**1. Introduction**

- 준비한 dataset : 각 환자들의 정보(특징)들에 따라서 그 환자가 심장질환을 갖고 있는지의 유무를 분류한 데이터

- 본 프로그램의 목적 : 이 dataset을 활용해서 환자가 심장질환을 가질지에 대한 여부를 예측하고자 한다. logistic regression(로지스틱 회귀)알고리즘을 사용한다.

\* 로지스틱 회귀 알고리즘을 사용하는 이유(참과 거짓 중에 하나를 내놓는 과정이기 때문)

선형회귀와 마찬가지로 적절한 선을 그려나가는 과정. 다만, 직선이 아니라 참(1)과 거짓(0)사이를 구분하는 S자 형태의 선을 그어 주는 작업

이 S자 그래프 형태가 바로 시그모이드 함수(a는 경사도, b는 그래프의 좌우 이동을 결정)

그렇다면 a와 b의 값을 어떻게 구할 것인가? 🡪 “경사하강법” (근데 이것도 역시 오차를 구한 다음 오차가 작은 쪽으로 이동시키는 방법임, 그래서, 오차 공식을 도출하기 위해 로그함수를 사용하게 됨)

- 라이브러리

numpy: 행렬 연산을 위한 함수, 다차원 배열 함수, 과학연산을 해결할 때 씀

Pandas: 프로그래밍 언어를 위한 분석도구 제공

Matplotlib: 다양한 차트와 플롯 제공(한마디로 자료들을 바탕으로 그래프나 차트 등으로 시각화)

Seaborn: matplotlib을 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 패키지.

sklearn: 머신러닝 교육을 위한 파이썬 패키지. 다양한 머신 러닝 모형(알고리즘)을 하나의 패키지에서 모두 제공하고 있다는 점이 최대 장점

- sklearn.linear\_model

- sklearn.model\_selection

os모듈: 운영체제에서 수행하는 여러 기능을 파이썬에서 수행할 수 있게 해준다.

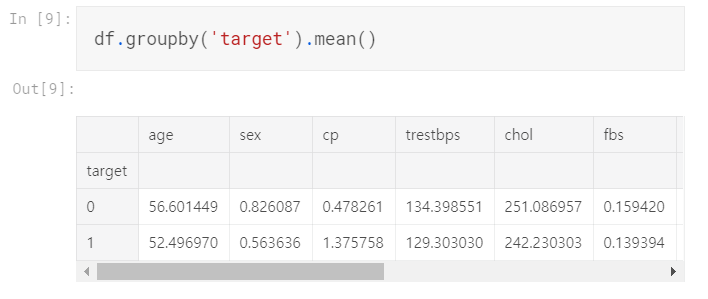
(우리의 경우 데이터를 입력할 때 경로 복사해서 입력한 거)

**2. Read Data**

(Data contains : age, sex, chest pain type[cp], resting blood pressure(trestbps), serum cholestoral in mg/dl(chol) … target(have disease or not : 1 = yes, 0 = no)

**3. Data Exploration**





이 외에도 다양한 데이터 시각화 과정을 해봄

**4. Creating Dummy Variables**

‘cp(chest pain type)’,

‘thal(maximum heart rate achieved)’,

‘slope(the slope of the peak exercise ST segment ; 심전도)’는 명목 상 이름뿐인 변수이기 때문에

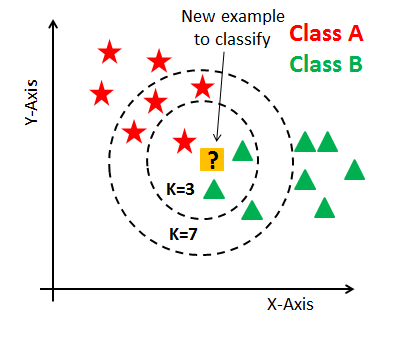
🡺 dummy variable로 취급하는 작업을 한다.

**5. Many Algorithm**

**KNN(K-Nearest Neighbors, 최근접 이웃)**

가장 간단한 “분류”알고리즘. 알고리즘 제작은 단순하게 데이터셋을 저장하는 과정만 있으면 된다.

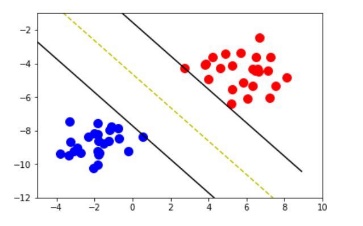
대상을 분류하는 방법은 대상에 근접한 그룹이 정해진 데이터셋들 중 가장 많은 그룹으로 대상을 분류



예를 들어 위의 사진을 보면 먼저 K의 수를 정한 다음 가장 근접한 K개의 데이터들의 그룹을 비교해 가장 많은 그룹으로 대상을 분류하며 예시에서는 K=3일 때 대상은 B로 분류되고 K=7일 때 대상이 A로 판명되는 모습을 볼 수 있다.

