Data Science

분석 결과의 해석

분석 모형의 평가 및 개선

강사 양석환

분석 모형의 평가 지표와 진단

분석 모형 평가 지표

- **지도학습: 분류모델 평가 지표**
  - **분석모델의 답과 실제 답과의 관계를 오차행렬(혼동행렬, Confusion Matrix)을 통해 평가함**
  - **평가 지표에는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 Score 등이 있음**

- **오차행렬(=혼동행렬, Confusion Matrix)**
  - 훈련을 통한 예측 성능을 측정하기 위해 예측값과 실제값을 비교하기 위한 표

		사실	
		참 (Positive)	거짓 (Negative)
실험 결과	참 (Positive)	TP (True Positive)	FP (False Positive)
	거짓 (Negative)	FN (False Negative)	TN (True Negative)

TP (True Positive): 실제 True인 답을 True라고 예측(정답)
FP (False Positive): 실제 False인 답을 True라고 예측(오답)
FN (False Negative): 실제 True인 답을 False라고 예측(오답)
TN (True Negative): 실제 False인 답을 False라고 예측(정답)

분석 모형 평가 지표

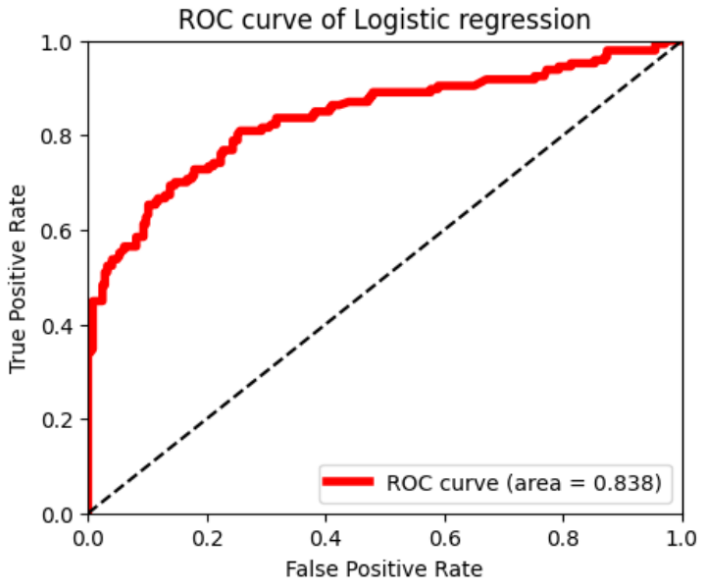
- **지도학습: 분류모델 평가 지표**
  - **정확도 (Accuracy): 실제 데이터와 예측 데이터를 비교하여 같은지 판단함**
    - $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$
  - **정밀도 (Precision): Positive로 예측한 대상 중에 실제와 예측값이 일치하는 비율**
    - $Precision = \frac{TP}{TP+FP}$
  - **재현율 (Recall): 실제 Positive인 대상 중에 실제와 예측값이 일치하는 비율**
    - $Recall = \frac{TP}{TP+FN}$
  - **F1 Score: 정밀도와 재현율을 결합한 조화평균 지표로 값이 클수록 모형이 정확하다고 판단할 수 있음**
    - $F1\ Score = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} = 2 \times \frac{Precision \cdot Recall}{Precision+Recall}$

분석 모형 평가 지표

- **지도학습: 분류모델 평가 지표**
  - **ROC (Receiver Operating Characteristic) 곡선**
    - FPR(False Positive Rate)이 변할 때, 민감도인 TPR(True Positive Rate)이 어떻게 변화하는지를 나타내는 곡선
      - 예: 임계값이 1이면 확률이 1일 때 True로 예측하므로 FP가 0이 되고 FPR도 0이 됨
    - 임계값을 1~0 범주 이내 값으로 조정하면서 FPR에 따른 TPR을 계산하면서 곡선을 그림
    - TPR 값과 FPR 값이 0.5인 기본 모델 위에 ROC가 위치할 경우 성능이 기본 모델보다 나음을 의미함
    - $FPR = \frac{FP}{FP+TN}, TPR = \frac{TP}{TP+FN}$

