교육 제목	데이터 기반 인공지능 시스템 엔지니어 양성 과정_ 머신러닝
교육 일시	2021년 10월 20일
교육 장소	YGL C-6 학과장 & 자택(디스코드 이용한 온라인)
교육 내용	
오전	지난 시간 Review & 기계학습(Machine Learning ML) 혼공머(혼자 공부하는 머신러닝) 04-1. 로지스틱 회귀

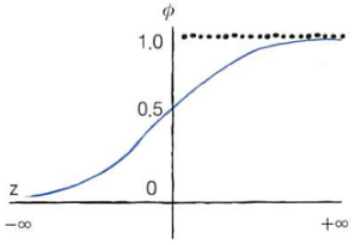
로지스틱 회귀logistic regression는 이름은 회귀이지만 분류 모델입니다. 이 알고리즘은 선형 회귀와 동일하게 선형 방정식을 학습합니다. 예를 들면 다음과 같습니다.

$$z = a \times (Weight) + b \times (Length) + c \times (Diagonal) + d \times (Height) + e \times (Width) + f$$

여기에서 a, b, c, d, e는 가중치 혹은 계수입니다. 특성은 늘어났지만 3장에서 다룬 다중 회귀를 위한 선형 방정식과 같습니다. z는 어떤 값도 가능합니다. 하지만 확률이 되려면 0~1(또는 0~100%) 사이 값이 되어야 합니다. z가 아주 큰 음수일 때 0이 되고, z가 아주 큰 양수일 때 1이 되도록 바꾸는 방법은 없을까요? **시그모이드 함수**sigmoid function(또는 **로지스틱 함수**logistic function)를 사용하면 가능합니다.

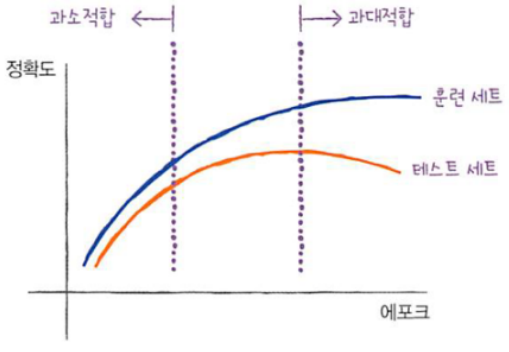
$$\phi = \frac{1}{1+e^{-z}}$$

시그모이드 함수



시그모이드 그래프

04-2. 확률적 경사 하강법



이 그래프는 에포크가 진행됨에 따라 모델의 정확도를 나타낸 것입니다. 훈련 세트 점수는 에포크가 진행될수록 꾸준히 증가하지만 테스트 세트 점수는 어느 순간 감소하기 시작합니다. 바로 이 지점이 모델이 과대적합되기 시작하는 곳입니다. 과대적합이 시작하기 전에 훈련을 멈추는 것을 **조기 종**

05-1. 결정 트리

- **결정 트리**는 예 / 아니오에 대한 질문을 이어나가면서 정답을 찾아 학습하는 알고리즘입니다. 비교적 예측 과정을 이해하기 쉽고 성능도 뛰어납니다.
- **불순도**는 결정 트리가 최적의 질문을 찾기 위한 기준입니다. 사이킷런은 지니 불순도와 엔트로피 불순도를 제공합니다.
- **정보 이득**은 부모 노드와 자식 노드의 불순도 차이입니다. 결정 트리 알고리즘은 정보 이득이 최대화되도록 학습합니다.
- 결정 트리는 제한 없이 성장하면 훈련 세트에 과대적합되기 쉽습니다. **가지치기**는 결정 트리의 성장을 제한하는 방법입니다. 사이킷런의 결정 트리 알고리즘은 여러 가지 가치지기 매개변수를 제공합니다.
- **특성 중요도**는 결정 트리에 사용된 특성이 불순도를 감소하는데 기여한 정도를 나타내는 값입니다. 특성 중요도를 계산할 수 있는 것이 결정 트리의 또다른 큰 장점입니다.

