결정나무의 가지치기

* 이과제에서는 사전가지치기와 사후가지치기를 적용하여 결정나무를 생성하고
* 두 가지치기 방법에 대하여 성능 비교를 한다.
* Data:
  + 이전 과제(결정나무 생성)의 데이터 사용

1. Amazon Mechanical Turk
   1. Amazon Mechanical Turk 란 무엇인가?
2. EDA
   1. 데이터에 대하여 간단한 데이터 탐색과정 – EDA 를 실행하라.
3. 데이터 전처리
   1. 결측치(missing data)에 대하여 조사하고 결측치를 적절히 채워 넣을 것.
4. 사전 가치치기를 적용하여 결정나무 모델 생성 - Pre-pruning
   1. Maxtreedepth 등의 rpart 파라미터를 최적화
   2. 파라미터 값의 최적 조합을 결정하라.
5. Full tree 로 모델을 성장 시킨 후 사후 가지치기를 적용하여 결정나무 모델 생성 - Post-pruning
   1. Full tree 로 모델을 성정 시키기 위해서는 파라미터를 어떻게 정해야 하겠는가?
   2. 가지 치기 - cp를 최적화
6. 위에서 생성된 두 예측 모델(사전, 사후 가지치기)의 성능을 비교하라.
   1. 각 모델의 정확도, 민감도, 특이도를 비교하라.
7. Plot, Learning curve 등을 이용하여 모델이 최적화되었음을 보일 것.
   1. 여러가지 learning curve plot – x 축은 최적화할 대상, y 축은 정확도 또는 에러율
8. 위의 두 모델 중 하나를 선택하여 rule set을 추출한다.
   1. The rpart.plot package version 3.0 (July 2018) 의 rpart.rules 함수 이용
   2. <https://stackoverflow.com/questions/36401411/extracting-information-from-the-decision-rules-in-rpart-package/51680932#51680932>
   3. 추출된 rule 중 가장 중요한 rule 은 무엇인가?  
      이 rule 이 적용되는 사례의 비율을 참조
   4. 추출된 rule 중 가장 신뢰도가 높은 rule은 무엇인가?  
      이 rule 이 적용되는 사례들의 예측 정확도를 참조

- 끝 -