분석 모형 평가 지표

- **지도학습: 분류모델 평가 지표**
  - **AUC (Area Under Curve)**
    - 평가 모델의 ROC 곡선의 하단 면적을 의미하며, 랜덤일 때 0.5 값으로, ROC 곡선이 직선에서 멀어질수록 성능이 더 뛰어난 것으로 해석함



분석 모형 평가 지표

- **지도학습: 회귀모델 평가 지표**
  - **회귀의 평가를 위한 지표는 실제값과 회귀 예측값의 차이를 기반으로 성능지표를 수립, 활용함**
  - **SSE(Sum Squared Error)**
    - 실제값과 예측값의 차이를 제곱하여 더한 값
    - $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - **MSE(Mean Squared Error)**
    - 실제값과 예측값의 차이의 제곱에 대한 평균을 취한 값. 평균제곱 오차
    - $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

분석 모형 평가 지표

- **지도학습: 회귀모델 평가 지표**
  - **RMSE(Root Mean Squared Error)**
    - MSE에 루트를 취한 값. 평균제곱근 오차
    - $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$
  - **MAE(Mean Absolute Error)**
    - 실제값과 예측값의 차이의 절대값을 합한 후 평균을 취한 값
    - $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$

분석 모형 평가 지표

- **지도학습: 회귀모델 평가 지표**
  - **결정계수 R^2**
    - 회귀모형이 실제값에 대해 얼마나 잘 적합하는지에 대한 비율
    - $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}$
  - **Adjusted R^2 (수정된 결정계수)**
    - 다변량 회귀분석에서 독립변수가 많아질수록 결정계수가 높아지는데 이를 보완한 결정계수
    - 표본크기(n)와 독립변수의 개수(p)를 추가적으로 고려하여 분모에 위치시킴으로써 결정계수 값의 증가도를 보정
      - $R_a^2 = 1 - \frac{(n-1)(1-R^2)}{n-p-1} = 1 - \frac{(n-1)\left(\frac{SSE}{SST}\right)}{n-p-1} = 1 - (n-1)\frac{MSE}{SST}$

분석 모형 평가 지표

- **지도학습: 회귀모델 평가 지표**
  - **MSPE(Mean Square Percentage Error)**
    - MSE를 퍼센트로 변환한 값
    - $MSPE = \frac{100\%}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{y_i}}{y_i} \right)^2$
  - **MAPE(Mean Absolute Percentage Error)**
    - MAE를 퍼센트로 변환한 값
    - $MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y_i}}{y_i} \right|$

분석 모형 평가 지표

- **지도학습: 회귀모델 평가 지표**
  - **RMSLE(Root Mean Squared Logarithmic Error)**
    - RMSE에 로그를 취한 값으로 이상치에 덜 민감함
    - $RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y_i} + 1))^2}$
  - **AIC(Akaike Information Criterion)**
    - 최대 우도에 독립변수의 개수에 대한 손실(Penalty)분을 반영하는 목적으로 모형과 데이터의 확률 분포 차이를 측정하는 것
    - AIC의 값이 낮을수록 모형의 적합도가 높아짐
    - $AIC = -2logL + 2K$

분석 모형 평가 지표

- 지도학습: 회귀모델 평가 지표
  - BIC(Bayes Information Criterion)
    - AIC와 동일한 목적을 가지며 주어진 데이터에서 모형의 우도(likelihood)를 측정하기 위한 값에서 유도된 지표
    - 변수 개수가 많을수록 AIC보다 더욱 강하게 페널티를 가하는 성격을 가짐
    - $BIC = -2logL + Klogn$

분석 모형 평가 지표

- 비지도학습: 군집분석 평가 지표
  - 비지도학습은 지도학습과 달리 실측자료에 라벨링이 없으므로 모델에 대한 성능평가가 어려움
  - 군집분석에 한해 다음과 같은 성능평가 지표를 사용함

  - 실루엣 계수(Silhouette)
    - $a(i)$ 는 i 번째 개체와 같은 군집에 속한 요소들 간 거리들의 평균
    - $b(i)$ 는 i 번째 개체가 속한 군집과 가장 가까운 이웃 군집을 선택하여 거리를 계산한 값
    - $a(i)$ 가 0이면 하나의 군집에 모든 개체들이 붙어있는 경우
    - 실루엣 계수가 0.5보다 클 경우, 적절한 군집 모델로 볼 수 있음
    - $s(i) = \frac{b(i)-a(i)}{max\ a(i), b(i)}$

분석 모형 평가 지표

- **비지도학습: 군집분석 평가 지표**
  - **Dunn Index**
    - 군집 간 거리의 최소값을 분자, 군집 내 요소 간 거리의 최대값을 분모로 하는 지표
    - 군집 간 거리는 멀수록, 군집 내 분산은 작을수록 좋은 군집화 결과임
    - Dunn Index 값은 클수록 좋음
    - $I(C) = \frac{min_{i \neq j} d_c(C_i, C_j)}{max_{1 \leq l \leq k} \Delta(C_l)}$

분석 모형 진단

- 정규성 가정
  - 통계적 검정, 회귀분석 등 분석을 진행하기 전에 데이터가 정규분포를 따르는지 검정하는 것
  - 데이터 자체의 정규성을 확인하는 과정
  - 데이터셋이 정규분포를 따른다는 귀무가설(H0)을 기각하고 대립가설이 채택된다면($p<0.01$ 또는 $p=0.05$) 해당 데이터셋은 정규분포를 따르지 않는 것으로 증명됨

분석 모형 진단

- 정규성 가정
  - 중심 극한 정리(Central Limit Theorem)
    - 동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 이론
    - 이때 표본분포의 평균은 모집단의 모평균과 동일함
    - 표준 편차는 모집단의 모 표준편차를 표본 크기의 제곱근으로 나눈 것과 같음
  - 정규성 검정 종류
    - Shapiro-Wilks Test : 표본 수(n)가 2,000개 미만인 데이터셋에 적합함
    - Kolmogorov-Smirnov Test : 표본 수(n)가 2,000개 초과인 데이터셋에 적합함
    - Quantile-Quantile Plot(Graphic Test)
      - 데이터셋이 정규분포를 따르는지 판단하는 시각적 분석 방법
      - 표본 수(n)가 소규모일 경우에 적합함

분석 모형 진단

- 잔차 진단
  - 회귀분석에서 독립변수와 종속변수의 관계를 결정하는 최적의 회귀선은 실측치와 예측치의 차이인 잔차를 가장 작게 해주는 선으로, 이때 잔차의 합은 0임
  - 잔차는 추세, 특정 패턴을 가지고 있지 않음

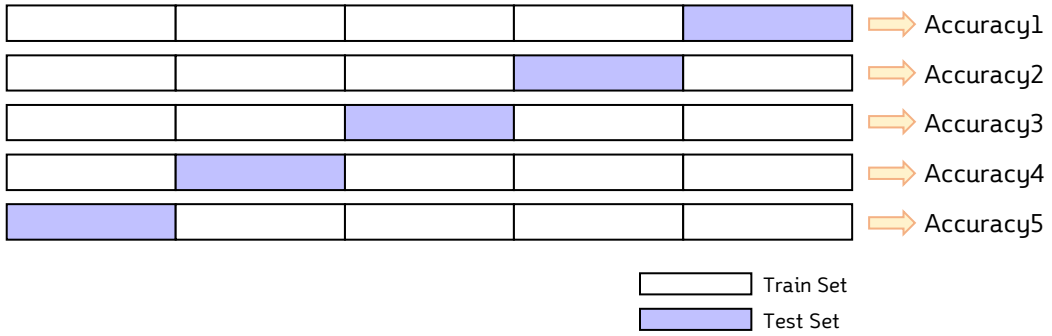
분석 모형 진단

- 잔차 진단
  - 잔차의 정규성 진단
    - 신뢰구간 추정과 가설검증을 정확하게 하기 위하여 Q-Q Plot과 같은 시각화 도표를 통해 정규분포와 잔차의 분포를 비교함
  - 잔차의 등분산성 진단
    - 잔차의 분산이 특정 패턴을 가지지 않고 순서와 무관하게 일정한지 등분산성을 진단함
  - 잔차의 독립성 진단
    - 잔차의 독립성: 자기상관(Auto Correlation)의 여부를 판단하는 것
    - 시점 순서대로 그래프를 그리거나 Durbin-Watson 검정을 적용했을때 패턴이 발견되지 않는다면 독립성을 충족한다고 할 수 있음
    - 만약 독립성이 위배된다면 시계열 분석을 통해 회귀분석을 진행해야 함

분석 모형의 검증

분석 모형 검증

- k-폴드 교차검증(k-fold Cross Validation)
  - 고정된 훈련 데이터셋과 테스트검증 데이터셋으로 평가를 하여 반복적으로 튜닝하게 될 경우, 테스트 데이터셋에 과적합되어버리는 결과가 생길 수 있는데, 이를 방지하고자 나온 방법이 교차검증 기법임
  - k-폴드 교차검증 기법은 전체 데이터셋을 k개의 서브셋으로 분리하여 그 중에서 k-1개를 훈련데이터로 사용하고 1개의 서브셋은 테스트 데이터로 사용함. 테스트셋을 중복없이 병행 진행한 후 평균을 내어 최종 모델의 성능을 평가함



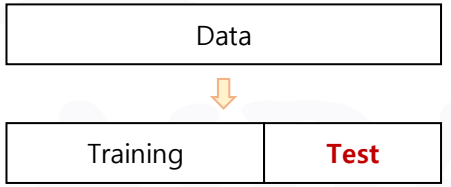
k-폴드 교차검증 예시

교차검증은 모든 데이터셋을 평가에 활용하여 과적합을 방지할 수 있으나 반복 횟수 증가에 따른 모델 훈련과 평가·검증 시간이 오래 걸릴 수 있음

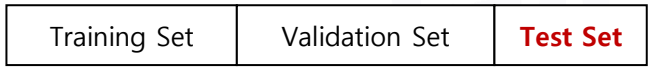
분석 모형 검증

- **홀드아웃 기법(Holdout Method)**
  - **일반적인 모델의 훈련방식을 적용할 경우, 동일한 테스트데이터를 계속 사용한다면 훈련데이터화 하여 과적합이 발생하게 됨**
  - **홀드아웃 기법**
    - **훈련데이터, 검증데이터, 테스트데이터를 일정 비율로 지정한 뒤,**
    - **먼저 훈련데이터로 학습하되**
    - **훈련데이터 내에서 일정 부분 검증데이터를 두어 학습과정에서 모델 성능을 높이는 검증을 진행**
    - **최종적으로는 테스트 데이터를 통해 성능을 평가할 수 있음**
  - **데이터셋의 크기가 작을수록 데이터를 나누는 방식에 따라 모델 성능 추정에 민감한 영향을 미칠 수 있음**

일반적인 모델의 훈련 방식
훈련 데이터셋과 테스트검증 데이터셋으로 구분한 뒤, 훈련데이터로 모델을 학습하고, 학습된 모델을 선택하여 테스트데이터로 성능을 확인, 증가시키는 방법



홀드아웃 교차검증(기본)



과적합 방지를 위한 홀드아웃 교차검증

분석 모형 검증

- 적합도 검정(Goodness-of-fit)
  - 일반적인 적합도 검정 방법으로 정규성 검정이 있으며
  - 모집단의 분포를 정규분포로 가정하는 분석기법(t-test, ANOVA, 회귀분석)이 적용될 경우 데이터가 정규분포를 따르는지 확인할 때 사용됨
  - 그 외에도 카이제곱 검정, 콜모고로프 스미르노프(Kolmogorov-Smirnov) 검정이 있음

분석 모형 검증

- 적합도 검정(Goodness-of-fit)
  - 카이제곱 검정
    - 기댓값과 관측값을 이용한 방법
    - k개의 범주로 나뉘어진 관측치들과 이와 동일한 범주의 가정된 분포 사이의 적합도를 검정함
    - 범주형 값 k가 나와야 할 횟수의 기댓값 와 실제 나온 횟수 의 차이를 이용하여 검정통계량을 계산함
    - $\sum_{k=1}^{K} \frac{(x_k - m_k)^2}{m_k}$

분석 모형 검증

- **적합도 검정(Goodness-of-fit)**
  - **콜모고로프 스미르노프 검정(Kolmogorov-Smirnov Test, K-S Test)**
    - 관측된 표본분포와 가정된 분호 사이의 적합도를 검사하는 누적분포함수의 차이를 이용한 검정법
    - 연속형 데이터에도 적용 가능
    - 관측된 자료의 크기를 나열, 관측치들의 누적확률을 구하여 가정된 분포의 누적확률과 비교하는 순서로 진행함

분석 모형의 개선

분석 모형 개선

- 과대적합 방지
  - 훈련 시에는 높은 성능을 보이지만 테스트데이터에 대해서는 낮은 성능을 보여주는 과대적합(Over Fitting)을 방지하고, 일반화된 모델을 생성하기 위해 다음과 같은 방향을 제시함

  - 모델의 낮은 복잡도
    - 훈련데이터를 더 많이 획득할 수 없다면 정규화, 드롭아웃 등을 활용하여 적절한 복잡도를 가진 모델을 자동으로 탐색함
    - 학습을 하면서 지속적으로 바뀌는 가중치 매개변수가 아니라 상수값인 하이퍼파라미터(학습률, 각 층의 뉴런 수 등)는 과대적합의 위험을 줄이기 위해 제약을 가하는 규제의 양을 결정할 수 있는 인수로 큰 값을 지정할수록 복잡도가 낮은 모델을 얻게 됨

분석 모형 개선

- **과대적합 방지**
  - **가중치 감소**
    - 학습과정에서 큰 가중치에 대해서는 큰 패널티(규제)를 부과하여 가중치의 절대값을 가능한 한 작게 만듦
    - 규제란 과대적합이 되지 않도록 모델을 강제로 제한하는 것을 의미하며, L1, L2 규제가 있음
    - L2 규제
      - $L2: \|w\|_2^2 = \sum_{j=1}^{m} w_j^2$
      - 손실함수에 가중치에 대한 L2 Norm의 제곱을 더한 패널티를 부여하여 가중치 값을 비용함수 모델에 비해 작게 만들어냄
      - 손실함수가 최소가 되는 가중치 값인 중심점을 찾아 큰 가중치를 제한하는데, 이때 람다로 규제의 강도를 크게 하면 가중치는 0에 가까워짐
      - 회귀 모델에 L2 규제를 적용한 것이 릿지(Ridge) 모델임

분석 모형 개선

- **과대적합 방지**
  - **가중치 감소**
    - **L1 규제**
      - $L2: \|w\|_1 = \sum_{j=1}^{m} |w_j|$
      - L2 규제의 가중치 제곱을 절대값으로 바꾸는 개념
      - 손실함수에 가중치의 절대값인 L1 Norm을 추가, 적용함으로써 희소한 특성벡터가 되어 대부분의 특성 가중치를 0으로 만듦
      - 회귀 모델에 L1규제를 적용한 것이 라쏘(Lasso) 모델임
  - **편향-분산 트레이드오프**
    - 과대적합과 과소적합 사이의 적절한 편향-분산 트레이드오프(절충점)을 찾음

분석 모형 개선

- **매개변수의 최적화**
  - **확률적 경사 하강법(Stochastic Gradient Descent, SGD)**
    - 최적의 매개변수 값을 찾기 위해 매개변수에 대한 손실함수의 기울기를 이용함
    - 손실함수의 기울기를 따라 조금씩 내려가다 최종적으로 손실함수가 가장 작은 지점에 도달하도록 하는 알고리즘
    - 배치 경사 하강법과 비교하면 랜덤으로 선택한 하나의 데이터로만 계산하는 단순하고 명확한 구조가 임
    - 다만 최소값인 (0, 0)까지 지그재그로 이동, 매개변수가 방향에 따라 다른 기울기를 가지는 비등방성 함수인 경우에는 비효율적인 움직임을 보임
      - $W \leftarrow w - \eta \frac{\partial L}{\partial W}$
    - 알고리즘 수식은 갱신할 가중치 매개변수인 W, dL/dw, 매개변수에 대한 손실함수의 기울기와 학습률(η)로 설명