pandas

- info()는 데이터프레임의 요약된 정보를 출력합니다. 인덱스와 컬럼 타입을 출력하고 널(null)이 아닌 값의 개수, 메모리 사용량을 제공합니다. verbose 매개변수의 기본값 True를 False로 바꾸면 각 열에 대한 정보를 출력하지 않습니다.
- describe()는 데이터프레임 열의 통계 값을 제공합니다. 수치형일 경우 최소, 최대, 평균, 표준편차와 사분위값 등이 출력됩니다.

  문자열 같은 객체 타입의 열은 가장 자주 등장하는 값과 횟수 등이 출력됩니다.

  percentiles 매개변수에서 백분위수를 지정합니다. 기본값은 [0.25, 0.5, 0.75]입니다.

scikit-learn

- DecisionTreeClassifier는 결정 트리 분류 클래스입니다.

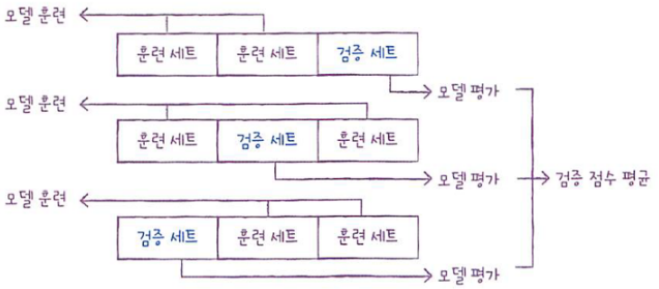
  criterion 매개변수는 불순도를 지정하며 기본값은 지니 불순도를 의미하는 'gini'이고 'entropy'를 선택하여 엔트로피 불순도를 사용할 수 있습니다.

  splitter 매개변수는 노드를 분할하는 전략을 선택합니다. 기본값은 'best'로 정보 이득이 최대가 되도록 분할합니다. 'random'이면 임의로 노드를 분할합니다.

  max_depth는 트리가 성장할 최대 깊이를 지정합니다. 기본값은 None으로 리프 노드가 순수하거나 min_samples_split보다 샘플 개수가 적을 때까지 성장합니다.

  min_samples_split은 노드를 나누기 위한 최소 샘플 개수입니다. 기본값은 2입니다.

  max_features 매개변수는 최적의 분할을 위해 탐색할 특성의 개수를 지정합니다. 기본값은 None으로 모든 특성을 사용합니다.

	• plot_tree()는 결정 트리 모델을 시각화합니다. 첫 번째 매개변수로 결정 트리 모델 객체를 전달합니다. max_depth 매개변수로 나타낼 트리의 깊이를 지정합니다. 기본값은 None으로 모든 노드를 출력합니다. feature_names 매개변수로 특성의 이름을 지정할 수 있습니다. filled 매개변수를 True로 지정하면 타깃값에 따라 노드 안에 색을 채웁니다.
오후	혼공머(혼자 공부하는 머신러닝) 05-2. 교차검증과 그리드 서치 교차 검증은 검증 세트를 떼어 내어 평가하는 과정을 여러 번 반복합니다. 그다음 이 점수를 평균하여 최종 검증 점수를 얻습니다. 이 과정을 그림으로 보면 이해가 쉽습니다. 다음은 3-폴드 교차 검증 그림입니다.  • **검증 세트**는 하이퍼파라미터 튜닝을 위해 모델을 평가할 때, 테스트 세트를 사용하지 않기 위해 훈련 세트에서 다시 떼어 낸 데이터 세트입니다. • **교차 검증**은 훈련 세트를 여러 폴드로 나눈 다음 한 폴드가 검증 세트의 역할을 하고 나머지 폴드에서는 모델을 훈련합니다. 교차 검증은 이런 식으로 모든 폴드에 대해 검증 점수를 얻어 평균하는 방법입니다. • **그리드 서치**는 하이퍼파라미터 탐색을 자동화해 주는 도구입니다. 탐색할 매개변수를 나열하면 교차 검증을 수행하여 가장 좋은 검증 점수의 매개변수 조합을 선택합니다. 마지막으로 이 매개변수 조합으로 최종 모델을 훈련합니다. • **랜덤 서치**는 연속된 매개변수 값을 탐색할 때 유용합니다. 탐색할 값을 직접 나열하는 것이 아니고 탐색 값을 샘플링할 수 있는 확률 분포 객체를 전달합니다. 지정된 횟수만큼 샘플링하여 교차 검증을 수행하기 때문에 시스템 자원이 허락하는 만큼 탐색량을 조절할 수 있습니다.

scikit-learn

• cross_validate()는 교차 검증을 수행하는 함수입니다.

첫 번째 매개변수에 교차 검증을 수행할 모델 객체를 전달합니다. 두 번째와 세 번째 매개변수에 특성과 타깃 데이터를 전달합니다.

scoring 매개변수에 검증에 사용할 평가 지표를 지정할 수 있습니다. 기본적으로 분류 모델은 정확도를 의미하는 'accuracy', 회귀 모델은 결정계수를 의미하는 'r2'가 됩니다.

cv 매개변수에 교차 검증 폴드 수나 스플리터 객체를 지정할 수 있습니다. 기본값은 5입니다. 회귀일 때는 KFold 클래스를 사용하고 분류일 때는 StratifiedKFold 클래스를 사용하여 5-폴드 교차 검증을 수행합니다.

n_jobs 매개변수는 교차 검증을 수행할 때 사용할 CPU 코어 수를 지정합니다. 기본값은 1로 하나의 코어를 사용합니다. -1로 지정하면 시스템에 있는 모든 코어를 사용합니다.

return_train_score 매개변수를 True로 지정하면 훈련 세트의 점수도 반환합니다. 기본값은 False입니다.

• GridSearchCV는 교차 검증으로 하이퍼파라미터 탐색을 수행합니다. 최상의 모델을 찾은 후 훈련 세트 전체를 사용해 최종 모델을 훈련합니다.

첫 번째 매개변수로 그리드 서치를 수행할 모델 객체를 전달합니다. 두 번째 매개변수에는 탐색할 모델의 매개변수와 값을 전달합니다.

scoring, cv, n_jobs, return_train_score 매개변수는 cross_validate() 함수와 동일합니다.

• RandomizedSearchCV는 교차 검증으로 랜덤한 하이퍼파라미터 탐색을 수행합니다. 최상의 모델을 찾은 후 훈련 세트 전체를 사용해 최종 모델을 훈련합니다.

첫 번째 매개변수로 그리드 서치를 수행할 모델 객체를 전달합니다. 두 번째 매개변수에는 탐색할 모델의 매개변수와 확률 분포 객체를 전달합니다.

scoring, cv, n_jobs, return_train_score 매개변수는 cross_validate() 함수와 동일합니다.

혼공머(혼자 공부하는 머신러닝)
05-3. 트리의 앙상블

- **앙상블 학습**은 더 좋은 예측 결과를 만들기 위해 여러 개의 모델을 훈련하는 머신러닝 알고리즘을 말합니다.
- **랜덤 포레스트**는 대표적인 결정 트리 기반의 앙상블 학습 방법입니다. 부트스트랩 샘플을 사용하고 랜덤하게 일부 특성을 선택하여 트리를 만드는 것이 특징입니다.
- **엑스트라 트리**는 랜덤 포레스트와 비슷하게 결정 트리를 사용하여 앙상블 모델을 만들지만 부트스트랩 샘플을 사용하지 않습니다. 대신 랜덤하게 노드를 분할해 과대적합을 감소시킵니다.
- **그레이디언트 부스팅**은 랜덤 포레스트나 엑스트라 트리와 달리 결정 트리를 연속적으로 추가하여 손실 함수를 최소화하는 앙상블 방법입니다. 이런 이유로 훈련 속도가 조금 느리지만 더 좋은 성능을 기대할 수 있습니다. 그레이디언트 부스팅의 속도를 개선한 것이 **히스토그램 기반 그레이디언트 부스팅**이며 안정적인 결과와 높은 성능으로 매우 인기가 높습니다.

