 분리한다고 생각해보자.

다양한 변수가 있다면 3,4차원의 그래프로 표현될 수 있지만,
X축의 디자인 중요도, Y축의 품질 중요도로 2차원 그래프로 예를 들어보겠다.

어떤 제품을 볼 때, 품질을 보는 사람이 있는지, 디자인을 중요하게 보는 사람이 있는지 본다고 가정하겠다.
디자인만 중요시 하는 사람도 있고, 품질을 중요시 하는 사람도 있을 거고,
품질과 디자인 둘 다를 중요시 하는 사람도 있을 것이다.



그룹을 가장 가까이 묶인 것들끼리 그룹이 나뉠 수 있다.
이렇게 군집을 나누면, 특성을 4개의 그룹으로 분리할 수 있다.

* 군집분석의 절차

단계1) 데이터로부터 군집분석에 사용할 주요변수 추출

단계2) 계층적 군집분석을 통한 대략적인 군집의 수 결정

단계3) 계층적 군집분석에 대한 타당성 검증

단계4) 비계층적 군집분석을 통한 군집분류

단계5) 분류된 군집의 특성 파악 및 군집명 결정

고객DB => 알고리즘 => 군집 알고리즘을 통해서(패턴으로) 근거리 모형으로 군집형성 - 규칙(rule)

변수에 의해서 그룹핑 된다.

변수 적용 : 상품카테고리, 구매금액, 총거래금액

유사객체를 묶어준다.

유사성 거리 : 유클리드 거리

y변수가 없는 데이터 마이닝 기법

예) 몸, 키 관점에서 묶음 => 3개 군집 <= 3개 군집의 특징 요약

유클리드 거리(EUCLIDEAN DISTANCE)

공간에서 두 점 사이의 거리를 계산하는 방법으로 이 거리를 이용하여 유클리드 공간을 정의한다. 우리는 쉽게 x축과 y축으로 구성된 2차원에 두 점이 있고 그 두 점 사이의 거리를 측정하는 것은 피타고라스 정의를 이용해 쉽게 할 수 있다. 하지만 다차원 좌표에서의 두 점의 거리를 재는 것은 쉽게 할 수 없다. 이 때 이 "유클리디안 거리"공식을 사용하면 된다. 참고로 거리를 계산하는 다른 방법으로 맨하탄 거리(Manhattan Distance)도 있다.

K-means Clustering

군집화는 아무런 정보가 없는 상태에서 데이터를 분류하는 방법이다. 이 중 가장 유명하고 간단한 K-means Clustering을 예시로 살펴볼 것이다.

K-means Clustering 이란 데이터 분류 종류를 K개 라고 했을 때 입력한 데이터 중 임의로 선택된 K 개의 기준과 각 점들의 거리를 오차로 생각하고 각각의 점들은 거리가 가장 가까운 기준에 해당한다고 생각하는 것이다. 그리고 이제 각각 기준에 해당하는 점들 모두의 평균을 새로운 기준으로 갱신해 나가게 된다. 이렇게 해서 가장 적절한 중심점들을 찾는 것이다. 이렇게 학습을 반복하면 깔끔하게 데이터를 분류할 수 있게 된다.

클러스터링은 일반적으로 4개의 유형으로 구분된다

- 클러스터 중심(centroid) 또는 평균 기반 클러스터링 k-means
- 빈도수가 많은 중간점(medoid) 기반 클러스터링 k-medoids
- 밀도 기반 클러스터링
- 계층적 클러스터링

K-MEANS CLUSTERING

몇 개의 그룹으로 나누어야 할 지가 관건이다. 이를 결정하는 방법으로 아래의 두 가지가 대표적이다

방법1) 엘보우(elbow) 기법

k-means 클러스터링은 클러스터 내 오차제곱합(SSE)의 합이 최소가 되도록 클러스터의 중심을 결정해 나가는 방법이다. 만약 클러스터의 개수를 1로 두고 계산한 SSE 값과, 클러스터의 개수를 2로 두고 계산한 SSE 값을 비교했을 때, 클러스터의 개수를 2로 두고 계산한 SSE 값이 더 작다면 1개의 클러스터 보다 2개의 클러스터가 더 적합하다고 볼 수 있다. 이런 식으로 클러스터의 개수를 늘려가면서 계산한 SSE를 그래프로 그려보면 SSE의 값이 점점 줄어들다가 어느 순간 줄어드는 비율이 급격하게 작아지는 부분이 생기는데, 그래프의 모양을 보면 팔꿈치 끄트머리 처럼 보이는 부분이 있는데 이 부분이 우리가 구하려는 최적의 클러스터 개수가 된다.

방법2) 실루엣(silhouette) 기법

클러스터링의 품질을 정량적으로 계산해 주는 방법이다. 자세한 수식은 인터넷 참조하자.

클러스터의 개수가 최적화되어 있으면 실루엣 계수의 값은 1에 가까운 값이 된다.

실루엣 기법은 k-means 클러스터링 기법 이외에 다른 클러스터링에도 적용이 가능하다