데이터마이닝(DataMining)

Chapter 2.2. 모형평가

최적의 부분모형 선택을 위한 기준

- 최적의 부분모형(best subset regression) 선택은 예측변수들의 모든 가능한 부분집합을 예측 변수로 하는 회귀모형(all possible subset regression)을 적합하고, 이 가운데 아래의 기준에 가장 잘 부합하는 모형을 찾는 방법
- 결정계수 : $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- 수정 결정계수 : $adj - R^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}$
- 평균제곱오차 : $MSE = \frac{SSE}{n-p-1} = \frac{\sum(y_i - \widehat{y_i})^2}{n-p-1}$
- Mallow's C_p : $C_p = p + \frac{(MSE - \widehat{\sigma^2})(n-p)}{\widehat{\sigma^2}}$

- p= 모수의 수(절편항 포함)
- SSE= p 개의 예측변수로 적합한 모형의 오차제곱합
- $\widehat{\sigma^2}$= 모든 예측변수를 포함한 적합모형의 평균제곱오차

- 모형평가

(i) 결정계수는 p에 대해 증가함수이므로 증가가 둔화되는 시점의 를 선택 (모형의 단순성을 위해)

(ii) 수정결정계수는 결정계수의 단점을 보완하여 설명력이 약한 예측변수가 추가될 때는 오히려 감소. 따라서 가장 큰 값을 가지는 p를 선택

(iii) 평균제곱오차가 최소가 되는 p를 선택

(iv) C_p의 값이 와 가장 가까운 값을 가지는 p를 선택한다

정보기준과 PRESS

- 회귀모형의 비교를 위해 다음의 여러 가지 정보기준(information criterion) 통계량이 사용
- 정보기준이 통계량 C_p보다 더 현실적인 방법으로 생각
- Akaike's 정보기준 (information critetion) : $AIC = n \ln\left(\frac{SSE}{n}\right) + 2p$
- Bayesian 정보기준 (information critetion) : $BIC = n \ln\left(\frac{SSE}{n}\right) + p \ln(n)$
- Amemiya's 예측기준 (Prediction Criterion) : $\text{APC} = \frac{n+p}{n(n-p)} SSE$
- 세 종류의 정보기준은 모두 작은 값을 가질수록 우수
- BIC는 AIC에 비해 모수의 수에 더 큰 벌점(penalty)을 부여하므로, 좀 더 단순한 모형을 선호
- AIC 기준이 과대적합의 경향이 있다는 일부 비판에 대한 보완으로 생각

- PRESS (prediction sum of squres)는 모형의 예측력을 통해 평가하는 방법
- $\widehat{y_{i(i)}}$: i -번째 자료를 제외하고 적합한 모형으로부터 i -번째 값을 추정한 것
- $PRESS = \sum_{i=1}^{n} \left(y_i - \widehat{y_{i(i)}} \right)^2$
- PRESS의 값이 작을수록 예측력이 우수

- 예측 결정계수 (predicted R^2) 값은 PRESS 보다 더 직관적으로 해석
- $R^2_{pred} = 1 - \frac{PRESS}{SST}$
- PRESS와 예측 결정계수는 모형 추정에 포함되지 않은 자료를 이용하여 계산되므로 과적합 (overfitting)을 방지하는데 도움
- 과적합은 모형 적합에 사용된 데이터에 대해서는 우수한 적합을 제공하나, 새로운 관측치에 대 해서는 유용한 적합을 보이지 못하는 것을 의미

교차타당법

- 교차타당법(cross-validation method)은 데이터셋을 모형구축에 사용될 훈련용셋(training set)과 예측력 평가에 사용될 평가용셋(validation or prediction set)으로 나누어 모형을 평가 하는 방법
- 데이터 양이 충분히 많은 경우에는 두 데이터셋의 비율을 50%:50%로 랜덤하게 나누어 적용
- k-중첩(fold) 교차타당법
  - 데이터의 양이 충분치 않은 경우에는 전체 데이터셋을 (동일한 크기의) 조각으로 나누고, 이 가운데 한 조각을 제외한 나머지 $(k-1)$ 조각으로 모형을 구축한 뒤, 남겨둔 한 조각에 대해 예측을 수행
  - 남겨지는 조각을 바꾸어 가며 이 절차를 k 번 반복
  - 각 조각에 대한 제곱예측오차를 더하여 교차타당 법의 측도로 사용

- Leave-one-out 교차타당법
  - k -중첩 교차타당법에서 $k = n$인 경우에 해당한다. 즉, 한 개의 데이터만 남기고 모형을 구축한 후 남겨진 한 개를 추정하는 과정을 반복
  - 계산량이 다소 많아질 수 있음
  - Leave-one-out 교차타당법 을 이용한 예측오차의 추정값은 PRESS와 동일

데이터마이닝에서의 모형 평가

- 대용량의 자료를 다루는 데이터마이닝에서 예측(분류)모형에 대한 평가는 보통 훈련용(training) 자료에 의해 구축된 모형을 검증용(test) 자료에 대해 적용하여 평가
- 모형평가에 사용되는 측도로는 범주형 반응변수에 대해서는 정오분류표(confusion matrix)에 기반한 여러 가지 측도(정분류율, 민감도, 특이도)들이 사용
- 연속형 반응변수에 대해서는 평균절대오차, 평균제곱오차 등이 사용

- 범주형 반응변수의 경우

		예측집단	
		C_1	C_2
실제집단	C_1	f_{11} (true positive)	f_{12} (false negative)
	C_2	f_{21} (false positive)	f_{22} (true negative)

