

Big Data 분석

Regression Model

- 단순 선형회귀분석: 설명변수가 1개이고 목표변수와 선형관계를 갖는 회귀모델
 - $y = \beta_0 + \beta_1 x + \epsilon$
- 다중 선형회귀분석: 설명변수가 다수이고 목표변수와 선형관계를 갖는 회귀모델
 - $y = b_0 + b_1 x_0 + \dots + b_n x_n + \epsilon$
- 비선형회귀분석: 설명변수와 목표변수가 비선형 관계를 갖는 회귀모델(2차, 지수 등)
 - $y = b_0 + b_1 x_1 + b_2 x_2^2 + \dots + b_n x_n^n + \epsilon$

단순 선형회귀분석

설명변수가 1개이고 목표변수와 선형관계를 갖는 (확률적) 회귀 모델

모델의 회귀계수는 Least Square Method를 활용하여 추정

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- β_0 : 절편으로서 모델이 y 축을 통과하는 점(**intercept**)
- β_1 : 모델의 기울기로 x 가 1 단위 증가함에 따른 y 의 변화량(**coefficient**)
- ϵ : 모델의 잔차. 정규분포를 따르며, 등분산이고 서로 독립이다.
- 결정계수(R^2 , coefficient of determination): 설명력. 독립변수들이 종속변수를 얼마나 설명하냐를 보여주는 계수.
 - 전체 변동 중에서 모델(회귀선)에 의해 설명되는 변동의 크기로 판단 ($0 \leq R^2 \leq 1$)

$$\min \sum e_i^2 = \min \sum (Y_i - \hat{Y}_i)^2 \rightarrow \hat{Y}_i = a + bX_i$$

다중 선형회귀분석

목표변수와 다수(2개 이상)의 설명변수와의 선형관계를 분석하는 (확률적) 모델

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_n x_{ni} + \epsilon, \quad i = 1, \dots, n, \quad e_i \sim N(0, \sigma^2)$$

- β_0 : 절편으로서 모델이 y 축을 통과하는 점(**intercept**)

- β_j : j 번째 설명변수 x_j 의 회귀계수로 다른 모든 설명변수 값을 고정한 상태에서 x_j 가 1 단위 증가함에 따른 y 의 변화량(**coefficient**)
- ϵ : 모델의 잔차. 정규분포를 따르며, 등분산이고 서로 독립이다.
- 수정결정계수(*Adjusted R²*): 조정된 설명력. 다중선형회귀에서 사용한다.
 - 결정계수는 독립변수 개수가 많아질수록 그 값이 커지게 된다. 따라서 종속변수의 변동을 별로 설명해 주지 못하는 변수가 모형에 추가된다고 하더라도 결정계수값이 커질 수 있다.

다중공선성(Multicollinearity)

회귀분석에서 **설명변수 간의 높은 상관관계**로 발생하는 문제. 독립적으로 존재해야 하는 설명변수 X 들이 선형적 관계를 가지는 것.

Least Square Method로 추정된 모델의 회귀계수에 대한 신뢰성이 떨어짐. 따라서 **설명변수의 상대적 중요도(회귀계수) 및 목표변수에 대한 설명력의 수준이 문제가 됨.**

진단

1. **산점도** 또는 **상관계수**를 확인하여 설명변수 간의 선형관계를 파악한다.
2. **VIF**(Variation Inflation Factor, 분산팽창계수)가 10보다 크면 multicollinearity가 존재한다고 판단됨.
 - a. 단, VIF가 높다고 하더라도, 목표변수를 잘 설명하는 설명변수라면 무작정 제거하면 안된다.

처리

1. 상관관계가 높은 설명변수 중 **일부 변수를 제거**
2. 변수를 변환하거나 새로운 데이터 추가
3. **주성분 분석**(PCA, Principle Component Analysis)를 이용한 diagonal matrix의 형태로 공선성을 없애준다.

