교 육 일 지

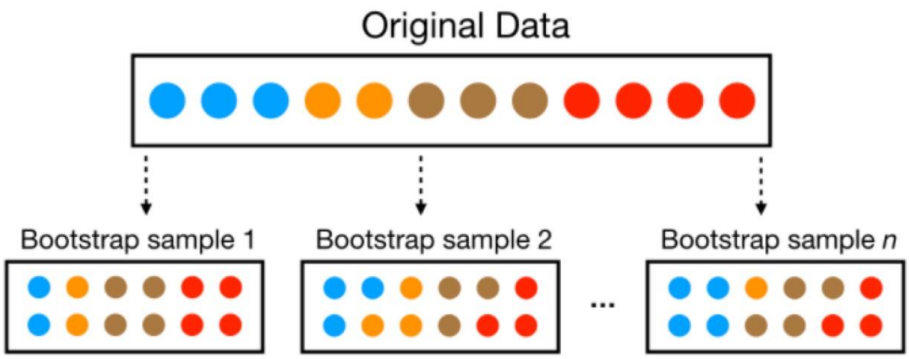
교육 제목	**머신러닝 모델링 프로세스**
교육 일시	2021.10.14
교육 장소	YGL-C6
교육 내용	

1. Bootstrapping

   통계학에서는 test를 하거나 metric을 계산하기 전에 Random Sampling을 적용하는 방법을 일컫는다.

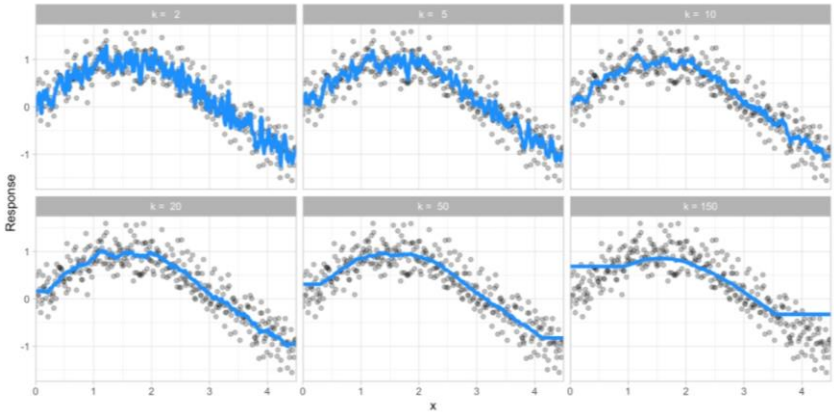
   머신러닝에서의 Bootstrapping은 랜덤 샘플링을 통해 training Data를 늘리는 방법이다.

   Training_set 내의 데이터 분포가 고르지 않는 경우 데이터가 적은 범주의 error는 무시되는 방향으로 학습도기 쉽다. 이때 bootstrapping을 사용해 적은 범주의 데이터를 늘려 Training set을 구성하면 된다.



2. HyperParameter Tuning(초매개변수 조절)

   - 초매개변수는 학습 과정을 제어하는데 사용되는 변수를 의미한다.
   - 초매개변수는 모델 학습과정이 아닌 개발자에 의해 지정된다.

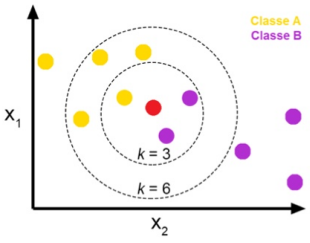


초 매개변수 k 에 따라 모델이 달라지는 것을 확인 할 수 있다.

3.

K-nearest neighbors classification

- 지도학습으로서 분류(Classification) 나 회귀(Regression)에 사용되는 비모수적 방법
- 파라메터 학습을 위한 훈련과정이 없으나 훈련집합은 필요
- 각 데이터 간에 거리를 계산하기 위한 거리척도가 필요
- 초매개변수 k를 설정해야 함
- 거리에 대한 가중치



https://bkshin.tistory.com

기하학적 거리 (Geometric distance measures)

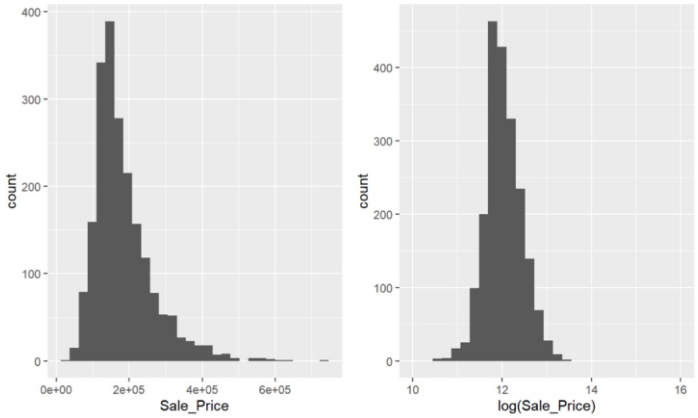
- Euclidean: $d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}, \quad \vec{x},\vec{y} \in p$
- Manhattan: $d(\vec{x},\vec{y}) = \sum_{i=1}^{p}|x_i - y_i|$
- Minkowski: $d(\vec{x},\vec{y}) = (\sum_{i=1}^{p}|x_i - y_i|^q)^{\frac{1}{q}}$
- Gower: Manhattan(Continuous) + Dice coefficient(Nominal)

4. 반응변수 전처리 (Target Engineering)

주로 parametric model 에서 예측 및 모델 적용을 위해서 사용
(Gaussian Distribution , Ordinary Linear Regression)

- Log Transformation

  · 오른쪽으로 치우친 분포 (Right skewed)가 정규 분포로 변환



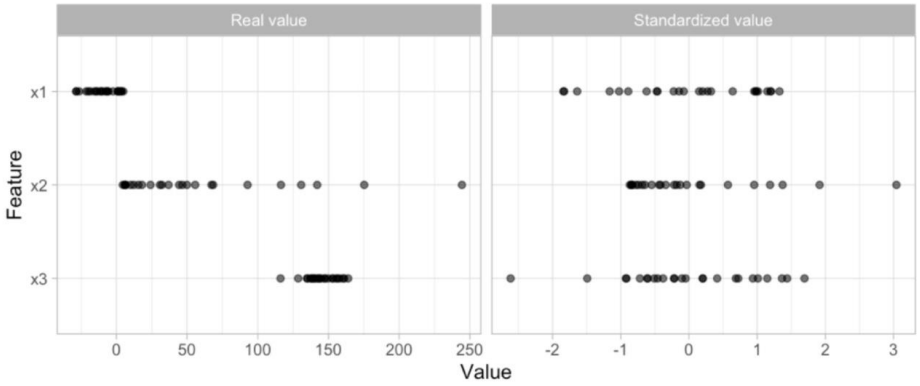
- Box-cox Transformation

5. Feature 표준화 (Standardiztion)

각 Feature의 측정 단위에 대한 보정

Feature의 단위가 각각 다를 경우 특정 Feature에 의해 다른 Feature가 학습에 적용 되지 않을 수 있다.

Centering and scaling을 통해서 평균이 0, 표준편차가 1이 되도록 변환해준다.



6. 결측치의 처리(Missing Data Processing)

6.1 결측치의 종료

- 무작위 결측치(Random missing value)
  - 완전 무작위 결측치(MCAR : Missing Completely At Random)
  - 무작위 결측치(MAR : Missing At Random)
- 비무작위 결측치(NAMR : Not Missing At Random)

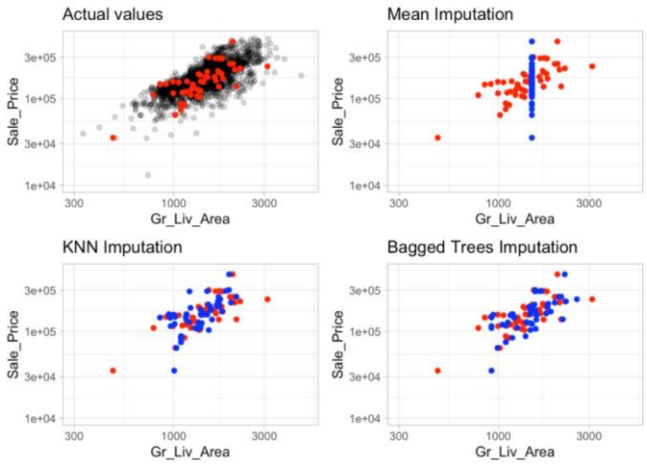
6.2 결측치의 대체 (Imputation)

