1.4.3 사례 기반 학습과 모델 기반 학습

◎ 사례 기반 학습

유사도를 기반으로 새로운 데이터와 학습한 샘플을 비교하는 방법

국가	1인당 GDP(US달러)	삶의 만족도
헝가리	12,240	4.9
대한민국	27,195	5.8
프랑스	37,675	6.5
호주	50,962	7.3
미국	55,805	7.2

○ 유사도

1. 유클리드 거리

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$

2. 마할라노비스 거리

$$d(x,y) = \sqrt{(x-y)\Sigma^{-1}(x-y)^T}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$

$$\Sigma = \begin{pmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{pmatrix}$$

$$Cov(X,Y) = E[(X_i - \bar{X})(Y_i - \bar{Y})]$$

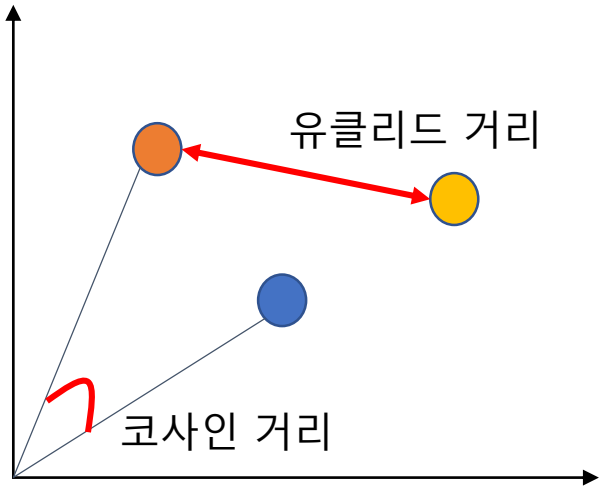
○ 유사도

3. 피어슨 상관계수

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. 코사인 거리

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$



○ 유사도

5. 자카드 계수

평가 점수가 이진형(예를 들어, 1: 구매, 0: 비구매)으로 주어지는 경우에 활용하는 유사성 측도

$$\text{자카드 계수} : J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

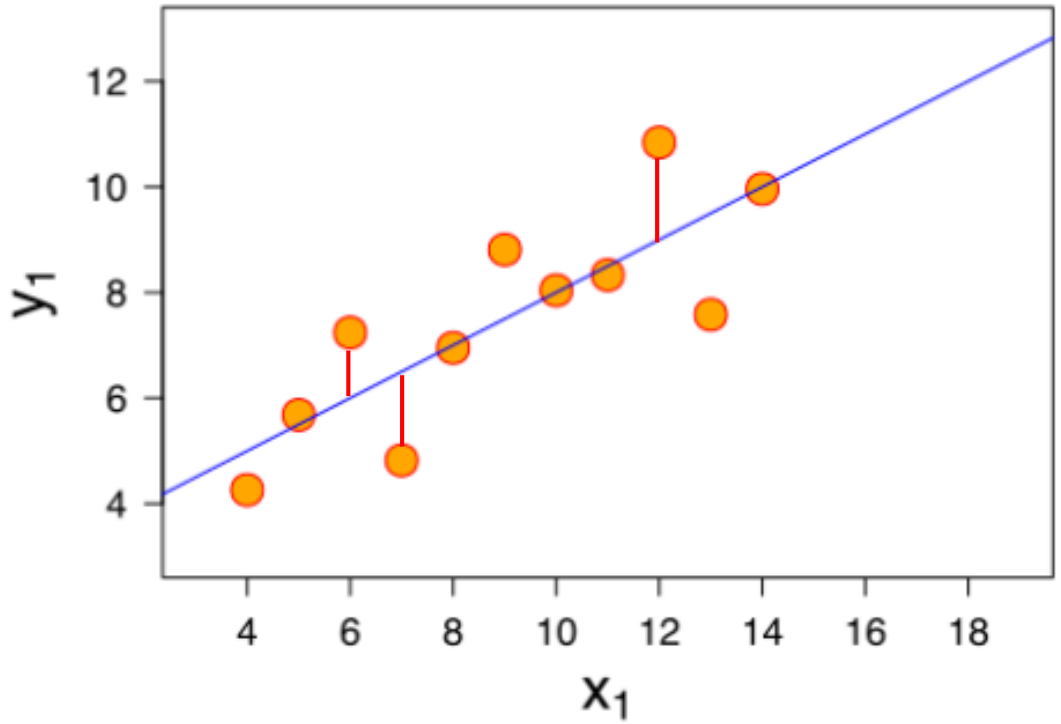
$|A \cup B|$: 사용자 A 또는 B가 구매한 항목의 수

$|A \cap B|$: 사용자 A와 B가 모두 구매한 항목의 수

사례 기반 학습과 모델 기반 학습

◎ 모델 기반 학습

훈련 샘플들을 활용하여 모델을 만들어 예측하는 방법



$$\widehat{y_1} = \theta_0 + \theta_1 \times x_1$$

○ 비용함수

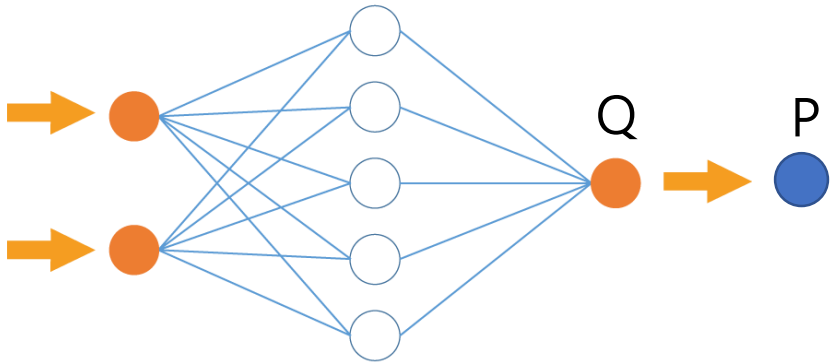
1. Mean Squared error(MSE, 평균 제곱 오차)

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

2. CrossEntropy

머신러닝 결과를 활용하여 만든 엔트로피

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$



1.5 머신러닝의 주요 도전 과제

1.5.1 충분하지 않은 양의 훈련 데이터

현재 기준으로는 많은 양의 데이터가 없다면 머신러닝 알고리즘 제대로 작동 불가