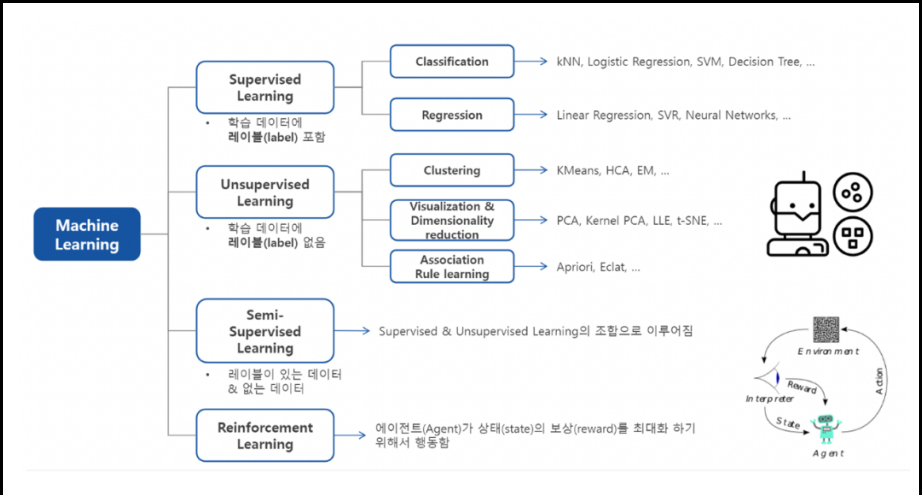
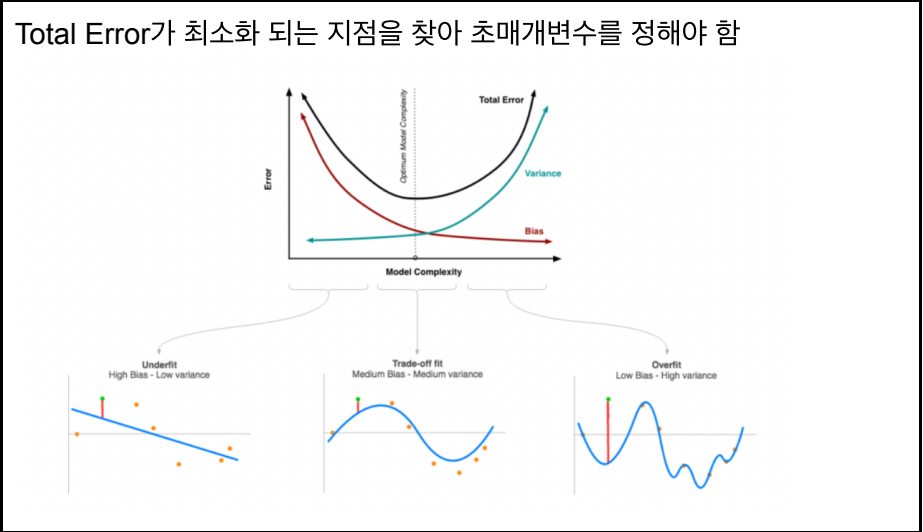


교육 제목	데이터 기반 인공지능 시스템 엔지니어 양성 과정_ 머신러닝
교육 일시	2021년 10월 15일
교육 장소	YGL C-6 학과장 & 자택(디스코드 이용한 온라인)
교육 내용	
오전	<p>지난 시간 Review &amp; 기계학습(Machine Learning ML)</p> <p>1. 기계학습의 분류</p>  <p>The diagram classifies Machine Learning into four main types: Supervised Learning (Classification: kNN, Logistic Regression, SVM, Decision Tree, ...; Regression: Linear Regression, SVR, Neural Networks, ...), Unsupervised Learning (Clustering: KMeans, HCA, EM, ...; Visualization &amp; Dimensionality reduction: PCA, Kernel PCA, LLE, t-SNE, ...; Association Rule learning: Apriori, Eclat, ...), Semi-Supervised Learning (Supervised &amp; Unsupervised Learning의 조합으로 이루어짐), and Reinforcement Learning (에이전트(Agent)가 상태(state)의 보상(reward)을 최대화 하기 위해서 행동함). A small diagram on the right shows an Agent interacting with an Environment to receive a Reward.</p> <p>2. Training set and test set</p> <ul style="list-style-type: none"> <li>ML 모델의 성능평가를 위해서 자료를 분할</li> <li><b>Training set:</b> 모델의 알고리즘 learning, 모델에 사용될 feature들을 결정, 초매개변수 조절 (약 전체 자료수의 70% 로 설정) <ul style="list-style-type: none"> <li><b>Training set:</b> 모델의 알고리즘 learning</li> <li><b>Validation set:</b> 모델에 사용될 feature들을 결정, 초매개변수 조절, 과적합 (Over-fitting) 방지</li> </ul> </li> <li><b>Test set:</b> 최종 선택된 모델의 성능평가 (약 전체 자료수의 30% 로 설정), 자료의 수가 적을 경우 생략 가능</li> </ul> <p>3. ML 모델의 치우침(Bias)과 분산(variance)</p> <p>Total Error가 최소화 되는 지점을 찾아 초매개변수를 정해야 함</p>  <p>The graph shows Error vs. Model Complexity. The Total Error curve is U-shaped, with Bias decreasing and Variance increasing as model complexity increases. The minimum Total Error occurs at the 'Optimum Model Complexity'. Below the graph, three plots illustrate: Underfit (High Bias - Low variance), Trade-off fit (Medium Bias - Medium variance), and Overfit (Low Bias - High variance).</p>