- 결측치를 "최사의 추측"값으로 대체

  Estimated statistic(e.g Mean, Median, Mode, Regression)

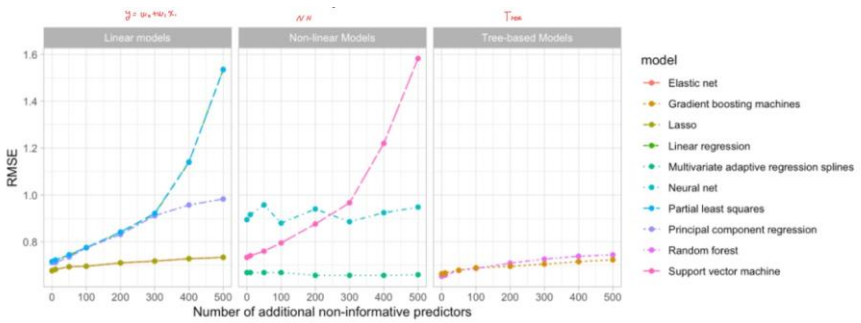
  k-nearest neighbor

  Tree-based



7. 중요하지 않은 Feature 제거 (Filtering)

의미없는 변수들(Non-Informative predictors)을 포함했을 때 RMSE의 변화

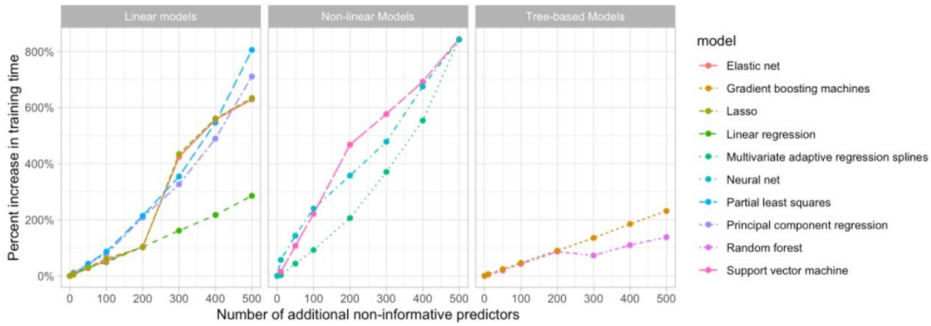


$MSE = \frac{1}{n} SSE$

$SSE = \sum_i^n (y_i - \hat{y}_i)^2$

$RMSE = \sqrt{MSE}$

의미없는 변수들 (non-informative predictors) 을 포함했을 때 학습시간변화



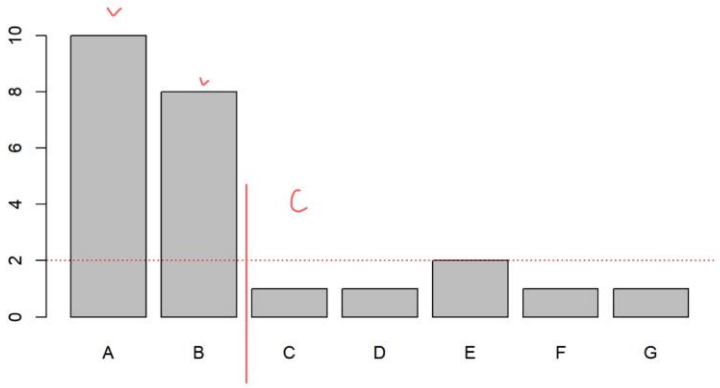
8. 제로 분산 Feature (Zero variance Features)