분석 모형 개선

- 매개변수의 최적화
  - 모멘텀(Momentum)
    - 모멘텀은 운동량을 뜻함
    - 미분계수가 0인 지점에서 더 이상 이동하지 않는 한계점을 가진 확률적 경사 하강법에 속도 개념인 기울기 방향으로 힘을 받으면 물체가 가속되는 관성 물리법칙을 적용함
      - $V \leftarrow \alpha V - \eta \frac{\partial L}{\partial W}, W \leftarrow W + v$
    - v(속도)항에 기울기 값이 누적되고, 누적된 값이 가중치 갱신에 영향을 주면서 이 기울기 값으로 인해 빠른 최적점 수렴이 가능함

분석 모형 개선

- 매개변수의 최적화
  - AdaGrad(Adaptive Gradient)
  - - 개별 매개변수에 적응적으로 학습률을 조정하면서 학습을 진행하는 알고리즘
  - - 첫 부분에서는 크게 학습하다가 최적점에 가까울수록 학습률을 점차 줄여가며 조금씩 작게 학습시킴
    - $h \leftarrow h + \frac{\partial L}{\partial W} \odot \frac{\partial L}{\partial W}$, $W \leftarrow W - \eta \frac{1}{\sqrt{h}} \frac{\partial L}{\partial W}$
  - - 새로나온 변수 h는 기존 기울기 값을 제곱하여 계속 더해줌. 매개변수를 갱신할 때 h의 제곱근을 나눠줌으로써 모든 가중치가 이전에 갱신되었던 크기에 맞게 학습률이 조정됨

분석 모형 개선

- **매개변수의 최적화**
  - **Adam(Adaptive Moment Estimation)**
    - 모멘텀과 AdaGrad를 결합한 방법론
    - 학습률, 일차 모멘텀 계수, 이차 모멘텀 계수의 3가지 초매개변수들을 설정함
    - 최적점 탐색 경로의 전체적인 경향은 모멘텀과 같이 공이 굴러가는 듯 하다가 AdaGrad로 인해 갱신 강도가 조정되어 좌우 흔들림이 줄어들게 됨

분석 모형 개선

- **매개변수의 최적화**
  - **초매개변수(Hyper Parameter) 최적화**
    - 초매개변수(Hyper Parameter)란 사람이 직접 설정해주어야 하는 매개변수를 말함
    - 뉴런의 수, 배치(batch)의 크기, 학습률(Learning Rate), 은닉층 개수 등이 있음
    - 학습률(Learning Rate)
      - 기울기(Gradient) 방향으로 얼마나 빠르게 이동할지를 결정함
      - 학습률이 작으면 학습시간이 길어지고 학습률이 커지면 발산하여 학습이 제대로 이루어지지 않을 수 있음

분석 모형 개선

- 매개변수의 최적화
  - 초매개변수(Hyper Parameter) 최적화
    - 미니배치(Mini-Batch) 크기
      - 전체 훈련 데이터셋을 신경망에 넣게 되면 리소스가 비효율적으로 사용되고 시간이 오래 걸리게 되므로 배치 개념을 사용해서 해결함
      - 미니배치는 전체 학습 데이터를 주어진 배치 크기로 나눈 것
      - 미니배치 크기가 큰 경우, 병렬연산 구조를 사용할 때 효과적일 수 있음
      - 미니배치 크기가 작은 경우, 더 많은 가중치 업데이트를 수행할 수 있음
    - 훈련 반복(Epoch) 횟수
      - 전체 훈련 데이터셋이 신경망을 통과한 횟수
      - 1 Epoch은 1회 학습만으로 통과했다는 의미가 됨
      - 학습의 조기 종료를 결정하는 변수로도 사용됨