잔차 분석

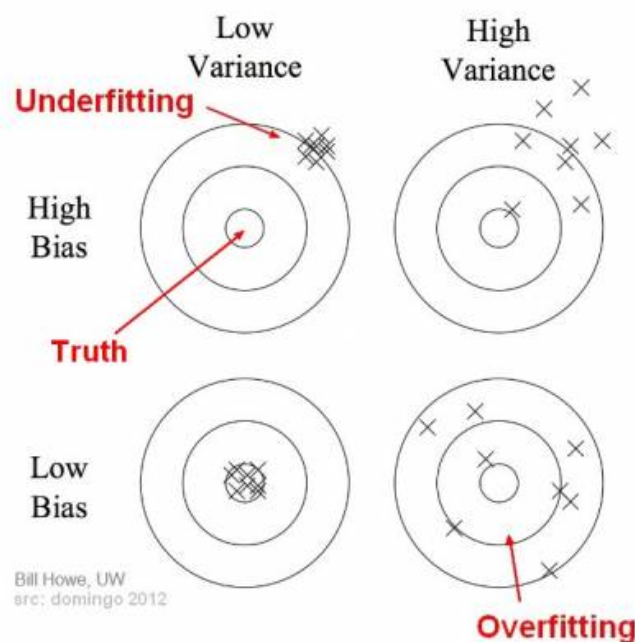
1. 등분산성: 잔차가 $y = 0$ 을 기준으로 Random하게 산포되어 있어야 한다.
2. 정규성: 정규분포 직선을 중심으로 분포해야 한다.
3. 독립성: 잔차가 $y = 0$ 을 기준으로 관리상하한(UCL, LCL)을 벗어나지 않고, Random하게 산포되어 있어야 한다.

Regularization

회귀 모델의 회귀계수가 가질 수 있는 값에 대한 제약조건을 부여하여 모델의 **variance**를 감소시키고 모델의 일반화(안정성) 성능을 높이는 기법

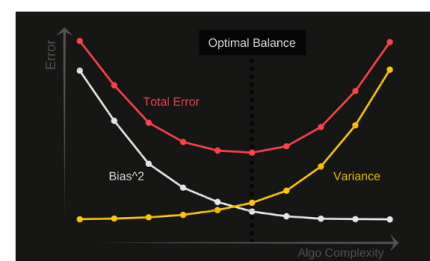
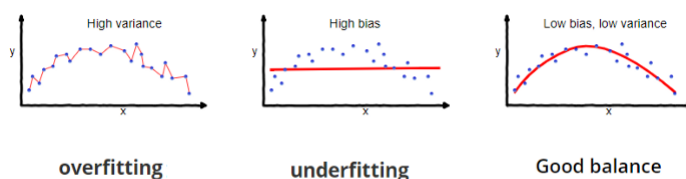
- 변수 선택(subset selection): 중요한 변수는 선택하고 중요하지 않은 변수는 제거
- 계수 축소(shrinkage): 덜 중요한 변수의 계수의 절대값을 낮춤

bias-variance tradeoff



bias-variance tradeoff

- bias(편향): 예측값의 평균과 실제값의 차이(편차). bias가 높으면 training data의 영향을 덜 받는다.
- variance(분산): 예측값과 예측치 평균과의 차이 제곱(산포). variance가 높으면 training data에 민감하게 영향을 받는다.



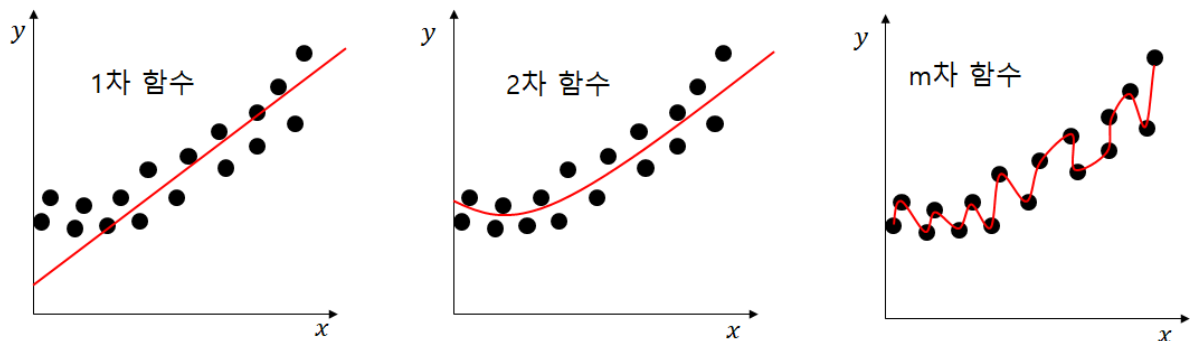
bias가 높으면 underfitting된 모델이며, 모델의 복잡도를 늘려 개선할 수 있다.

varaince가 높으면 overfitting된 모델이며, 모델의 복잡도를 낮춰 개선할 수 있다.

따라서 최적의 모델을 찾기 위해서는 bias와 variance가 동시에 낮아지는 최적점을 잘 찾아야 한다.

