DS 04주차 팀과제

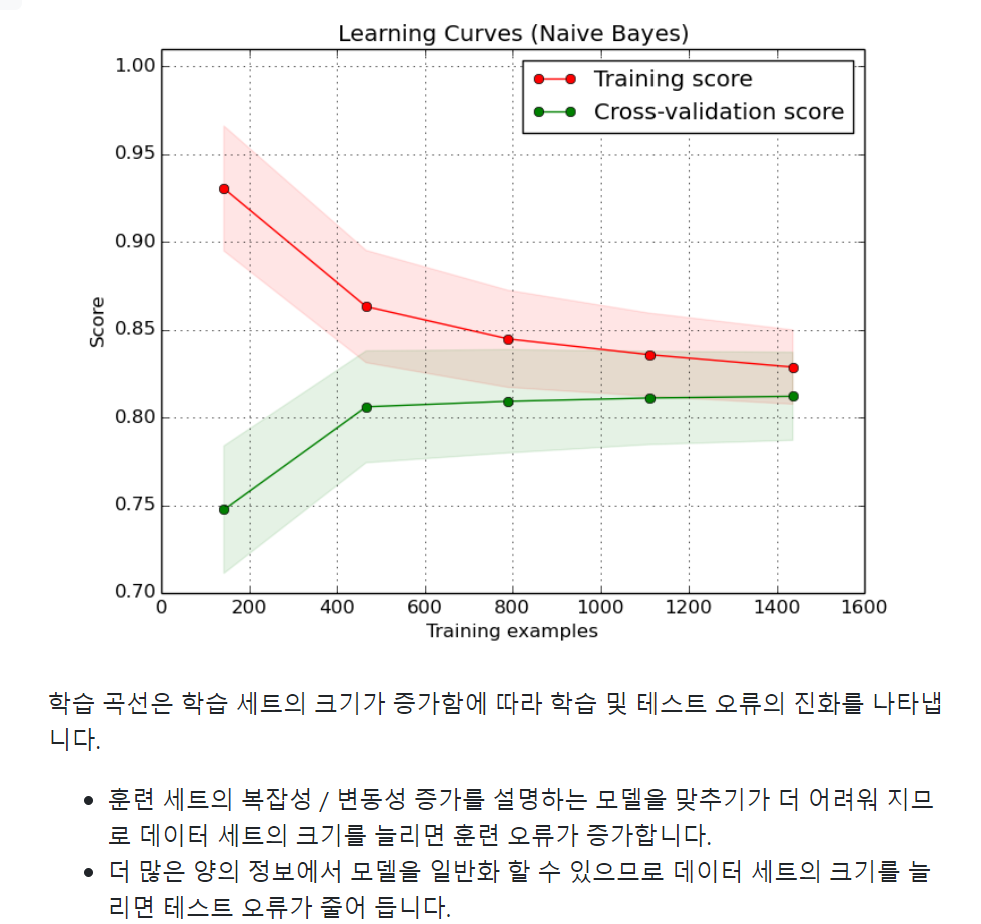
\* 아래 제시된 문항은 최소한의 가이드라인이며 과제는 아래 내용으로 제한하지 않습니다.

1. German Credit, 또는, Titanic, Iris 이외의 Dataset으로 의사결정나무 생성하기
   1. 단, 이러한 데이터와 상응하는 크기와 복잡도를 가지는 데이터셋을 선택할 것.
   2. Dataset에 대한 출처 및 설명 필요함
   3. 위의 데이터로 먼저 연습해 보기 바랍니다. – 예제를 Googling 해 보세요. – 과제 불포함
   4. 여러 예제들을 참고하는 것은 좋으나 예제를 그대로 답습하여 과제로 작성하는 것은 삼가합니다.
   5. 모델 생성을 위하여 적절한 최적화 과정이 필요하며 이 과정을 설명하여야 합니다.
   6. 목표변수를 적절하게 선택하여야 합니다.
   7. 적절한 데이터 전처리, 탐색과정을 거쳐야 합니다.
   8. 데이터 탐색 과정에서 분류 예측 정확도에 가장 영향력 있는 변수를 찾을 수 있는 방법은 무엇이며 이 방법으로 찾아 낸 변수는 무엇인가?
   9. 데이터를 학습용, 시험용 데이터 셋으로 나누기 위하여 목표변수를 기준으로 층화 추출한다.
   10. rpart 함수의 파라미터 중 minsplit, minbucket, maxdepth, method, cp 등은 각각 무엇을 의미하는가? 모델의 최적화를 위하여 여러가지 파라미터 값을 사용하여 모델 생성을 시도해 보고 최고의 성능을 나타내는 파라미터를 선택하기 바랍니다.
   11. 적절한 예측변수 집합을 선정하고 Decision tree 모델을 생성하여 정확도를 측정하라.
   12. 모든 변수를 포함하여 모델을 생성하고 정확도를 측정하고 위 문항에서의 결과와 비교하라.
   13. Split criterion의 종류, Pruning 여부를 선택함에 따라 tree의 형태 및 정확도는 어떻게 변화하는가? 정확도를 최대로 하는 최적 파라미터는 무엇인가?
   14. 위의 문항에서 찾아낸 최적 파라미터로 학습데이터 크기에 대한 학습곡선(Learning Curve)을 그려라. 학습 곡선 그래프의 x와 y축 은 각각 학습용 데이터의 크기 및 모델의 정확도이다. 곡선의 변화를 보이기 위하여 x 축에서 적절한 간격을 선택해야 한다.

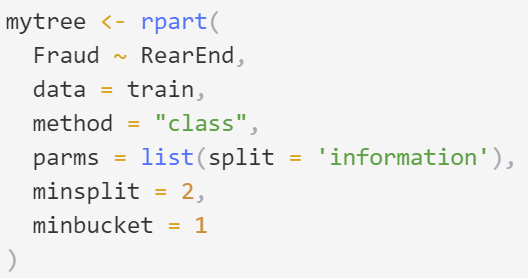
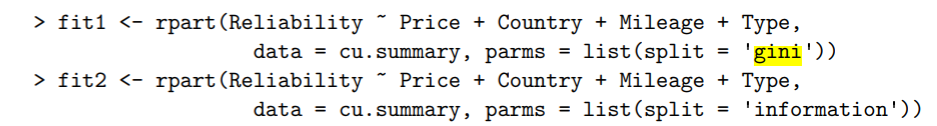
- 끝 -

참고 자료

1. 학습 곡선



1. Split criterion 설정   
   split의 default value는 “gini”

Reference: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf