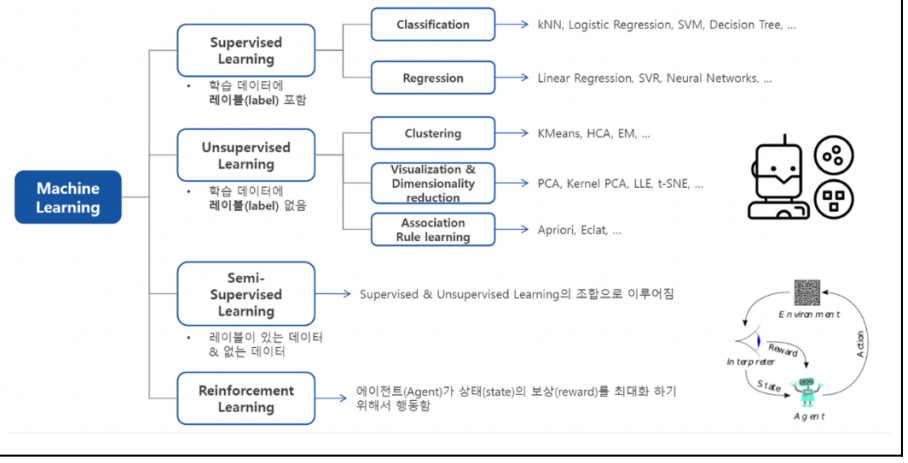
교육 제목	데이터 기반 인공지능 시스템 엔지니어 양성 과정_ 머신러닝
교육 일시	2021년 10월 14일
교육 장소	YGL C-6 학과장 & 자택(디스코드 이용한 온라인)
교육 내용	
오전	지난 시간 Review & 선형회귀 선형회귀(Linear regression) 1. 최소제곱추정량(OLS : Ordinary least square estimation) Linear model, $\hat{y}(x_i) = \hat{\theta}_0 + \hat{\theta}_1 x_i$ 2. 손실함수(Loss function, J) $J(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\theta}_0 - \hat{\theta}_1 x)^2 = \sum_{i=1}^{n} e_i^2$ - $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$ - Sum of Squares of the Errors (SSE) = $\sum_i^n e_i^2$ • Goal is to solve for $\hat{\theta}_0$ and $\hat{\theta}_1$ to minimize the objective function. $\hat{\theta}_o, \hat{\theta}_1 = argmin_{\theta_0,\theta_1} \sum_i^n e_i^2 = argmin_{\theta_0,\theta_1} J(\theta)$ 기계학습(Machine Learning ML) 1. 기계학습의 분류  2. Training set and test set

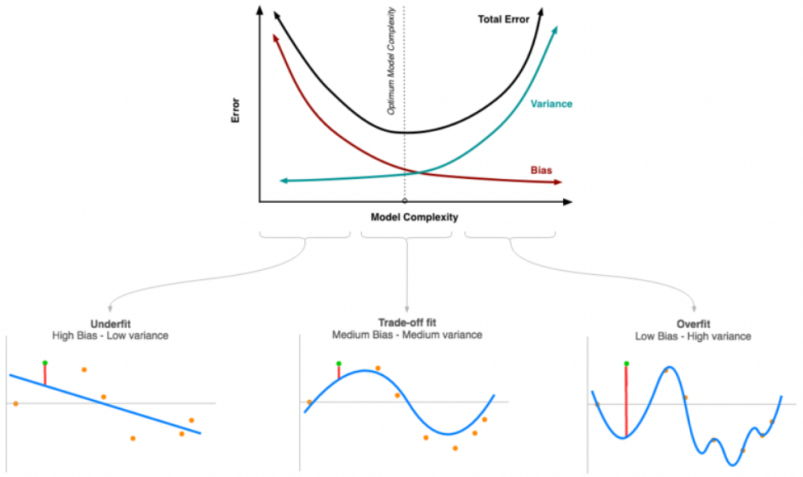
- ML 모델의 성능평가를 위해서 자료를 분할

- **Training set**: 모델의 알고리즘 learning, 모델에 사용될 feature들을 결정, 초매개변수 조절 (약 전체 자료수의 70% 로 설정)
  - **Training set**: 모델의 알고리즘 learning
  - **Validation set**: 모델에 사용될 feature들을 결정, 초매개변수 조절, 과적합 (Over-fitting) 방지
- **Test set**: 최종 선택된 모델의 성능평가 (약 전체 자료수의 30% 로 설정), 자료의 수가 적을 경우 생략 가능

3. ML 모델의 치우침(Bias)과 분산(variance)

> 예측값들과 정답이 대체로 멀리 떨어져 있으면 결과의 편향(bias)이 높다고 말하고,
> 예측값들이 자기들끼리 대체로 멀리 흩어져있으면 결과의 분산(variance)이 높다고 말합니다.