4. 기계학습 모델 평가 : k-fold 교차검증(k-fold cross validation(CV))
5. 기계학습 모델 평가 : Bootstrapping : 중복 추출
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
  - 파라미터 학습을 위한 훈련과정이 없으나 훈련 집합은 필요
  - 각 데이터 간 거리를 계산하기 위한 거리 척도 필요
  - 초매개변수 k를 설정해야 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. 모델평가 지표 (Model evaluation metrics)
  - 회귀분석모델

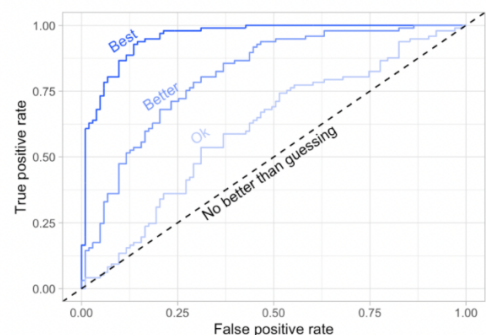
- MSE (Mean squared error) =  $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- RMSE (Root mean squared error) =  $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
- MAE (Mean absolute error) =  $\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$

- Confusion matrix(혼동행렬, 분류결과표)

		Predicted Class		
		Positive	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



<p>오후</p>	<p>혼공머(혼자 공부하는 머신러닝)</p> <p>01-3. 마켓과 머신러닝</p> <ul style="list-style-type: none"> <li>● KNN(k-최근접 이웃: k-nearest neighbor)알고리즘 : 어떤 데이터의 답을 구할 때 주위의 다른 데이터를 보고 다수를 차지하는 것을 정답으로 사용 (범주형) <ul style="list-style-type: none"> <li>○ 어떤 규칙을 찾는 것이 아닌 전체 데이터를 메모리에 가지고 있는 것</li> </ul> </li> <li>● 머신러닝 알고리즘에서 데이터가 규칙을 찾는 과정을 훈련이라고 한다. 사이킷런에서 fit()메서드가 하는 역할</li> <li>● 머신러닝 프로그램에서 알고리즘이 구현된 객체를 모델이라고 함, 종종 알고리즘 자체가 모델이라고 부름</li> <li>● 정확도는 정확한 답의 비율을 백분율로 나타낸 값. 사이킷런에서는 0~1 사이의 값으로 출력</li> </ul> <p>02-1. 훈련 세트와 데이터 세트</p> <ul style="list-style-type: none"> <li>● 지도학습은 입력과 타겟을 전달하여 모델을 훈련한 다음 새로운 데이터를 예측하는데 활용</li> <li>● 비지도학습은 타겟 데이터가 없음. 무엇을 예측하는 것이 아닌 데이터의 어떤 특징을 찾음</li> <li>● 훈련 세트는 모델을 훈련하는데 사용하는 데이터로 훈련 세트가 클수록 좋음</li> <li>● 테스트 세트는 전체 데이터의 20~30%를 사용하는 경우가 많음</li> </ul> <p>02-2. 데이터 전처리</p> <ul style="list-style-type: none"> <li>● 데이터 전처리란 머신러닝 모델에 훈련 데이터를 주입하기 전에 가공하는 단계</li> <li>● 표준 점수 : 훈련세트의 스케일을 바꾸는 대표적 방법. 표준점수는 특성의 평균을 빼고 표준편차로 나눔. ! 반드시 테스트 세트도 훈련세트의 평균과 표준편차 사용</li> <li>● 브로드캐스팅 : 크기가 다른 넘파이 배열에서 자동으로 사칙연산을 모든 행이나 열로 확장하여 수행하는 기능</li> </ul> <p>03-1. k-최근접 이웃 회귀</p> <ul style="list-style-type: none"> <li>● 회귀는 임의의 수치를 예측하는 문제로 타겟값도 임의의 수치임</li> <li>● k-최근접 이웃 회귀는 knn알고리즘을 사용하여 회귀문제를 해결함. 가장 가까운 이웃을 찾고 이웃들의 타겟값을 평균하여 예측</li> <li>● 결정계수(<math>R^2</math>) : 회귀 문제의 성능 측정 도구. 1에 가까울 수록 좋고, 0에 가까울 수록 나쁨.</li> <li>● 과대적합 : 모델의 훈련 세트 성능이 테스트 세트 성능보다 훨씬 높을 때 발생(모델이 훈련 세트에 집착하여 데이터에 내재된 거시적 패턴을 감지하지 못함) 과소적합은 반대 경우로 더 복잡한 모델을 사용하여 훈련 세트에 잘 맞는 모델을 만들어야 함</li> </ul>
-----------	--