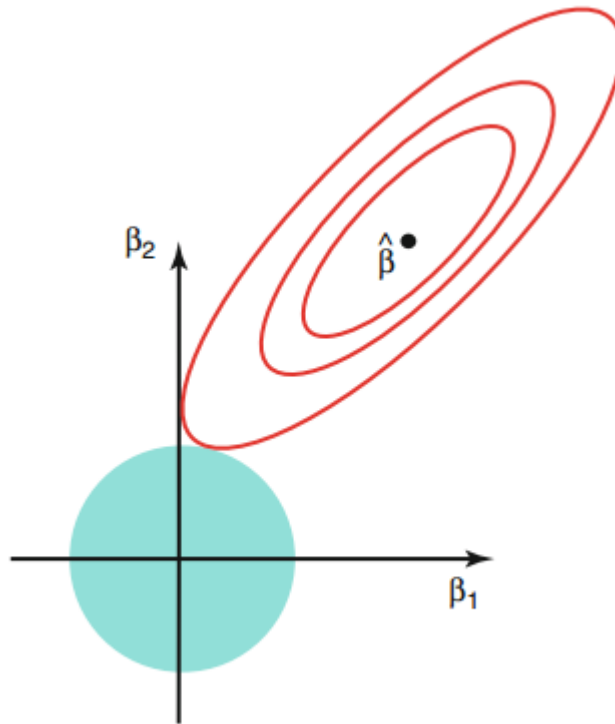
규제화의 종류

고차원의 비선형 회귀모델을 사용하면 오른쪽 그림과 같이 variance가 증가하며, overfitting이 발생하게 된다. 이러한 경우 특정 회귀계수가 가질 수 있는 값에 제약을 걸어, overfitting을 방지하는 방법이 Regularization.



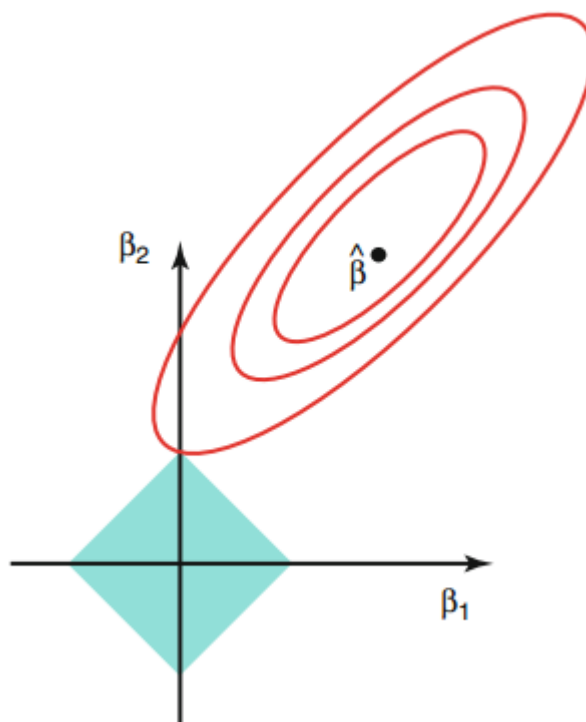
<https://velog.io/@cleansky/인사이드-머신러닝-대표적인-규제Regularization기법들-Ridge-Lasso-Elastic-Net>

Ridge Regression(L2 Regularization)



- 변수간 상관관계가 높아도 좋은 성능을 유지할 수 있다.
- 회귀계수가 0에 가까워지지만, 계수를 제거할 순 없다.

LASSO Regression(L1 Regularization)



- 변수 선택을 통해 높은 해석력을 가진다.
- 미분 불가능한 지점이 있어, Ridge보다 덜 매끄럽게 최적점에 도달한다.

