머신러닝 치트시트

머신러닝의 종류

- 지도 학습 (Supervised Learning): 입력(X)과 정답(Y) 데이터 쌍을 학습하여, 새로운 입력에 대한 정답을 예측.
  - `회귀 (Regression)`: 연속적인 값을 예측. (예: 집값 예측)
  - `분류 (Classification)`: 이산적인 범주를 예측. (예: 스팸 메일 분류)
- 비지도 학습 (Unsupervised Learning): 정답이 없는 데이터를 학습하여 데이터의 숨겨진 구조나 패턴을 발견.
  - `군집화 (Clustering)`: 유사한 데이터들을 그룹으로 묶음.
  - `차원 축소 (Dimensionality Reduction)`: 데이터의 특징(feature) 수를 줄임.
- 강화 학습 (Reinforcement Learning): 에이전트(agent)가 환경(environment)과 상호작용하며 보상(reward)을 최대화하는 행동(action)을 학습.

지도 학습 알고리즘

- 선형 회귀 (Linear Regression): 데이터에 가장 잘 맞는 직선을 찾아 회귀 모델을 만듦.
- 로지스틱 회귀 (Logistic Regression): 시그모이드 함수를 사용하여 클래스에 속할 확률을 예측하는 분류 모델.
- 서포트 벡터 머신 (SVM, Support Vector Machine): 클래스 간의 경계(margin)를 최대화하는 결정 경계(decision boundary)를 찾는 분류/회귀 알고리즘.
- 결정 트리 (Decision Tree): 데이터를 질문에 따라 나누어 가며 예측하는 나무 구조의 모델. 해석이 용이함.
- K-최근접 이웃 (K-Nearest Neighbors, KNN): 새로운 데이터 주변의 K개 이웃들의 다수결로 클래스를 예측.
- 나이브 베이즈 (Naive Bayes): 베이즈 정리를 기반으로 한 분류 알고리즘. 특징들이 서로 독립이라고 가정.

비지도 학습 알고리즘

- K-평균 군집 (K-Means Clustering): 데이터를 K개의 군집(cluster)으로 묶는 알고리즘. 각 데이터는 가장 가까운 군집의 중심(centroid)에 속하게 됨.
- 주성분 분석 (PCA, Principal Component Analysis): 데이터의 분산을 가장 잘 설명하는 새로운 축(주성분)을 찾아 데이터의 차원을 축소.
- 계층적 군집 분석 (Hierarchical Clustering): 개별 데이터에서 시작하여 유사한 클러스터를 점차 합쳐나가는 방식(agglomerative) 또는 전체에서 나눠나가는 방식(divisive)으로 군집을 형성.

앙상블 학습 (Ensemble Learning)

여러 개의 약한 학습기(weak learner)를 결합하여 더 강력한 모델을 만드는 기법.

- 배깅 (Bagging, Bootstrap Aggregating): 원본 데이터에서 부트스트랩 샘플링(중복 허용 샘플링)하여 여러 모델을 학습시키고 결과를 평균 또는 다수결로 합산.
  - `랜덤 포레스트 (Random Forest)`: 배깅을 적용한 결정 트리의 앙상블. 각 트리를 학습할 때 특징의 일부만 무작위로 선택하여 다양성을 높임.
- 부스팅 (Boosting): 이전 모델이 잘못 예측한 데이터에 가중치를 주어 다음 모델이 더 잘 학습하도록 순차적으로 모델을 개선.
  - `AdaBoost`, `Gradient Boosting Machine (GBM)`, `XGBoost`, `LightGBM`

모델 평가 지표

- 회귀 (Regression):
  - `MAE (Mean Absolute Error)`: |실제값 - 예측값|의 평균.
  - `MSE (Mean Squared Error)`: (실제값 - 예측값)^{2}의 평균.
  - `R² (결정 계수)`: 모델이 데이터의 분산을 얼마나 설명하는지 나타냄. 1에 가까울수록 좋음.
- 분류 (Classification): (혼동 행렬 기반)
  - `정확도 (Accuracy)`: (TP+TN) / (TP+TN+FP+FN). 전체 중 올바르게 예측한 비율.
  - `정밀도 (Precision)`: TP / (TP+FP). Positive로 예측한 것 중 실제 Positive인 비율.
  - `재현율 (Recall)`: TP / (TP+FN). 실제 Positive인 것 중 Positive로 예측한 비율.
  - `F1 점수 (F1 Score)`: 정밀도와 재현율의 조화 평균. `2 * (Precision * Recall) / (Precision + Recall)`.
  - `AUC (Area Under the ROC Curve)`: ROC 곡선 아래의 면적. 1에 가까울수록 좋은 모델.

과적합 방지 (Regularization)

- L1 Regularization (Lasso): 가중치의 절댓값 합을 비용 함수에 추가. 일부 가중치를 0으로 만들어 특징 선택 효과.
- L2 Regularization (Ridge): 가중치의 제곱 합을 비용 함수에 추가. 가중치를 0에 가깝게 만들어 모델을 부드럽게 함.
- 드롭아웃 (Dropout): (딥러닝에서) 학습 시 신경망의 일부 뉴런을 무작위로 비활성화.
- 조기 종료 (Early Stopping): 검증 데이터셋의 성능이 더 이상 향상되지 않을 때 학습을 중단.