scikit-learn

- RandomForestClassifier는 랜덤 포레스트 분류 클래스입니다.

n_estimators 매개변수는 앙상블을 구성할 트리의 개수를 지정합니다. 기본값은 100입니다.

criterion 매개변수는 불순도를 지정하며 기본값은 지니 불순도를 의미하는 'gini'이고 'entropy'를 선택하여 엔트로피 불순도를 사용할 수 있습니다.

max_depth는 트리가 성장할 최대 깊이를 지정합니다. 기본값은 None으로 지정하면 리프 노드가 순수하거나 min_samples_split보다 샘플 개수가 적을 때까지 성장합니다.

min_samples_split은 노드를 나누기 위한 최소 샘플 개수입니다. 기본값은 2입니다.

max_features 매개변수는 최적의 분할을 위해 탐색할 특성의 개수를 지정합니다. 기본값은 auto로 특성 개수의 제곱근입니다.

bootstrap 매개변수는 부트스트랩 샘플을 사용할지 지정합니다. 기본값은 True입니다.

oob_score는 OOB 샘플을 사용하여 훈련한 모델을 평가할지 지정합니다. 기본값은 False입니다.

n_jobs 매개변수는 병렬 실행에 사용할 CPU 코어 수를 지정합니다. 기본값은 1로 하나의 코어를 사용합니다. -1로 지정하면 시스템에 있는 모든 코어를 사용합니다.

- ExtraTreesClassifier는 엑스트라 트리 분류 클래스입니다.

  n_estimators, criterion, max_depth, min_samples_split, max_features 매개변수는 랜덤 포레스트와 동일합니다.

  bootstrap 매개변수는 부트스트랩 샘플을 사용할지 지정합니다. 기본값은 False입니다.

  oob_score는 OOB 샘플을 사용하여 훈련한 모델을 평가할지 지정합니다. 기본값은 False입니다.

  n_jobs 매개변수는 병렬 실행에 사용할 CPU 코어 수를 지정합니다. 기본값은 1로 하나의 코어를 사용합니다. -1로 지정하면 시스템에 있는 모든 코어를 사용합니다.

- GradientBoostingClassifier는 그레이디언트 부스팅 분류 클래스입니다.

  loss 매개변수는 손실 함수를 지정합니다. 기본값은 로지스틱 손실 함수를 의미하는 'deviance'입니다.

  learning_rate 매개변수는 트리가 앙상블에 기여하는 정도를 조절합니다. 기본값은 0.1입니다.

  n_estimators 매개변수는 부스팅 단계를 수행하는 트리의 개수입니다. 기본값은 100입니다.

  subsample 매개변수는 사용할 훈련 세트의 샘플 비율을 지정합니다. 기본값은 1.0입니다.

  max_depth 매개변수는 개별 회귀 트리의 최대 깊이입니다. 기본값은 3입니다.

- HistGradientBoostingClassifier는 히스토그램 기반 그레이디언트 부스팅 분류 클래스입니다.

  learning_rate 매개변수는 학습률 또는 감쇠율이라고 합니다. 기본값은 0.1이며 1.0이면 감쇠가 전혀 없습니다.

  max_iter는 부스팅 단계를 수행하는 트리의 개수입니다. 기본값은 100입니다.

  max_bins는 입력 데이터를 나눌 구간의 개수입니다. 기본값은 255이며 이보다 크게 지정할 수 없습니다. 여기에 1개의 구간이 누락된 값을 위해 추가됩니다.