데이터과학 HW

제출일 : 2021/05/05

학번/이름 : 201400875/김용준

1-1 지도학습 vs 비지도학습

▶ 가장 큰 차이는 “답을 알려주는가?” 에 대한 대답이다. 지도학습은 Yes, 비지도학습은 No 이다.
지도학습을 하지 않을 때에는 지도학습처럼 답을 알려주기 위해 구해가는 과정을 우선적으로 처리하는 것을 비지도 학습이라고 생각해도 될 것 같다.

1-2 회귀분석과 분류에 대해 비교하여 설명하시오

▶ 회귀분석과 분류는 둘 다 지도학습의 일부이다. 회귀와 분류의 가장 큰 차이는 무엇을 구하고 싶은가? 이다. 회귀는 ‘숫자’를 예측하고 싶을 때, 분류는 ‘이름이나 문자’를 과거의 경험을 토대로 구할 때 사용한다고 생각하면 된다.

1-3 단순 선형 회귀분석은 무엇을 최적화하여 어떤 문제를 푸는 것인지 설명하시오.

▶ $y = \beta 0 + \beta 1x + \varepsilon$ 의 식을 가짐으로써 최소제곱추정(오차 제곱합을 최소로 함)을 최적화하여 정형화된 데이터나 단순한 키-몸무게 간의 산점도 정도의 문제를 푼다.

1-4 회귀분석에서 정규화를 쓰는 이유를 설명하시오

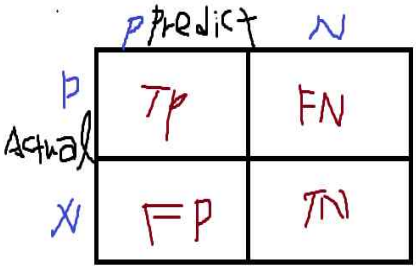
▶ 가장 큰 이유는 과적합(Overfitting)을 해결하기 위해서 사용한다. 모든 특성을 사용하되, 파라미터의 값을 줄이고 특성이 많아도 잘 동작하게 해주기 위해 사용을 한다.

1-5 로지스틱 회귀가 어떻게 분류 문제를 해결하는지 설명하고 그 예를 한가지 들어보시오.

▶ 로지스틱 회귀분석은 어떤 사건(event)이 발생할지에 대한 직접 예측이 아니라 그 사건이 발생할 확률을 예측하는 것이다. 금융권에서 고객의 신용도 평가를 통해 고객의 신용도가 우량일지 불량일지를 알 수 있는 기법이다.

1-6 Confusion Maxrix를 그리고, Precision, recall, F1-Score를 설명해보시오.

▶ True Postive, False Negative, False Postive, True Negative 4개가 있으며
Precision(정밀도)는 TP/(TP+FP). 즉 맞다고 판단했을 때 그것이 진짜일 정도
Recall(재현도)는 TP/(TP+FN). 즉 예측한 것들 중에 그것이 진짜일 정도
F1-Score는 (Precision*Recall) * 2 / (Precision+Recall) 이다.



1-7 교차검증을 하지 않을 때 데이터 분석에 실패하는 예를 들어 보시오.

▶ 고정된 Train Set과 Test Set으로 평가를 하고 반복적으로 모델을 튜닝하다보면 Test Set에만 과적합(Over Fitting)이 되어버리는 결과가 생긴다.

1-8 Naive Bayes에서 Naive란 이름이 붙는 이유에 대해 설명하고 계산이 더 편리한 이유를 설명하시오.

▶ 베이즈정리는 기본적으로 조건부확률을 계산하는 방법 중에 하나이다. 나이브베이즈는 이 정리를 이용하여 Text분류를 수행한다. 이름이 붙여진 이유는 단순해서(빠르다!)이다. 계산이 더 편리한 이유는 모든 값이 독립임을 가정하기 때문이다.

1-9 Ensemble Learning의 장점에 대해 설명하고, Random Forest의 동작 방식에 대해 설명해보시오.

