|  |  |
| --- | --- |
| 교육 제목 | 머신러닝 |
| 교육 일시 | 2021년 10월 20일 |
| 교육 장소 | 영우글로벌러닝 |
| **교육 내용** | |
| 오전 | \*복습1시간  \*트리 알고리즘: 먼저 데이터를 가져와 많은 데이터가 있으면 StandardScaler 클래스를 사용해 특성을 표준화하고 훈련세트와 테스트세트로 나누고나서 훈련세트를 전처리한후 테스트세트를 변환하고 변환된 훈련세트와 테스트세트를 사용해 로지스틱 회귀모델을 훈련해봅니다. 그럼 생각보다 점수가 높지 않게나옵니다. 그럼여기서 결정 트리 모델을 사용해보곘습니다. 사이킷런 모델안에있는 DecisionTreeClassifier 을 사용해서 정확도를 편가해 보면 값을 잘 나오는것을 볼수있는데 이것을 plot.tree() 함수를 사용해 결정트리를 이해하기 쉬운 트리 그림으로 출력해줍니다. 그림은 그려보면 엄청난 트리가 만들어졌습니다. 너무 복잡하니 plot\_tree() 함수에서 트리의 깊이를 제한해서 출력하려면 max\_depth 매개변수를 1로주면 하나의 노드를 더확장하여 그림을 그립니다. 또 filled 매개변수에 클래스에 맞게 노드의 색도 칠할수있습니다. 그려보면 간단하게 그림이 출력되고 훨신 편안하게 이해할수있습니다. 해석해보자면 샘플값이 5197개라고 표시되고 그밑에 value 값이 나타나는데 음성 클랫스가 1258개 양성 클래스가 3939개로 표시됩니다. 그리고 루트노드 당도가 -0.239와 같거나 작으면 왼쪽가지로 그렇기 않으면 오른쪽가지로 이동 합니다. 다시 왼쪽으로 가서 노드를 살려보면 -0.802와 같거나 낮다면 다시 왼쪽가지로 그렇지 않으면 오른쪽가지로 이동합니다. 음성과 양성 클래스 샘플개수는 각가 1177개와 1745개 입니다. 즉 화이트 와인의 비율이 크게 줄었습니다. 그이유는 오른쪽 노드를 보면 알수 있는데 오른쪽노드는 음성 클래스가 81개 양성이 2194개로 대부분 화이트 와인 샘플이 이 노드로 이동했습니다. 결정 트리에서 예측하는 방법은 간단합니다. 리프 노드에서 가장 많은 클래스가 예측 클래스가 됩니다. 앞에서 보았던 K-최근접 이웃과 매우 비슷합니다. 만약 이결정트리의성장을 여기서 멈춘다면 왼쪽 노드에 도달한 샘플과 오른쪽 노드에 도달한샘플은 모두 양성 클래스로 예특됩니다. 두 노드 모두 양성 클래스의 개수가 많기 때문이죠 |
| 오후 | ,.\* 불순도 gini는 지니 불순도를 의미합니다. DecisionTreeClassifier 클래스의 criterion 매개변수의 기분값이 gini 입니다. 지니 불순도는 클래스 비율을 제공해서 더한다음 1에서 뺴면 도비니다. 지니 불순도 = 1-(음성클래스비율\*\* + 양성클래스비율\*\*) 이게끝입니다. 열매를 잘 맺기 위해 과수원에서 가지치기를 하는 것처럼 결정트리도 가지치기를 해야합니다 그렇지 ㅇ낳으면 무작정 끝까지 자나나는 트리가 만들어져서 훈련세트는 아주 잘맛곘지만 테스트 세트에서 점수는 그에 못 미치게되면 이를 두고 일반화가 잘 안될것 같다고 합니다. max\_depth 매개변수를 3으로 지정하여 모델을 만들면최대3개의 노드까지만 성장 가능합니다.  \*교차 검증과 그리드 서치 교차검증은 검증 세트를 떼어 내어 평가하는 과정을 여러번 반복하게되는데 그다음 이점수를 평균하여 최종 검증 점수를 얻습니다. 사이킷런에는 cross\_validate()라는 교차 검증 함수가 있습니다. 사용법은 간단한데 , 먼저 평가할 모델 객체를 첫 번째 매개변수로 전달합니다. 그다음 앞에서처럼 직접 검증 세트를 떼어내지 않고 훈련세트 전체를 cross\_validate() 함수에 전달합니다. 이함수는 fit\_time,score\_time,test\_score 키를 가진 딕셔너리를 반홥합니다. 그리고 cross\_validate 함수는 기본적으로 5\_폴드 교차 검증을 수행합니다. cv 매개변수로 폴드수를 바꿀수도 있습니다. 사이킷런의 분할기는 교차 검증에서 폴드를 어떻게 나눌지 결정해 줍니다. cross\_validate() 함수는 기본적으로 회귀 모델일 경우 KFold 분할기를 사용하고 분류 모델일 경우 타깃 클래스를 골고루 나누기위해 strafifiedfKFold 를 사용합니다. max\_depth의 최적의 값을 찾았다고 가정하고 max\_depth를 최적의 값으로고정하고 min\_sampled\_split을 바꿔가면 최적의 값을 찾습니다. 이렇게 한 매개변수의 최적값을 찾고 다른 매개 변수의 최적값을 찾아도 될까요? 아니요 틀렸ㅅ브니다. max\_depth 최적값은 min\_sample\_split 매개변수의 값이 바뀌면 함께 달라집니다. 즉 이 두 매개변수를 동시에 바꿔가며 최적의 값을 찾아야 하는거죠 개다가 매개변수가 많아짐녀 문제는 더복잡해집니다. 이러한문제를 사이킷런에서 제공하는 그리드 서치를 사용합니다. 사이킷런의 GridSearchCV 클래스는 친절하게도 하이퍼파라미터 탐색과 교차 검증을 한번에 수행합니다. 먼저 GridSearchCV 클래스를 임포트하고 탐색할 매개변수와 탐색할 값의 리스트를 딕셔너리로 만듭니다. 그다음 훈련세트에서 그리드 서치를 수행하여 최상의 평균 검증 점수가 나오는 매개변수 조합을 찾습니다. 이조합은 그리드 서치 객체에 저장되며 그리드 서치는 최상의 매개변수에서 전체 훈련 세트를 사용해 최종 모델을 훈련합니다. 이모델도 그리드 서치 객체에 저장됩니다. |