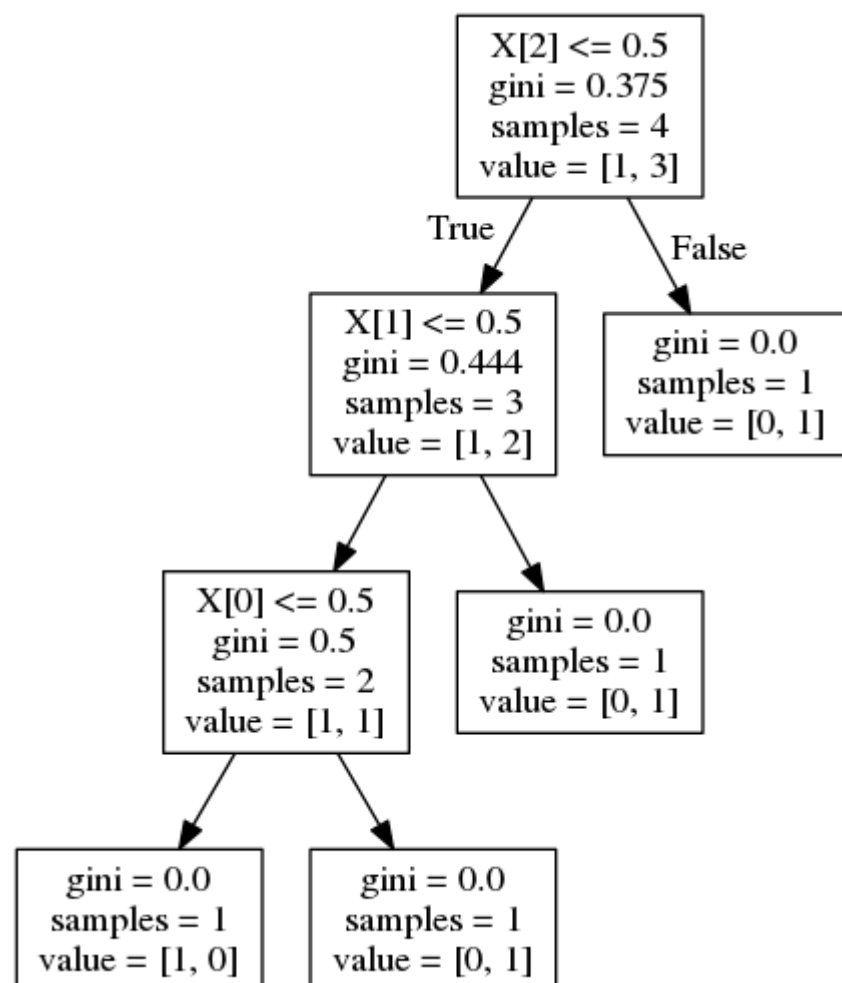
Elastic Net Regression

Ridge와 LASSO를 합친 상태.

- **grouping effect**: 변수간 상관관계가 큰 변수들에 유사한 가중치를 줌으로써 중요한 변수들은 똑같이 중요하게 취급하고 중요하지 않은 변수들은 모두 공평하게 중요하지 않게 취급한다.

Decision Tree

설명변수들의 규칙, 관계, 패턴 등으로 관심 대상인 목표변수를 분류하는 나무 구조의 모델을 만들고, 설명변수들의 관측값을 모델에 입력하여 목표변수를 분류/예측



장점

- 분석 결과가 직관적으로 제공되고 해석이 쉬움

단점

- 비연속적 분리로 인해 분리 경계점 근처에서 분리오류가 발생할 수 있음.

Metric

$$MSE = \frac{\sum_{i=1}^n (\hat{y} - y_i)^2}{n}$$

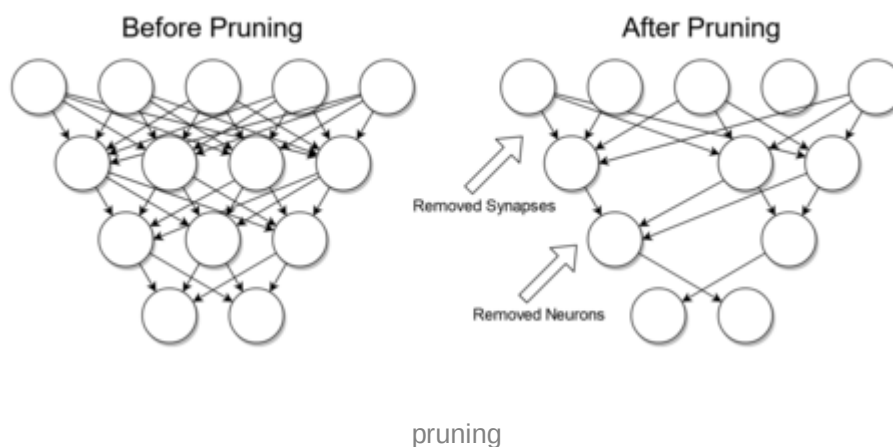
$$MAE = \frac{\sum_{i=1}^n |\hat{y} - y_i|}{n}$$

Overfitting

train data의 정보를 너무 많이 사용해서 **train data에 대해서는 정확하지만, test data나 새로운 데이터에는 일반화되기 어려워 정확도가 떨어지는 상태.**

Decision Tree에 대해서는 모델의 노드 수가 많아져서 train data에 적응하게 된다. 따라서 이를 막기 위해 pruning(가지치기)을 해야 한다.