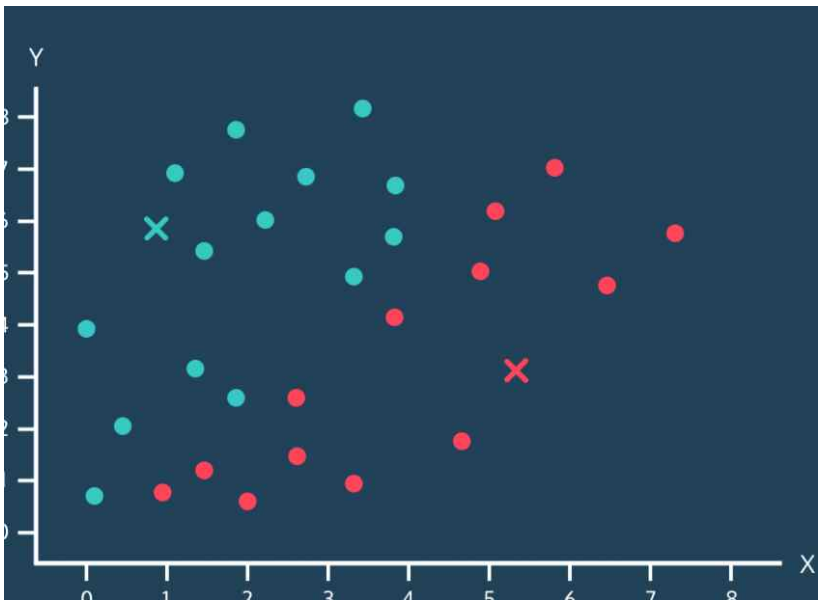
▶ Ensemble Learning은 일단 기본적으로 여러 가지 우수한 학습 모델을 조합하여 예측력을 향상시키는 학습법이다. 장점으로는 단일 모델에 비해서 분류 성능이 우수하다는 점이다.

▶ Random Forest는 여러 개의 결정 트리를 임의적으로 학습하는 Ensemble의 배깅유형으로, 별도의 튜닝이 필요 없이 분류, 회귀 분석에서 모두 사용이 가능하다.

1-10 k-means 알고리즘의 동작 과정을 요약하고 장/단점을 설명해보시오.
동일 데이터와 k값에 대해 k-means 알고리즘을 두 차례 수행한다면
결과가 같은가? 왜 그런지 대답에 대한 근거를 간략히 설명해보시오.

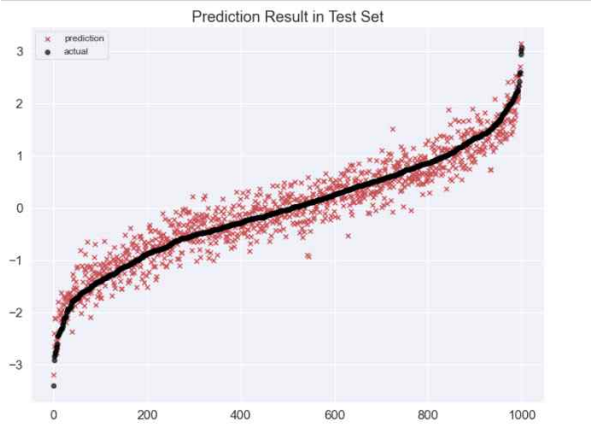
▶ 데이터포인트 n개중 k개를 선택하여 각 군집 중심(cluster center)으로 지정 -> 나머지 n-k개의 포인트들은 cluster center와 가장 가까운 군집으로 배정 -> cluster center를 군집 내 포인트 위치의 평균으로 업데이트 하고 과정 반복

▶ 같을 것이다. 다음 사진처럼 k를 정한 후 그 포인트를 중심으로 군집을 배정한 후 포인트럴 정하면서 움직이게 되는데, 똑같은 k값과 똑같은 데이터면 한번 더 수행해도 똑같을 것이다.



2. 회귀분석 실습 >> 다음 집값 예측모델을 학습하고 결과를 해석하라

▶ 집값 예측 모델을 돌려본 결과



▶ 집값과 상관이 있을법한 수치들의 상관관계

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
0	1.028660	-0.296927	0.021274	0.088062	-1.317599	-0.490081
1	1.000808	0.025902	-0.255506	-0.722301	0.403999	0.775508
2	-0.684629	-0.112303	1.516243	0.930840	0.072410	-0.490211
3	-0.491499	1.221572	-1.393077	-0.584540	-0.186734	0.080843
4	-0.807073	-0.944834	0.846742	0.201513	-0.988387	-1.702518

▶ 집값에 영향력이 있는 순서

	feature	coefficients
0	Avg. Area Income	0.653957
1	Avg. Area House Age	0.463246
4	Area Population	0.425856
2	Avg. Area Number of Rooms	0.344046
3	Avg. Area Number of Bedrooms	0.005391

▶ 대체적으로 집값의 좋은 영향을 끼치는 지표는 Income, Number Of Rooms, Bedrooms, Population이고 집이 오래될수록 부정적인 영향을 끼친다.

