

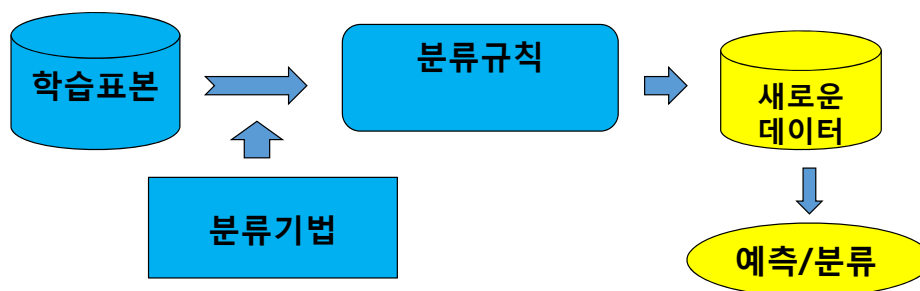
Wk9-3 : 데이터마이닝과 분류

- 분류규칙과 과적합-

2. 분류 (Classification)

- 분류분석 (classification analysis)은 다수의 속성 (attribute, variable)을 갖는 객체 (object)를 그룹 또는 범주 (class, category) 로 분류
- 학습표본 (training sample)으로부터 **효율적인** 분류규칙 (classification rule)을 생성.

오분류율 최소화 (cost function을 최소화)



2. 분류 (Classification) - 분류규칙

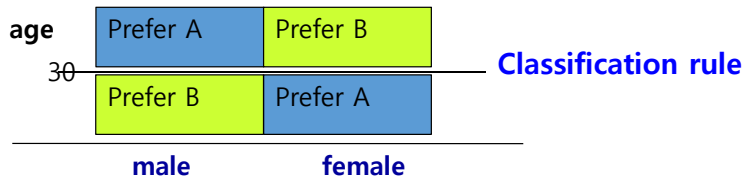
9.3 데이터마이닝-분류

- 이동통신회사 선호도 조사 (n=9), 타겟변수(선호통신사)=A, B

고객	1	2	3	4	5	6	7	8	9	New
성별	남	남	남	남	여	여	여	여	여	여
나이	20	23	35	41	19	24	25	33	39	50
선호회사	A	B	A	A	A	B	A	B	B	? B
규칙1	B	B	A	A	A	A	A	B	B	
규칙2	A	B	A	A	A	B	A	B	B	

- $Y=(A, B)$ - 범주가 2개인 분류문제, 속성 변수 2개 (성별, 나이)
- 위의 학습표본으로부터 다음과 같은 분류규칙을 얻었다고 가정하자.

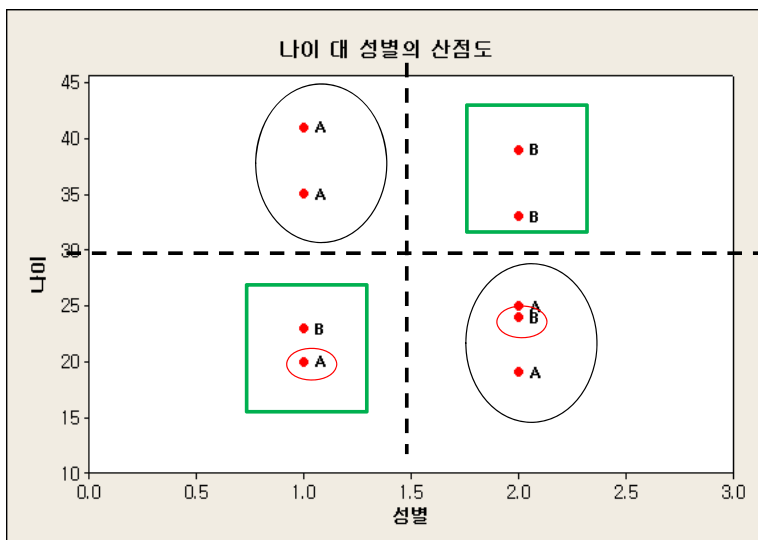
(분류 규칙 1) 남자이고 $Age \geq 30$ or 여자이고 $Age \leq 30 \Rightarrow$ 범주 1 (A)
그 밖의 경우이면 \Rightarrow 범주 2 (B)



2. 분류 - 오분류율(Misclassification rate)

9.3 데이터마이닝-분류

- 오분류율(misclassification rate) = 오분류 객체수/전체 객체수 = $2/9=0.22$



2. 분류 – 과적합(overfitting)