모델 생성 이전에 분리 정지조건을 지정하는 **사전 가지치기**와, 생성된 모델 결과에서 가지치기를 수행하는 **사후 가지치기**가 있다. (Python에서는 사전 가지치기만 지원)



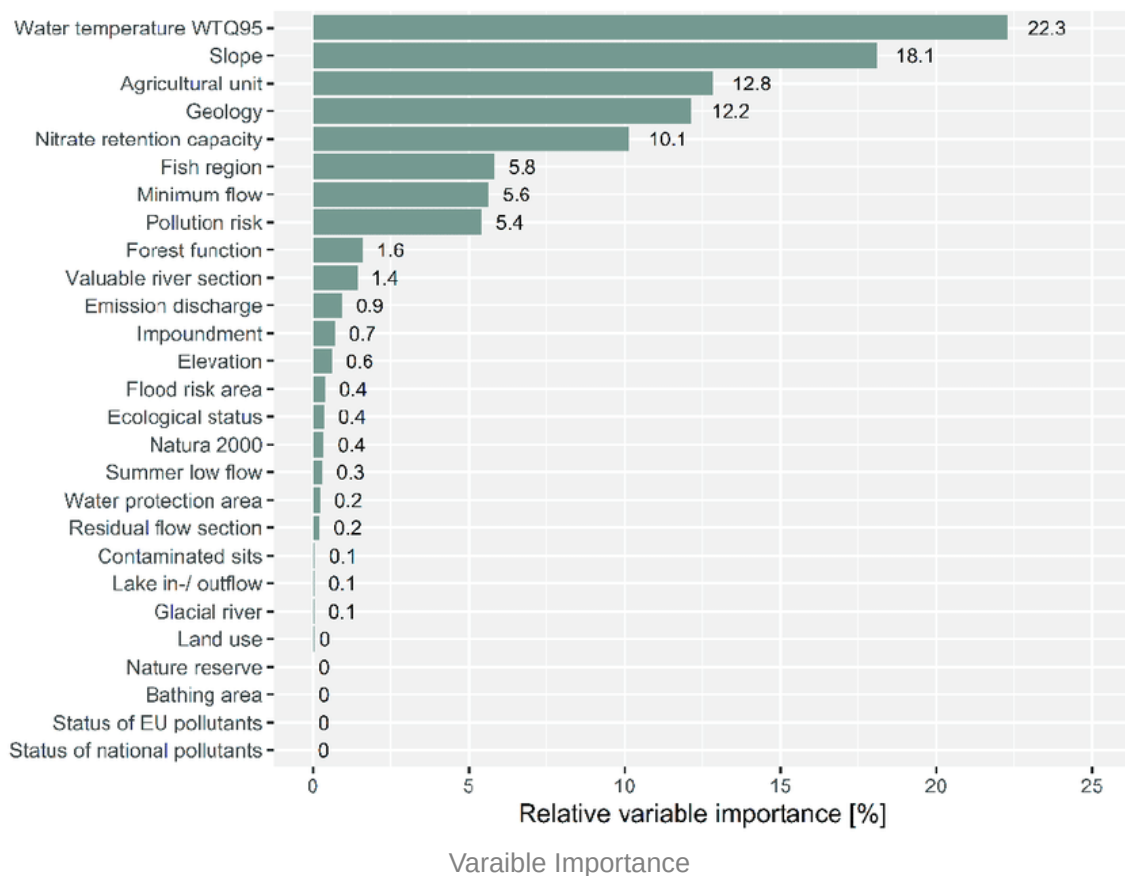
- Leaf Size: 리프 노드의 최소 데이터 수를 지정한다. **작으면 작을수록 더 세분화해서 분리하므로 overfitting.** `min_samples_leaf`

- Split Size: 분리 노드의 최소 데이터 수를 지정한다. **작으면 작을수록** 리프 노드가 많아 지므로 **overfitting**. `min_samples_split`
- Depth: 트리의 깊이를 지정한다. **높을수록 overfitting**.

Variable Importance

Tree 생성 과정에서 분리에 기여한 설명변수의 상대적 중요도(0~1)

상대적인 지표로, 값이 낮다고해서 해당 변수가 전혀 유용하지 않다는 뜻은 아니며 **다른 조건으로 나무를 생성하면 변수 중요도가 바뀔 수 있음**.



Ensemble

앙상블은 모델의 성능 개선을 목적으로 **다양한 알고리즘을 이용한 다수의 모델을 통합하는 방법**.