분석 모형 개선

- **매개변수의 최적화**
  - **초매개변수(Hyper Parameter) 최적화**
    - 이터레이션(Oteration)
      - 하나의 미니배치를 학습할 때, 1 iteration으로 1회의 매개변수(파라미터) 업데이트가 진행됨
      - 미니배치 개수와 이터레이션의 개수는 동일함
    - 은닉층(Hidden Layer) 개수
      - 은닉층 수가 많아질수록 특정 훈련데이터에 더 최적화시킬 수 있음
      - 모든 은닉층들의 뉴런 개수를 동일하게 유지하는 것이 은닉층 개수에 뉴런의 개수를 가변적으로 하는 것보다 효과적임
      - 첫 번째 은닉층에 있는 뉴런의 개수가 입력층에 있는 뉴런의 개수보다 큰 것이 효과적인 경우가 많음

분석 모형 개선

- **매개변수의 최적화**
  - **분석 모형 융합**
    - **앙상블 학습**
      - **주어진 자료를 이용하여 여러 가지 분석 예측모형들을 만들고 해당 예측 모형들을 결합하여 최종적인 하나의 예측모형을 만드는 방법**
      - **치우침이 있는 여러 모형의 평균을 취할 사, 균형적인 결과(평균)를 얻을 수 있음**
      - **여러 모형의 분석 결과를 결합하면 변동성 및 과적합의 여지가 줄어듦**
      - **배깅**
        - 복원 추출 방법으로 데이터를 샘플링, 모델링한 후, 전체 결합하여 결과를 평균하는 기법

분석 모형 개선

- **매개변수의 최적화**
  - **분석 모형 융합**
    - **앙상블 학습**
      - **부스팅**
        - 순서대로 모델들을 진행하는 방법
        - 이전 분류기의 학습 결과에 따라 다음 분류기의 학습 데이터의 샘플 가중치(잘못 분류한 데이터와 이용하지 않은 데이터에 대한 가중치)를 조정해 학습을 진행함
      - **랜덤 포레스트**
        - 배깅을 적용한 의사결정나무로 다수의 의사결정나무를 만들고
        - 각 나무들은 학습 데이터셋의 일부분을 추출해서 학습함
        - 나무를 구성하는 변수 역시 전체 변수들의 부분집합으로 선택됨

분석 모형 개선

- 매개변수의 최적화
  - 분석 모형 융합
    - 결합분석 모형
      - 두 종류 이상의 결과변수를 동시에 분석할 수 있는 방법
      - 결과 변수 간의 유의성, 관련성을 설명할 수 있음

분석 모형 개선

- 매개변수의 최적화
  - 최종 모형 선정
    - 회귀모형에 대한 주요 성능 평가 지표
      - SSE, R^2, MAE, MAPE
    - 분류모형에 대한 주요 성능 평가 지표
      - 특이도(Specificity): 음성 중 맞춘 음성의 수
      - 정밀도(Precision): 양성 판정 수 중 실제 양성 수. 해당 클래스 예측 샘플 중 실제 속한 샘플 수의 비율
      - 재현율(Recall): 통계용어로 민감도(Sensitivity). 전체 양성 수에서 검출 양성 수(양성 중 맞춘 양성의 수). 실제 속한 샘플 중 특정 클래스에 속한다고 예측한 표본 수의 비율
      - 정확도(Accuracy): 전체 수 중에서 양성과 음성을 맞춘 수. 전체 샘플 중에서 맞게 예측한 샘플 수의 비율

분석 모형 개선

- **매개변수의 최적화**
  - **최종 모형 선정**
    - **비지도학습 모형에 대한 주요 성능 평가 지표**
      - **군집분석**
        - 군집 타당성 지표(Clustering Validity Index)로 군집 간 거리, 군집의 지름, 군집의 분산 등을 고려함
      - **연관분석**
        - 연관규칙에서 지지도와 신뢰도가 모두 최소한도보다 높은 것으로 평가함
        - 일반적으로 최소 지지도를 정한 뒤, 이에 대한 이하 값들은 버리고 그 중에서 신뢰도가 어느정도 높은 결과들을 가져옴