9.3 데이터마이닝-분류

앞의 분류문제에 대해서 (분류규칙 2)

(분류규칙 2) [남자 & ($\text{Age} \geq 30$ or $\text{Age} \leq 20$)] 혹은 [여자 & $\text{Age} \geq 30$] => 범주 1 (A)
그 밖의 경우이면 => 범주 2 (B) (단, 여자이고 나이가 24이면 범주 2)

(분류규칙 2) 에 따르면 오분류율 $0/9=0$ 이 된다. 학습표본에 대해서 오분류율을 0으로 인위적으로 만드는 경우 과적합 (overfitting)이라고 한다



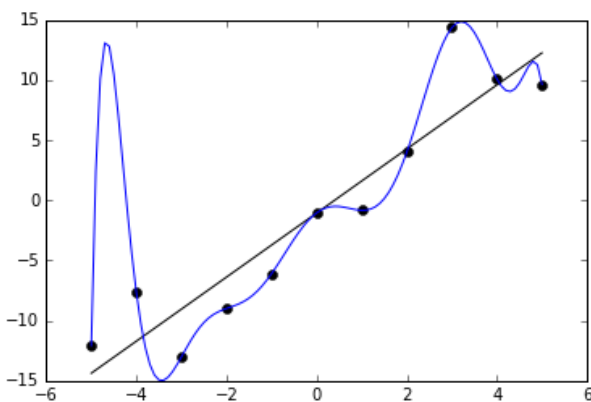
과적합 (overfitting)

- 분류모형에서 훈련데이터에 대한 과적합을 시킬 경우, 실제 데이터를 적용했을때 더 높은 오분류율 발생

2. 분류 – 과적합(overfitting)

9.3 데이터마이닝-분류

• 예측모형에서의 과적합



<https://en.wikipedia.org/wiki/Overfitting>

예측모형에서 훈련데이터에 대한 과한 적합모델을 선택하는 경우, 실제 데이터를 적용했을때 더 높은 오차를 발생



이런 과적합문제를 방지하기 위하여
학습데이터와 검증데이터를 5:5, 6:4, 7:3, 8:2 로 분리하여
모형의 성능을 비교평가한다

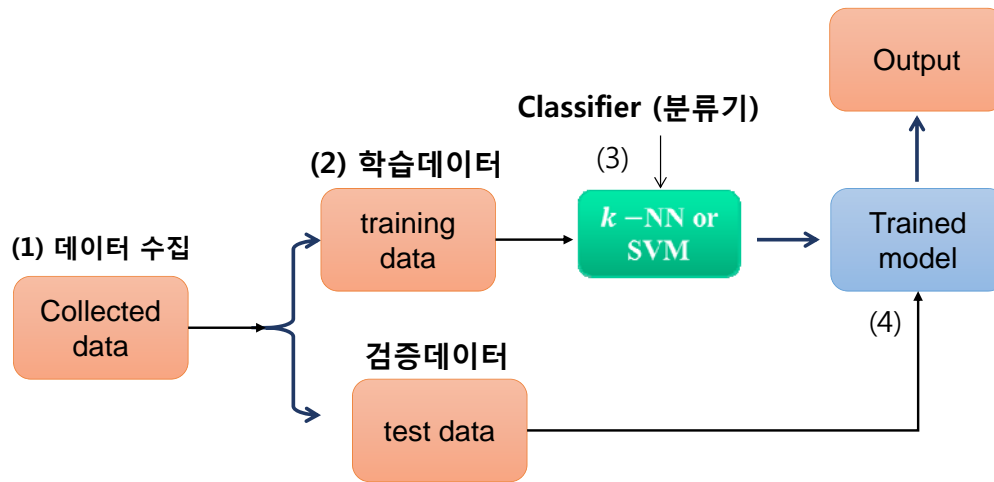
학습데이터 (70%)

검증데이터 (30%)

2. 분류- 교차검증(cross-validation)

9.3 데이터마이닝-분류

• 분류모형의 유효성 검증방법



2. 분류- 교차검증(cross-validation)

9.3 데이터마이닝-분류

• k-fold cross validation method 교차타당성 검증

5-fold cross-validation의 예제 : $n=100$ 이면 5등분으로 나누어 4등분은 학습데이터로 예측 모형을 구성하고, 나머지 5등분째 데이터로 검증한다