- 전체 데이터에서 **변수나 데이터를 샘플링**하여 만든 훈련용 데이터를 이용하여 **다수의 모델을 만들고 결합**

- Random Forest: 독립적으로 다수의 모델을 생성하고 결합하여 예측값 평균을 산출하는 모델링
- Gradient Boosting: 순차적으로 이전 모델의 파라미터를 조정하여 오차를 지속적으로 줄이는 모델링

장점

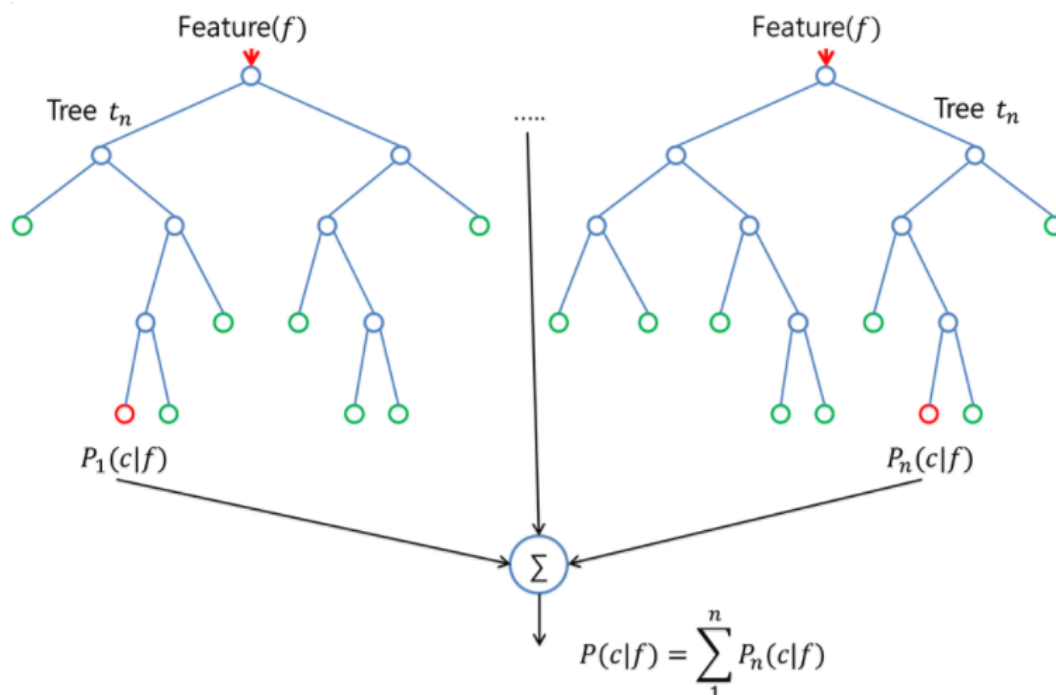
- 다수의 모델을 생성/결합하거나 이전 모델의 오차를 지속적으로 개선하여 **단일 모델보다 성능이 좋고 안정적**.
- 전체 분산을 감소시켜 오차를 줄이는 경향성을 보임.
 - 단일 모델보다 성능이 높아지므로, Bias와 Variance가 동시에 감소하게 됨.
- **이상치에 대한** 대응력이 높아져 **안정적인 결과**를 보임.
- 일반적으로 overfitting을 줄이는 경향을 보임.

단점

- 다수의 모델을 사용하기에 직관적이지 않고 해석의 어려움.

Random Forest

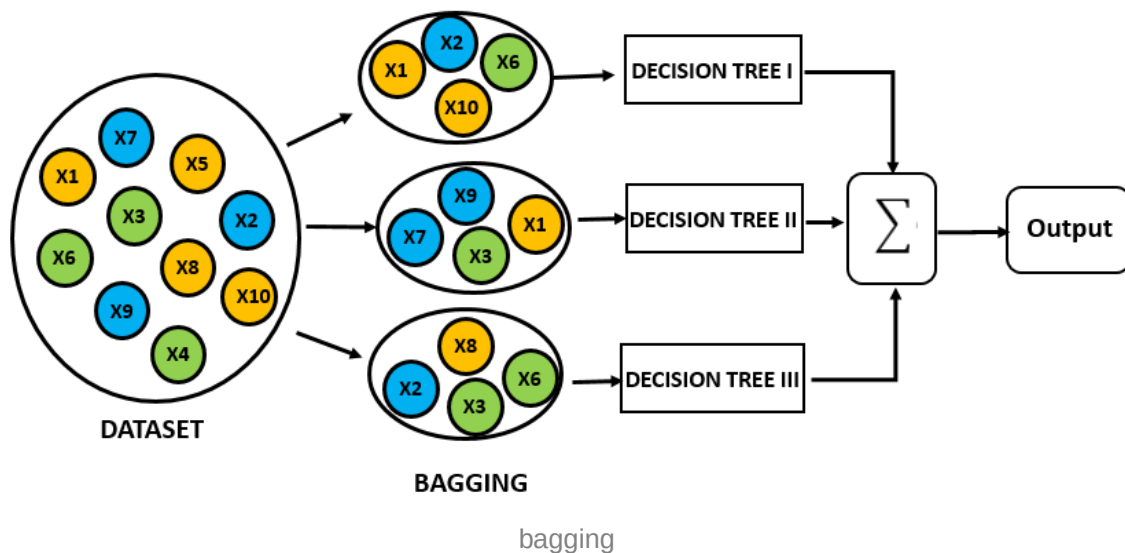
독립적으로 다수의 **Decision Tree**를 생성하고 그 결과를 **결합**한(bagging) 최종 모델을 생성함으로써 성능을 높이고 overfitting을 방지하여 안정적인 모델을 생성하는 앙상블 방법.