제로분산 Feature를 판단하는 일반적 기준

- 전체 샘플중에 서로 다른 관측값의 비율이 낮은 경우(10% 이하)
- 가장 빈도가 높은 관측값과 두 번째로 높은 관측값 과의 비가 높은경우 (약 20배 이상)

9. 범주형 데이터(Categorical Feature Engineering)

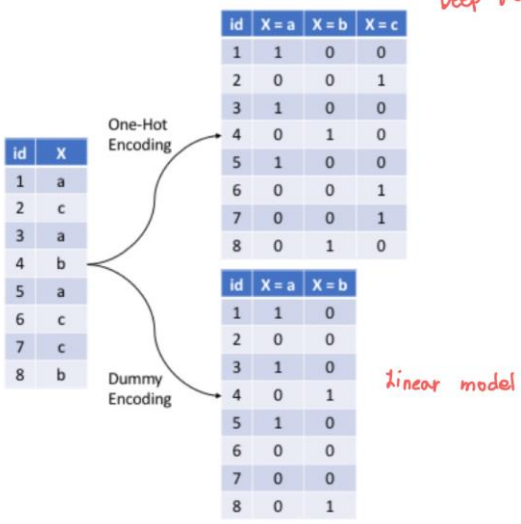
-재범주화 (Lumping) : 매우 작은 빈도를 갖는 범주를 모아서 하나의 범주로 재범주화



-One-hot & Dummy Encoding 각 범주를 1 또는 0(Bool)로 표시

True or False) 로 표시

Deep Learning



id	X
1	a
2	c
3	a
4	b
5	a
6	c
7	c
8	b

id	X = a	X = b	X = c
1	1	0	0
2	0	0	1
3	1	0	0
4	0	1	0
5	1	0	0
6	0	0	1
7	0	0	1
8	0	1	0

id	X = a	X = b
1	1	0
2	0	0
3	1	0
4	0	1
5	1	0
6	0	0
7	0	0
8	0	1

linear model

-Label Encoding : 각 범주 자료를 연속형 변수로 바꾸어 표현(순서형 자료의 경우)

- Replacing with the mean or proportion

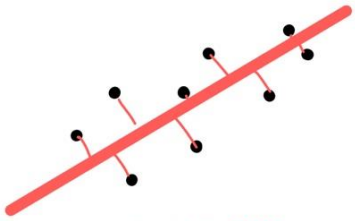
차원 축소 (Dimension reduction)

- 여러 개의 feature에서 불필요한 feature들 제거하는 방법
- 예) 주성분 분석 (PCA, principal components analysis)

정보의 손실 최소화

100개의 Features → 10 Features

ex) Eigen Value
Eigen vector



데이터 손실 최소화

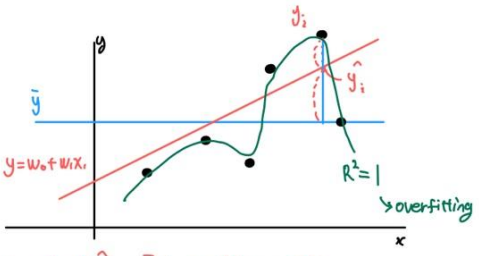
모델평가 지표 (Model evaluation metrics)

회귀분석 모델 (Regression models)

- MSE (Mean squared error) = $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- RMSE (Root mean squared error) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- MAE (Mean absolute error) = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

$= \frac{SSR+SSE}{SST} - \frac{SSE}{SST}$

$0 \le R^2 \le 1$



$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$

$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$ (=0)

$SST = SSR + SSE$

모델평가 지표 (Model evaluation metrics)

분류 모델 (Classification models)

- Misclassification
- Mean per class error
- MSE
- Cross entropy
- Gini Index

① High = 25 (6), Mid = 30 (6 miss), Low = 35 (4 miss) → $\frac{15}{90} = 17\%$ Misclassification Rate

② $\frac{6}{25}$, $\frac{6}{30}$, $\frac{4}{35} \rightarrow 14.5$

③ A ← 0.91, B ← 0.77, C ← 0.84 ⇒ $\frac{(1-0.91)^2 + (1-0.77)^2 + (1-0.84)^2}{3}$