- 정분류율(correct classification rate) 또는 정확도(accuracy): 전체에서 정확히 예측한 비율

$$\frac{f_{11} + f_{22}}{n}$$

- 민감도(sensitivity): 실제 참(true)인 것을 참(true)으로 제대로 분류한 비율

$$\frac{f_{11}}{f_{11} + f_{12}}$$

- 특이도(specificity): 실제 거짓(false)인 것을 거짓(false)으로 제대로 분류한 비율

$$\frac{f_{12}}{f_{21}+f_{22}}$$

- 정분류율은 다음과 같이 민감도와 특이도의 가중합으로도 표현 가능

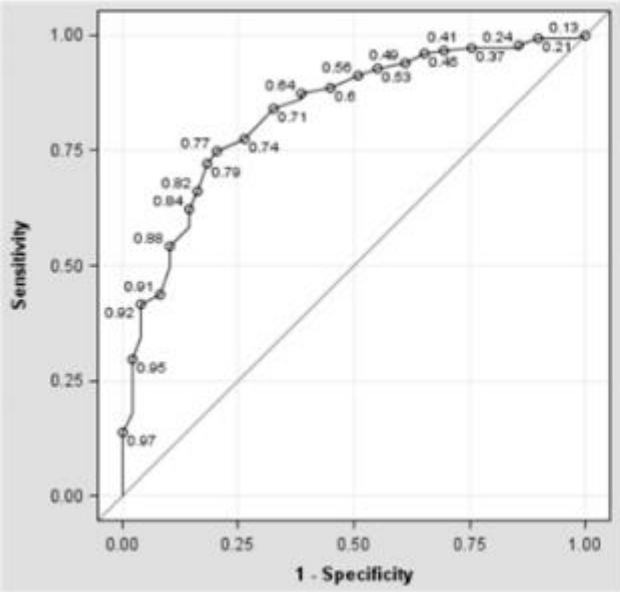
$$\frac{f_{11}+f_{12}}{n}\times\frac{f_{11}}{f_{11}+f_{12}}+\frac{f_{21}+f_{22}}{n}\times\frac{f_{12}}{f_{21}+f_{22}}$$

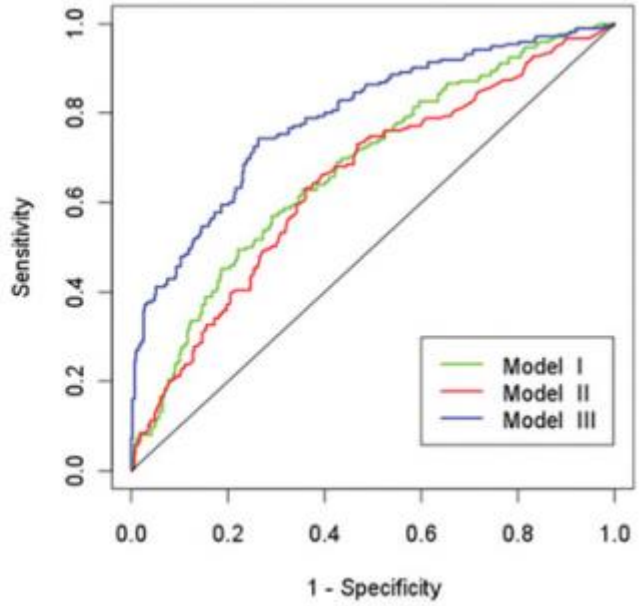
- 연속형 반응변수의 경우
- 평균절대오차(mean absolute error) : $\sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{n}$
- 평균제곱오차(mean squared error) : $\sum_{i=1}^{n} \frac{(y_i - \widehat{y_i})^2}{n}$

모형 선택을 위한 비교 방법

- 모형 선택을 위한 여러 가지 예측모형 간의 비교 방법에는 신뢰구간(또는 검정)을 이용하는 방법, ROC(Receiver Operating Characteristic) 그래프를 이용하는 방법
- 신뢰구간을 이용하는 방법은 두 개의 예측모형 간의 비교에 사용
- 예) 각 모형에 대해 k-중첩 교차타당도(k-fold cross-validation) 방법을 수행한 후 평균오차율을 추정하고, 이를 이용하여 오차율(=1-정확도)의 차이에 대한 신뢰구간을 구하여 비교하는 방법
- ROC 그래프는 로지스틱 회귀 또는 베이즈분류의 사후확률처럼 연속형의 값으로 주어질 때 유용
- 검증용 자료에 대해 예측값을 내림차순으로 정렬한 뒤, 분류를 위해 기준값을 선택하면 정오분류표를 얻게 된고, ROC 그래프는 기준값을 0→1로 변화시키면서 해당 정오분류표로부 터 (1-특이도)와 민감도의 값을 구하고, 이 값을 각각 x, y 좌표 상에 연결하여 그린 것

- ROC 그래프는 로지스틱 회귀 또는 베이즈분류의 사후확률처럼 연속형의 값으로 주어질 때 유용
- 검증용 자료에 대해 예측값을 내림차순으로 정렬한 뒤, 분류를 위해 기준값을 선택하면 정오분류표를 얻게 되고, ROC 그래프는 기준값을 0→1로 변화시키면서 해당 정오분류표로부 터 (1-특이도)와 민감도의 값을 구하고, 이 값을 각각 x, y 좌표 상에 연결하여 그린 것





- 데이터마이닝에서 예측모형에 대한 평가 방법으로 검증용 자료에 기반한 모형 평가(이를 예비법 (holdout method)이라 함)외에 예측모형(또는 분류기)을 평가하는 일반적인 방법으로 교차타당법, 붓스트랩(bootstrap) 방법 등이 있음