- 일반적으로 설명변수 및 표본을 무작위로 추출(복원추출, bagging)하여 모델 생성
- 생성된 다수 모델의 예측값을 평균하여 최종 예측값 산출. 예측의 일반화(안정성)가 향상됨
- overfitting 위험이 있는 하나의 Decision Tree보다 안정적이고 예측 성능이 높음
- OOB(Out Of Bag): 데이터를 복원추출하므로 선택되지 않는 데이터들이 존재함. 이를 test 데이터로 사용한다.

Bagging(Bootstrap + Aggregation)

training set의 부분집합에서 데이터를 복원추출하고, 그 데이터로 여러 트리를 학습시킨다. 이럴 경우 수학적으로 대략 36.7%는 선택되지 않는데, 이러한 데이터들은 OOB 평가에 이용된다.



장점

- 뛰어난 성능을 내며 파라미터 조정이 쉬움
- 데이터 Scale 변환이 불필요
- 큰 데이터에 적용 가능
- 일반화 및 성능이 좋은 모델

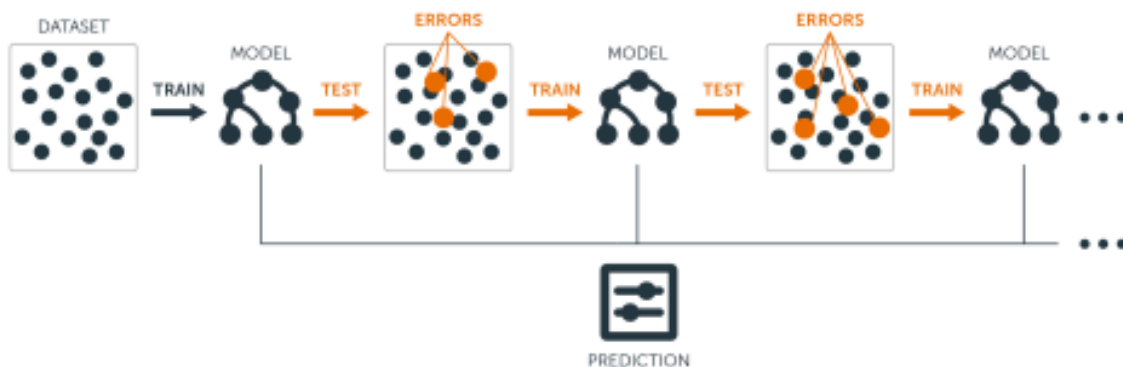
단점

- 개별 트리 분석이 어렵고 트리 분리가 복잡해지는 경향(overfitting)이 있음
- 차원이 크고 희소한 데이터는 성능 미흡(e.g. 텍스트 데이터)

- 선형 모델보다 **훈련 속도 느림**

Gradient Boosting

다수의 Decision Tree를 통합하여 강력한 모델을 만드는 앙상블 기법으로 **이전의 학습결과를 다음 학습에 전달**하여 이전의 오차(**잔여오차**)를 점진적으로 개선하는 기법(Boosting)



- 복잡하지 않은 모델에서 시작하여 이전 모델에 대한 학습을 통하여 모델 성능을 개선 (boosting 방식)
- 이전 모델의 (잔여)오차를 지속적으로 개선하는 모델로 다양한 분야에 적용되는 인기 있는 Machine Learning 방법
- hyperparameter 조정(특히 learning rate)에 민감하며 트리가 추가될수록 overfitting 이 될 수 있음

장점

- 뛰어난 성능을 내며 파라미터 조정이 쉬움
- 데이터 scale 변환이 불필요
- 대규모 데이터에 적용할 수 있고 속도가 빠름
- 일반화 및 성능이 좋은 모델

단점

- 개별 트리 분석이 어렵고 트리가 점진적으로 복잡해지는 경향이 있음. 이는 overfitting 을 야기함.
- 차원이 크고 희소한 데이터에 성능이 미흡함.(e.g. 텍스트 데이터)

- 파라미터(특히 **학습률**)에 민감하여 파라미터 조정에 따라 결과가 달라질 수 있음.

