|  |
| --- |
| **SSAC**  **파이썬을 활용한 비즈니스 빅데이터 분석가 양성 과정**  **(TEMPLATE) 자가학습 워크시트** |

|  |  |
| --- | --- |
| **이름** | 백승재 |
| **날짜** | 2021-11-16 |
| **온라인 콘텐츠(인강)**  **학습 범위** | * 직장인을 위한 Python : 예측을 위한 머신러닝 라이브러리 Ch 01.01~02 |
| **복습 노트 작성 (자유 양식)** | Ch 01. 머신러닝의 개요 – 01. 머신러닝의 개요  - 머신러닝이란 : 데이터 기반(Data) + 패턴(관계,데이터들(x,y)의 상관관계, hyphothesis(model))+ 결과 추론(예측 (prediction))  - 머신러닝의 모델(hyphothesis)는 User가 선택하고 머신러닝은 Weight와 b를 찾는다.  - 과거에는 사용자가 데이터를 기반으로 모델을 생성하여(전체 과정을 user가 담당) 예측을 하였지만, 머신러닝을 이용하면 훈련 데이터와 테스트 데이터가 있다면 데이터를 기반으로 사용자와 머신러닝과 협동하여 모델을 생성하여 예측할 수 있다.  1. 머신러닝의 종류  1) 지도 학습 : 입력값과 레이블 값이 있음  - 회귀 : 수치형 (집값, 가격, 온도)  - 분류: 범주형(스팸메일, 종류 판별, 암 진단)  2) 비지도 학습 : 레이블할 값이 없음  - 군집화 (뉴스 분류, 사용자 관심사)  - 차원 축소  2. 머신 러닝의 장 단점  1) 장점  a) 복잡한 패턴(상관관계)을 인지 (가중치와 편향계수)  b) 적절한 알고리즘, 다양한 양질의 데이터가 있다면, 좋은 성능  c) 도메인 영역에 대한 지식이 상대적으로 부족해도 가능하다. 모델을 이용하여 중요한 feature를 파악할 수 있다.  2) 단점  a) 데이터의 의존성이 크다. (데이터의 질이 낮으면 결과값도 좋지 않다. (Garbage in/out)  b) 과적합의 오류에 빠질 수 있다.  - 과한 학습 상태에서 새로운 값이 들어 올 때 맞추지 못한다. (일반화 오류)  c) 풍부한 데이터가 기본적으로 요구된다.  - 데이터가 적으면 편향적인 학습 -> 과적합의 오류  3. 모델 성능을 좋게 하기 위한 조건  a. 데이터의 질과 양  - 데이터 가공 = 전처리 (Pre-Processing)  b. 적합한 모델(알고리즘) 선택 : 사용자가 예측해야할 값에 맞는 모델 선택, 데이터의 탐색 단계와 도메인 지식이 필요함  Ch 01. 머신러닝의 개요 – 02. 가설 함수, 비용, 손실 함수, 경사하강법  - 머신러닝의 원리  1. 가설 함수 (Hyphothesis)  - 가설 함수를 통해서 나온 값 -> H(x) = Y Predict  2. 손실 함수(Cost, Loss Function)  - 손실(loss) : 예측값과 얼마나 떨어져 있는 가 (예측값- 실제값)  - H(x)(예측값) – Y(실제값)  - 손실의 총합의 오류 -> 손실의 총합이 0인 경우가 한 개가 아님 -> 손실 함수에 제곱을 취해줌 -> 데이터의 개수가 많아지면 손실이 커지므로 전체 손실의 평균을 구함(제곱오차(MSE))  3. 경사하강법 (Loss fucntion)  - 손실 함수의 기울기가 최소인 지점을 찾아 나가는 방법  - 과정:  1) 모델을 선정한다.  2) initial W를 설정한다  3) 손실 함수를 계산한다.  4) 손실 함수의 기울기를 업데이트하면서 기울기가 최소인 점을 찾는다. |
| **퀴즈 제출** | Q1.다음 선형회귀에 대한 설명 중 적절하지 않은 것은?  1) 선형회귀에는 Linear Regression,Ridge Regression, Lasso Regression, Elastic Net이 있으며 Hyphothesis는 동일하다.  2) 선형회귀에 Hyphothesis가 같으면 예측값도 동일하다.  3) Ridge와 Lasso의 차이는 규제를 가할 때 모델의 복잡도를 의미하는 weight를 조정하는데 이때 사용하는 weight의 크기 척도가 다르다. (Ridge는 L2 Norm을 Lasso는 L1 Norm을 사용한다. )  4) 과대 적합이 발생할 때 규제의 강도를 높여야 하며 이때 오차의 비율이 적어지며 오차가 완화된다.  정답 2:  - 선형회귀에서 Hyphothesis의 폼은 동일하지만 폼 안에 구성하는 weight와 bias가 다르기 때문에 예측값은 다르다.  Q2. 다음 선형회귀 모델의 평가에 대한 설명 중 적절하지 않는 것은?  1) 모델의 예측은 hyphothesis로 하고 평가는 cost함수로 한다.  2) mean\_squared\_error 예측값과 실제값의 차이를 제곱한 것으로 평균제곱오차(MSE)에 해당한다.  3) MSE는 오차가 제곱이 되어 실제 오류보다 과장 되기 때문에 이에 루트를 씌운 것이 RMSE이다.  4) 일반적으로 RMSE가 실제값과 비슷하므로 R2score보다 많이 사용된다.  정답 : 4  - 스케일링이 안된 상태에서 RMSE를 가지고 모델의 성능의 평가하기 힘든 경우가 있으므로 score의 점수가 0과 1사이에 점수로 산출되는R2 score를 일반적으로 많이 사용한다.  Q3. 머신러닝의 성능 향상을 위해서는 적절한 데이터와 모델 선정이 중요하다. 다음 모델 선정에 대한 내용으로 적절한 것은?  1) 주가가 오를 것인지 내릴 것인지 예측하기 위해서는 회귀모델을 사용한다.  2) 주가의 추세를 예측하기 위해서는 분류모델을 사용하며 대표적인 모델은 SVM이다.  3) 뉴스기사를 텍스트 분석하여 그룹핑하기 위해서 분류 모델을 사용한다.  4) 암이 양성인지 음성인지 확인하기 위해 분류 모델을 사용하며 로지스틱 회귀 모델을 사용한다.  정답 4:  1) 주가가 오를 것인지 내릴 것인지 예측하기 위해서는 분류 모델을 사용하며 대표적인 모델은 SVM이다.  2) 주가의 추세(주가의 시세)를 예측하기 위해서는 회귀 모델을 사용하며 대표적인 모델은 선형회귀 모델이다.  3) 뉴스 기사를 그룹핑하기 위해서는 라벨이 없는 상태에서 분류하는 것이므로 비지도학습이며 군집화 방법을 사용한다. |
| **교육생 사진 (교육생 모습 + 교육 화면)** |  |
| **화면 캡쳐**  **(교육 화면 + 교육일시)** |  |