x -> f -> Y

y = f(x) ; 여기서 f는 AI

* f가 지식과 경험 기반인 규칙 기반 알고리즘
* f가 데이터 기반 알고리즘

관계 검정 - 두 변수간에 관련 유무를 검정( 엄청나개 많은 X <=> Y)

* 관련 유무에 대한 정의를 독립 유무로 할 경우 실질적으로는 모집단의 두 변수 관계를 보는 것이지만, 모집단을 볼 수 없기 때문에 표본을 가지고 검정을 하게 됨.
* 두 변수 모두 Numeric일 경우 Correlation test를 하게 되는데, 주의할 건 corr=0이라고 해서 두 변수가 반드시 독립인 것은 아님.

모수적 방법과 비모수적 방법

* 모수적 방법 : 두 변수간의 관계를 가정
  + 장점 : 가정이 맞다면 다른 어느 방법보다 효율적임(더 적은 양의 데이터와 적은 계산량으로 결론 도출 가능)
  + 단점 : 가정이 틀리면 아무리 많은 데이터와 아무리 많은 계산량으로도 결론을 도출할 수 없음
  + ex) 선형회귀, 로지스틱 회귀
* 비모수적 방법 : 두 변수간의 관계를 가정하지 않음
  + 장점 : 언젠가 관계를 찾을 수 있음(엄청나게 많은 데이터와 계산량)
  + 단점 : 많은 양의 데이터와 계산량이 필요
  + ex) Randomforest(트리기반)

해석 가능성 vs 모델 복잡성(Trade-Off 관계)

설명도 잘하지만 예측도 좋은 모델에 대한 욕구

o 선형회귀

* OLS(ordinary least square) - 선형회귀모형의 최소자승법
* bo, b1을 찾은 후 => 모델의 성능 검증하는법 : RMSE(Roost MSE)
* RSS(Regression sum of square)
* R^2 = 1 - RSS/TSS(Total Sum of Square) ; 0~1사이의 값

o 단순회기, 다중회귀

o 범주형 변수가 존재하는 선형회귀 => 더미변수

* 범주형 변수의 레벨이 3개라면?(아시아인, 백인, 흑인) => 0, 1, 2 ? (대,중,소 가 아니라면 X!!!!!!!!!!!!!)
* 0, 1, 2으로 할 경우 아시아인과 백인의 차이가, 백인가 흑인의 차이와 같다고 하는 것
* 백인은 아시아인과 흑인 사이에 있다고 하는 것

(선형회귀는 0,1,2를 숫자로 인식하기 때문에)

* 결론적으로, 더미변수를 2개(n-1)를 만들어서 구분해야함(one-hot encoding)
* 문제는, 최근 카테고리 변수의 개수가 너무 많아지는 현상(ex, 카드상품 종류 1000개)
* 두 변수간의 상호작용이 존재할 때 => 넣었을 때 p-value 값으로 판단.

o kernel 기법(ex, SVM - 딥러닝 이전까지는 효과적이었음)

* 저차원의 데이터를 가지고 설명하기 힘들기 때문에, 차원을 늘려 데이터를 분석하고 다시 역함수를 통해 저차원으로 돌아옴