Data Scaling은 지난 자료 참고

Bayesian Ridge

데이터 분포 함수를 도출할 때, 베이지안 확률을 사용함

신뢰도가 높을 것 같은 데이터와, 신뢰도가 낮을 것 같은 데이터를 동등히 취급

가중치를 바꾸어 가면서 Linear Regression 진행

L2 정규화를 진행하기 때문에 극단적인 데이터 (이상feature)의 가중치를 0에 가깝게 만듬

Decision Tree에는 Classification Tree와 Regression Tree가 있다.

랜덤 포레스트는 앙상블 학습으로써, 여러 학습기를 융합 시켜 성능을 향상하는 방법이다.

데이터에서 베깅을 통해 n쌍의 무작위 데이터를 만듬.

각각의 Decision Tree를 생성, 그 결과를 합침

RFE, RFECV

Feature selection의 가장 단순한 방법이다. 모든 feature들로부터 하나하나 제거함.

학습하고 싶은 모델을 정하고 데이터를 학습했을 때 각 feature의 중요도를 도출.

그리고 중요도가 낮은 feature를 하나하나 제거

RFECV는 feature의 개수마다 F-fold validation등의 Cross Validation을 활용하여 각기 다른 성능을 도출한다. 각 feature의 개수별 성능을 평균을 내고, 가장 높은 성능을 가지는 feature 개수에 해당하는 feature들을 최종 Feature selection 결과로 사용한다.

SMOTE

오버샘플링 기법 중 합성데이터를 생성하는 방식으로 가장 많이 사용되고 있는 모델이다.

합성 소수 샘플링 기술로 다수 클래스를 샘플링하고 기존 소수 샘플을 보간하여 새로운 소수 인스턴스를 합성해낸다.