GOETOPOIS (Data Mining)

한국방송통신대학교 정보통계학과 장영재교수 10 강 /

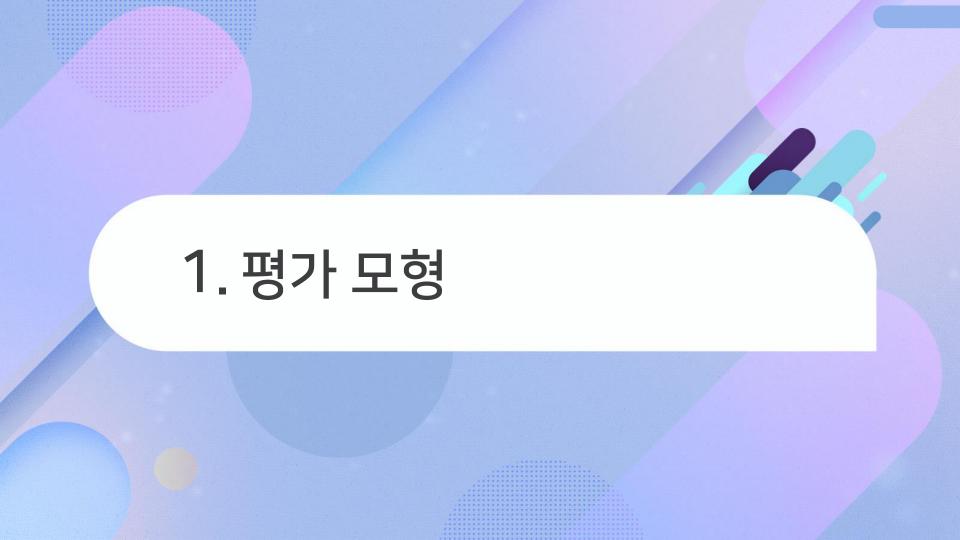
모형평가

MIIII

목차

10. 모형평가

- 1) 평가 모형
- 2) 평가 측도
- 3) 데이터 분할에 의한 타당도 평가



평가 모형

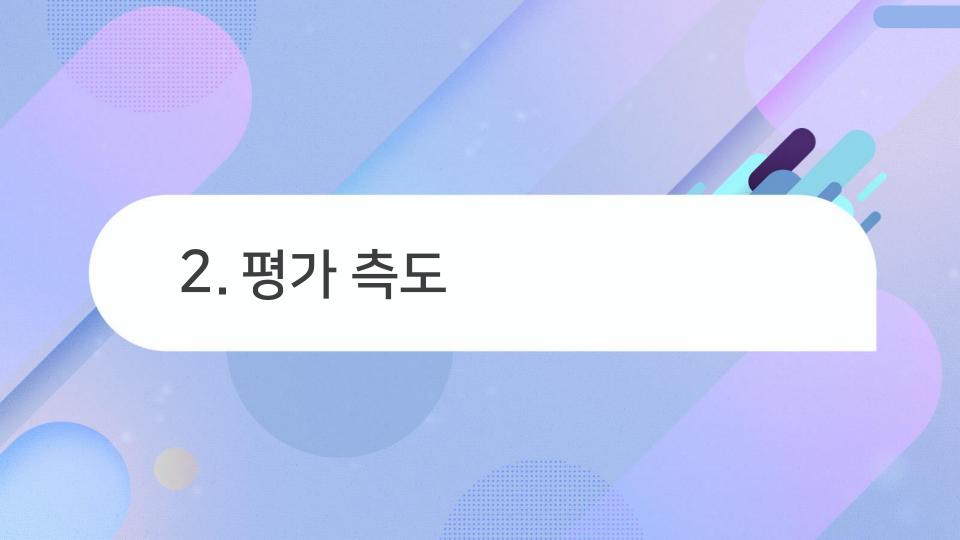
- ➢ 목표변수가 존재할 때 다양한 방법으로 모형을 구축하는 경우, 예측 값이 실제 값과 동일 또는 유사하다면 예측이 잘되었다고 평가
 - 데이터마다 예측력을 평가하여 최적의 모형을 선택

평가 모형

- 1) 연속형 목표 변수
 - 목표변수가 연속형일 때 선형회귀모형, 회귀나무모형, 또는 신경망 모형을 구축하여 각 객체의 목표변수의 예측 값 산출
 - 선형회귀모형, 회귀나무모형, 신경망모형, 그리고 랜덤포레스트에 의해 생성된
 모형을 비교 평가

평가 모형

- 2) 이항형 목표 변수
 - ➢ 목표변수가 이항형일 때 로지스틱회귀모형, 분류나무모형, 신경망모형, 앙상블 (배깅, 부스팅, 랜덤포레스트) 등을 사용해서 각 범주를 취할 확률 을 계산
 - 목표변수의 예측 값을 구하여 모형을 비교 평가