Residual error

이전 모델의 예측값과 실제값의 차이. 일반적인 모델과 달리 **이전 모델의 오차를 지속적으로 학습하고 개선하기** 때문에 “잔여오차” 라고 함

- overfitting되기 쉽기 때문에 정확도가 높다고 무조건 좋은 것은 아님. 일반화를 고려해야 함.
- overfitting을 방지하기 위한 조치(pruning)이 필요
- 정확도와 복잡도 사이의 적절한 균형이 중요.

모델 평가

일반적으로 Baseline 모델 대비 성능이 얼마나 우수한가를 평가한다.

좋은 모델이란, **기존 데이터도 잘 예측하면서 경험하지 않은 미래의 데이터에 대해서 좋은 예측 및 분류를 할 수 있는 모델**이다.

- 설명력: 분류/예측의 **성능이 우수한가?**
- 안정성/일반성: 동일 모집단의 **다른 데이터에 적용 시 안정적인 결과**가 나오는가?
- 효율성: **유사한 성능이라면 얼마나 적은 자원**(자료 및 설명변수 수)을 사용하는가?

Dimensionality Reduction

차원의 저주

데이터의 수가 차원의 수(변수의 수)보다 적을 때, 정보를 제대로 표현할 수 없음.

다중공선성이 높으면 예측모델의 성능이 저하되므로, 차원을 축소할 필요가 생긴다.


이 변수들을 단순히 제거하면 정보가 손실되므로, 차원을 줄이면서 정보의 손실을 최소화하는 방법이 PCA이다.

PCA(Principal Component Analysis)



주성분 분석(PCA)

PCA가 말하는 것: 데이터들을 정사영 시켜 차원을 낮춘다면, 어떤 벡터에 데이터들을 정사영 시켜야 원래의 데이터 구조를 제일 잘 유지할 수 있을까? ※ 본 article에서는 열벡터(column

 <https://angeloyeo.github.io/2019/07/27/PCA.html>



<https://www.youtube.com/watch?v=bEX6WPMiLvo>

Time Series Analysis

정상성

시계열 데이터의 확률적 성질(평균, 분산)이 시간의 변화에도 변하지 않는 (일정한) 상태 또는 성질

- 정상성을 만족하는 경우, 특정한 시점 t_i 와 t_{i-k} 의 데이터는 서로 독립적

비정상 시계열 데이터의 정상화

비정상 시계열 데이터를 변환을 통해 정상성을 확보하는 과정

차분을 이용하여 추세를 제거하거나 정규성을 만족시킬 수 있는 변환 적용 후 분석(ARMA) 필요

AR(Auto Regressive, 자기회귀) 모델

p 시점 전 데이터(설명변수)가 현재 시점 데이터(목표변수)에 직접적으로 영향을 주는 회귀 모델. 설명변수가 목표변수의 과거 데이터 값.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t$$

- ACF(Auto Correlation Function, 자기상관 함수): 현재와 p시점 전의 자기 데이터 간의 상관관계.
- PACF(Partial ACF, 부분 자기상관 함수): 모든 다른 시점 데이터들의 영향력을 배제하고 Z_t 와 Z_{t-k} 의 두 시점 데이터간의 상관관계
- AR 모델에서 일반적으로 ACF는 연속적으로 감소하며, PACF는 절단면을 가짐.

MA(Moving Average, 이동평균) 모델

일정 p구간의 데이터 평균이 현재 시점의 자신의 데이터에 영향을 주는 추세 분석 모델. 설명변수는 특정 구간에서의 자신의 데이터의 평균

$$Z_t = a_t - \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_p a_{t-p}$$

Classification Model

Logistic Classification

범주형 목표변수와 다양한 설명변수 관계를 모형화하여 목표변수를 분석하고 분류하는 통계적 기법

목표변수에 영향을 미치는 설명변수를 찾고 목표변수 발생 확률을 예측하고 분류

Logit function(검색해볼것)

logit 변환을 통하여 선형화하는 함수

Odds / Odds Ratio

Odds: 임의의 사건이 요인에 의해 발생하지 않을 확률 대비 발생할 확률

Odds Ratio: Odds 간의 비율에 따른 사건 발생 확률 비교

질병발생 요인노출	예	아니오	합계
예	a	b	a+b
아니오	c	d	c+d
합계	a+c	b+d	n=a+b+c+d

- 요인: 코로나 확진자와 식사를 함.
- 사건: 코로나에 확진 됨.

Likelihood(MLE)

Maximum Likelihood Estimation(MLE)

주어진 x 에 대해 가능도를 가장 크게 해주는 모수 θ 를 찾는 방법으로, 회귀 계수를 추정한다.