Confusion Matrix(혼돈 행렬)

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$ → Recall
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$ → 1-Specificity = False Positive Rate
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

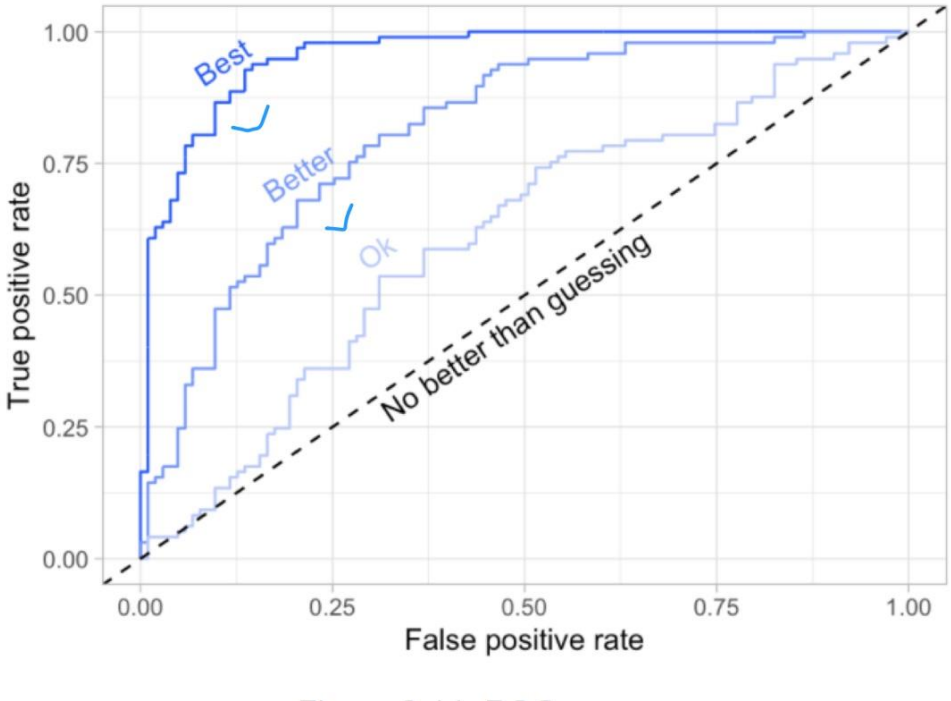
Confusion matrix 예제

		Predicted Class: Positive	Predicted Class: Negative	Recall
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

True	Test_yes	Test_No
Yes	8	2
No	1	5

ROC(Receiver Operating Characteristic curve) 와 AUC(Area Under the Curve)

- 좋은 분류 모델은 높은 정밀도와 감도를 가지게 되고 오분류율을 최소화한다.



ROC (Receiver Operating Characteristic curve) 예제

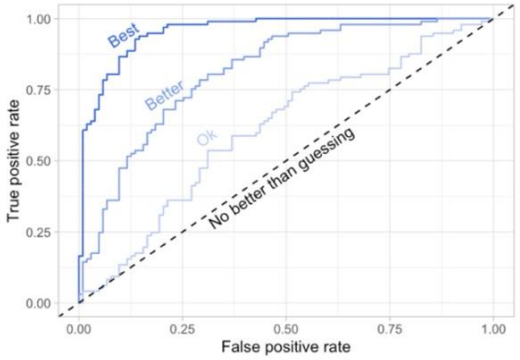
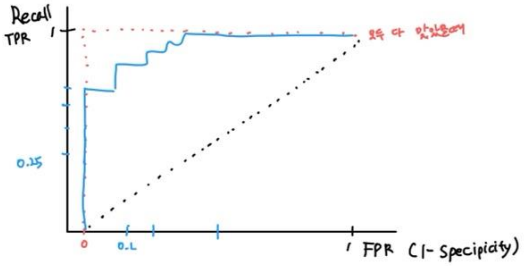


Figure 2.14: ROC curve.



https://bradleyboehmke.github.io/HOML