Total Error가 최소화 되는 지점을 찾아 초매개변수를 정해야 함



4. 기계학습 모델 평가 : k-fold 교차검증(k-fold cross validation(CV))

- k-fold 교차 검증 (일명 k-fold CV)은 훈련 데이터를 동일한 크기의 k 그룹 (k-fold)으로 무작위로 나누는 리샘플링 방법
- k-fold CV 추정치는 k 테스트 오류를 평균화하여 계산.

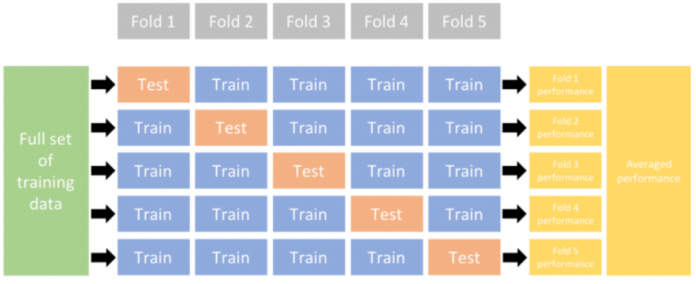
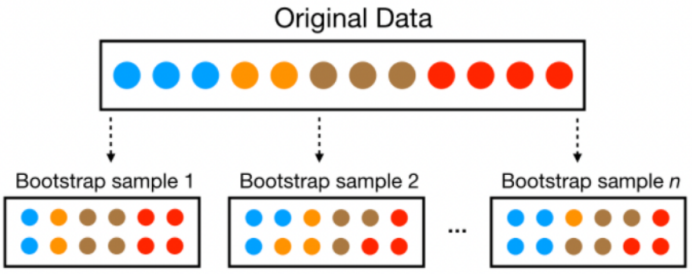


Figure 2.4: Illustration of the k-fold cross validation process.

오후

5. 기계학습 모델 평가 : Bootstrapping

- Bootstrapping 샘플은 복원추출을 이용한 데이터의 무작위 샘플.
- Bootstrapping은 선택한 샘플을 기반으로 모델을 구축하고 OOB (Out-of-Bag) 샘플을 이용하여 모델을 평가



6. 초매개변수 조절(Hyperparameter tuning)
   - 초매개변수는 학습 과정을 제어하는 데 사용되는 매개 변수를 의미
   - 초매개변수는 모델 학습과정이 아닌 모델 개발자에 의해서 지정됨
7. KNN (K-nearest neighbors classification)
   - 지도학습으로 분류나 회귀에 사용되는 비모수적 방법
   - 파라메터 학습을 위한 훈련과정이 없으나 훈련 집합은 필요
   - 각 데이터 간 거리를 계산하기 위한 거리 척도 필요
   - 초매개변수 k를 성정해야 함
   - 거리에 대한 가중치 고려
8. Feature 표준화(Standardization)
   - 각 feature의 측정 단위에 대한 보정 : 가격과 평수 비교를 위한 보정
   - Centering and scaling을 통해서 평균이 0, 표준편차 1이 되도록 변환
9. 결측치 처리
   - 결측치 종류 : 완전무작위(단순 결측), 무작위 결측(일부가 답을 안함), 비무작위 결측치(특정 부류에 의해 결측 발생)
   - 결측치 대체 : Estimated statistic(Mean, Median, Mode, Regression…), K-nearest neighbor, Tree-based
10. 중요하지 않은 Feature 제거(Filtering)
11. 제로분산 Feature(Zero variance Features)
12. 범주형 데이터 (Categorical Feature) Engineering
    - 재범주화(Lumping)
    - One-hot & dummy encoding
    - Label encoding
    - Replacing with the mean proportion
13. 차원축소(Dimension reduction) : 여러개의 feature 중 불요한 것을 제거
14. 모델평가 지표 (Model evaluation metrics)
    - 회귀분석모델

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

    - 분류 모델 (Classification models)

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

○ Confusion matrix(혼동행렬, 분류결과표

		Predicted Class: Positive	Predicted Class: Negative	
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** $\frac{TP+TN}{(TP+TN+FP+FN)}$

○ ROC (Receiver Operating Characteristic curve)와 AUC(Area under the curve)

- 좋은 분류모델은 높은 정밀도와 감도 가지게 되고 오분류율 (위양성 또는 위음성)을 최소화 함