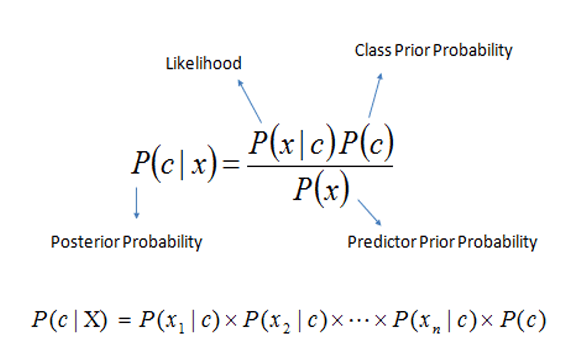
**SVM(Support Vector Machin, 서포트 벡터 머신)**

패턴을 인식하고 주어진 자료를 분석하는 모델이며 주로 분류와 회귀 분석을 위해 사용된다.



알기 쉽게 얘를 들어 설명하자면 위의 사진처럼 그룹이 정해진 데이터셋들이 주어져 있을 때 두그룹을 분류하는 기준선을 그은다(여기선 노란 점선). 이때 기준선은 두 그룹으로부터 가장 멀리 떨어져 있으면서도 양쪽과 거리가 같아야 한다. 그후 대상을 분류할 때 선을 기준으로 대상이 어떤 그룹으로 속하는 지 알 수 있다.

**Naive Bayes(나이브 베이즈 알고리즘)**



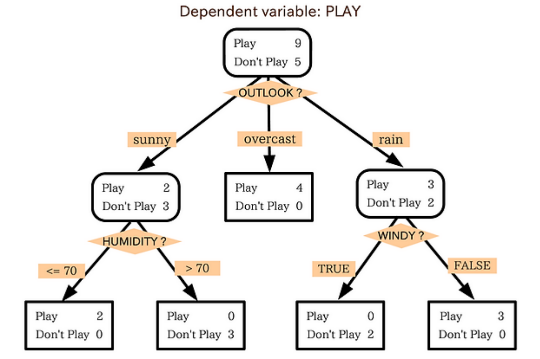
나이브 베이즈 알고리즘은 조건부 확률을 이용한 분류기다.

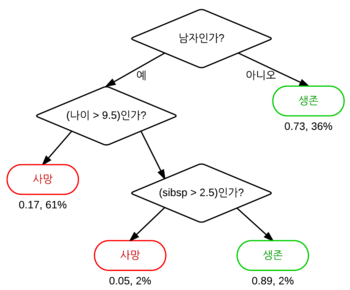
위의 사진은 x가 발생했을 때 c가 발생할 확률이며 예를 들자면 비가 오는데(x)우산이 없을(c) 확률을 생각해 볼 수 있다. P(우산이 없을 확률l비가 올 확률)=[P(xㅣc = 우산을 놓고 왔는데 비가 올 확률) x P(c=우산을 놓고 올 확률)] / P(x=비가 올 확률)이고 이것을 조건부 확률이라 한다.

이제 나이브 베이즈 알고리즘을 통한 예시를 들자면 ‘평소에 운동을 안하고 주로 고열량의 식단으로 식사를 해왔으며 스트레스를 많이 받아왔고 과한 음주습관이 있는 사람A’를 당뇨환자와 당뇨 환자가 아닌 사람으로 분류할 때 “주어진 상황에서 당뇨환자일 확률(조건부 확률)”을 통해 A를 당뇨 환자와 당뇨 환자가 아닌 사람으로 분류할 수 있다.

**decision tree(의사 결정 나무)**

여러가지 결정들의 경로와 결과를 나타내며 트리구조를 사용

[](http://imgur.com/ZKDnzOB)



트리구조 예시(타이타닉 호에서 생존할 확률)

간단히 말하자면 대상에 대한 여러가지 질문을 통해 대상이 어떤 그룹에 속하는지 분류

**Random forest(랜덤 포레스트)**

다수의 의사결정나무들로 이루어져 있으며 이 다수의 의사결정 나무들에서 도출된 결과(확률)을 종합해 대상을 분류

**6. Confusion matrix**

알고리즘의 성능을 평가하는 지표로 쓰입니다.

쉽게 말하면 알고리즘이 예측한 값과 실제 값을 비교해 얼마나 정확한지(아니면 오차의 정도)를 표현하며 주로 true positive + true false / true positive + true negative + false positive + false negative의 형태(1에 가까울수록 정확)로 알고리즘의 정확도를 알 수 있습니다.