- ➤ 모형 선택시 예측력(prediction power), 해석력(interpretability), 효율성(efficiency), 안정성(stability) 등 다양한 측면을 고려
 - 데이터마이닝의 주목적이 예측이기 때문에 예측력이 가장 중요한 측도
 - 의학연구의 경우 또는 신용평가 분야에서는 예측 뿐만아니라 질병 예방을 위한 입력변수의 해석 또한 중요한 요소
 - 응용분야에 따라 어떤 요소가 중요한지 고려하여 종합적으로 모형을 평가하여 선택하여야 함

- 1) 연속형 목표 변수
 - ➤ 목표변수가 연속형인 경우에 모형의 예측력 측도로서 PMSE(prediction mean squared error)를 주로 사용

$$PMSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n$$

- 관측값(\hat{y}_i)과 예측값(\hat{y}_i) 사이에 차이가 적을수록 예측을 잘한 것이므로 PMSE 가 작을수록 그 모형의 예측력은 높다고 할 수 있음
 - 관측값과 예측값을 가로 및 세로축으로 놓고 그린 산점도(scatter plot)가 45도 대각선을 중심으로 모여 있으면 예측력이 좋다고 할 수 있음

- 2) 이항형 목표 변수
 - ▶ 실제 범주와 예측 범주를 분류표로 만든 정오분류표를 통해 예측력을 평가
 - 목표변수의 예측 값을 구하여 모형을 비교 평가

		예측 범주		합계
		1	0	합계
실제 범주	1	n_{11}	n_{10}	n_{1+}
	0	n_{01}	n_{00}	n_{0+}
합계		$n_{\pm 1}$	$n_{\pm 0}$	n

2) 이항형 목표 변수

- 예측력의 측도로 민감도(sensitivity)와 특이도(specificity)를 계산
- 민감도 = $Pr(\hat{Y}=1|Y=1) = n_{11}/n_{1+}$
- 특이도= Pr(Ŷ=0|Y=0) = n₀₀/n₀₊
- 예측정확도= $Pr(\hat{Y}=1,Y=1)+Pr(\hat{Y}=0,Y=0) = (n_{11}+n_{00})/n$
- 오분류율= Pr(Ŷ≠1,Y=1)+Pr(Ŷ≠0,Y=0) = (n₁₀+n₀₁)/n

2) 이항형 목표 변수

- 민감도와 특이도는 임계치에 따라 달라지고, 임계치는 상황에 따라 다르게 결정
- 여러 가능한 임계치에 대해 (1-특이도)를 가로축에, 민감도를 세로축에 놓고 그린 그래프를 ROC(receiver operating characteristic) 곡선
- 민감도와 특이도가 모두 높을수록 예측력이 좋다고 할 수 있기 때문에 ROC 곡선이 좌상단에 가까울수록 ROC 곡선 아래 면적(AUC; area under the ROC curve)이 커지고, AUC가 클수록 예측력이 좋다고 평가

3.데이터 분할에 의한 타당도 평가

3. 데이터 분할에 의한 타당도 평가

- 수집한 데이터를 이용하여 어떤 모형을 구축하게 되고 이를 비교 평가 하여 최종 모형을 선택
 - 이미 모형을 구축하기 위해 사용한 데이터를 재사용하여 모형을 평가하게 되면 과대평가하게 되어 결국에는 예측 오차(prediction error)가 더 커지게 될 수 있음
 - 이를 방지하기 위해 데이터를 분할하여 모형 구축 및 평가의 역할을 분담 하게 하는 기술을 종종 사용

3. 데이터 분할에 의한 타당도 평가

- 1) 훈련데이터와 검증데이터
 - ➤ 데이터를 분할하여 모형 구축을 위한 훈련데이터(training data)와 모형 평가를 위한 검증데이터(test data)로 임의로 나누어 사용
 - 모형이 커지거나 복잡해진다고 해서 검증데이터를 사용한 예측오차 (prediction error)가 작아지지는 않기 때문에 적절한 크기의 모형이 가장 좋은 예측력을 가지게 되며 이를 선택

3. 데이터 분할에 의한 타당도 평가

2) 교차 타당도

- 훈련데이터와 검증데이터로 분할하여 모형을 구축하고 평가하는 방법은 손쉽고 계산이 간단
 - 우연히 특정 모형에 유리하게 분할될 가능성이 존재하므로 여러 부분으로 분할하여 반복 검증하는 방법을 사용(교차타당도(cross validation)평가)

2) 교차 타당도

- 데이터를 V 개의 부분집합으로 분할하여 첫 번째 부분집합을 검증용으로 남겨두고 나머지 V-1개의 부분집합 데이터로 모형을 구축하고 남겨둔 데이터를 이용하여 오차를 계산
- 두 번째 부분집합을 남겨두고 나머지 부분집합으로 모형 구축, 남겨둔 부분 집합으로 오차 계산
- 마지막 부분집합을 검증용으로 사용하여 오차를 계산할 때까지 반복
- 이렇게 계산한 오차를 종합하여 모형을 평가.

강의를 마쳤습니다. 다음시간에는...