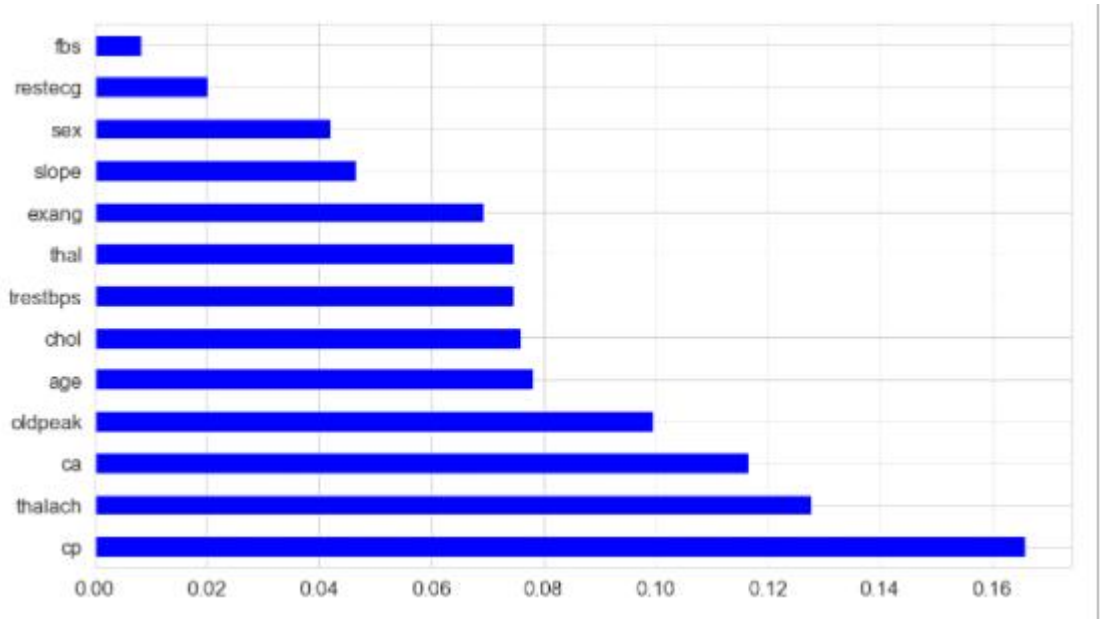
모델은 선형회귀 모델을 사용하였고, 변인들이 집값에 직접적으로 영향을 끼치기 때문에 사용하게 되었다.
linear_model.LinearRegression()
학습에 사용된 데이터는 위 5개, 테스트에 사용된 데이터는 집값 // Address는 제외
주어진 데이터로 집값을 예측한 결과와 실제 결과가 거의 일치하였다.

3. 심장병 분류 예측모델을 학습하고 결과를 해석해보세요

▶ Random Forest 모델로 심장병 분류 예측 모델을 학습한 결과

	precision	recall	f1-score	support
0	0.88	0.70	0.78	30
1	0.76	0.90	0.82	31
accuracy			0.80	61
macro avg	0.82	0.80	0.80	61
weighted avg	0.81	0.80	0.80	61

왜 Random Forest모델을 사용하게 되었냐면, 위에 정리해둔 내용중에 Ensemble기법에서도 분류, 회귀에서 모두 사용가능한 다재다능한 모델이기도 하고, 별도 튜닝이 필요가 없어서 언제든 사용이 가능하다는 점에서 사용하게 되었다.



각 기능별로 얼마나 모델에 영향을 끼쳤는지 알려주는 그래프이다.
충분히 기계학습으로 심장병인 환자를 분류할 수 있을 것이다.

코드는 주피터노트북 코드로 첨부하겠습니다.