8주차 예습과제_정윤지

Chapter 05 회귀

01 회귀 소개

회귀 분석
- 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 통계학 기법
- 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법을 통칭
- 독립변수의 값에 영향을 미치는 회귀계수
- 머신러닝 관점에서 독립변수는 피처에 해당되며 종속변수는 결정 값
- 머신러닝 회귀 예측의 핵심은 주어진 피처와 결정 값 데이터 기반에서 학습을 통해 최적의 회귀 계수를 찾아내는 것

회귀 계수가 선형이면 선형 회귀, 비선형이면 비선형 회귀
독립변수의 개수가 한 개인지 여러 개인지에 따라 단일 회귀, 다중 회귀로 나뉨

지도학습은 분류와 회귀로 나뉨
분류는 예측 값이 카테고리와 같은 이산형 클래스 값이고 회귀는 연속형 숫자 값

선형 회귀가 가장 많이 사용되며 선형 회귀는 실제 값과 예측 값의 차이를 최소화하는 직선형 회귀선을 최적화하는 방식

선형 회귀 모델은 규제 방법에 따라 다시 별도의 유형으로 나뉠 수 있음
규제는 일반적인 선형 회귀의 과적합 문제를 해결하기 위해 회귀 계수에 페널티 값을 적용하는 것

대표적인 선형 회귀 모델
- 일반 선형 회귀: 예측 값과 실제 값의 RSS를 최소화할 수 있도록 회귀 계수를 최적화, 규제를 적용하지 않음
- 릿지: 선형 회귀에 L2 규제를 추가한 회귀 모델
- 라쏘: 선형 회귀에 L1 규제를 적용한 방식
- 엘라스틱넷: L2, L1 규제를 함께 결합한 모델
- 로지스틱 회귀: 분류에 사용되는 선형 모델, 강력한 분류 알고리즘, 텍스트 분류와 같은 영역에 사용

02 단순 선형 회귀를 통한 회귀 이해

단순 선형 회귀
- 독립변수 1개, 종속변수 1개인 선형 회귀
- 1차 함수식으로 모델링이 가능
- 독립변수가 1개인 단순 선형 회귀에서는 기울기와 절편을 회귀 계수로 지칠
- 실제 값과 회귀 모델의 차이에 따른 오류 값을 남은 오류, 잔차라고 부름
- 최적의 회귀 모델을 만드는 것은 전체 데이터의 잔차 합이 최소가 되는 모델을 만든다는 의미, 동시에 오류 값 합이 최소가 될 수 있는 최적의 회귀 계수를 찾는다는 의미
- 오류 값을 계산할 때 절댓값을 취해 더하거나, 오류 값의 제곱을 구해, 더하는 방식을 취함

- RSS 방식으로 오류 합을 구함

$$Error^2 = \text{RSS}$$

- RSS를 최소로 하는 회귀 계수를 학습을 통해 찾는 것이 머신러닝 기반 회귀의 핵심 사항
- 일반적으로 RSS 학습 데이터의 건수로 나누어서 다음과 같이 정규화된 식으로 표현

$$RSS(w_0, w_1) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 * x_i))^2$$

(i는 1부터 학습 데이터의 총 건수 N까지)

- 회귀에서 RSS는 비용이며 w 변수(회귀 계수)로 구성되는 RSS를 비용 함수라고 함
- 최종적으로 감소하지 않는 최소의 오류 값을 구하는 것
- 비용 함수를 손실 함수라고도 함

03 비용 최소화하기 - 경사 하강법

경사 하강법
- 고차원 방정식에 대한 문제를 해결하면서 비용 함수 RSS를 최소화하는 방법을 직관적으로 제공
- 점진적으로 반복적인 계산을 통해 W 파라미터 값을 업데이트하면서 오류 값이 최소가 되는 W 파라미터를 구하는 방식
- 반복적으로 비용 함수의 반환 값, 즉 예측 값과 실제 값 차이가 작아지는 방향성을 가지고 W 파라미터를 지속해서 보정
- 오류 값이 더 이상 작아지지 않으면 그 오류 값을 최소 비용으로 판단, 그때의 W 값을 최적 파라미터로 반환

$$\mathrm{R(w)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 * x_i))^2$$

- R(W)를 미분하여 미분 함수의 최솟값을 구해야 하므로 편미분을 적용

$$\frac{\partial R(w)}{\partial w_1} = \frac{2}{N}\sum_{i=1}^{N} -x_t * (y_i - (w_0 + w_1 x_i)) = -\frac{2}{N}\sum_{i=1}^{N} x_i * (\text{실제값}_i - \text{예측값}_i)$$

$$\frac{\partial R(w)}{\partial w_0} = \frac{2}{N}\sum_{i=1}^{N} -(y_i - (w_0 + w_1 x_i)) = -\frac{2}{N}\sum_{i=1}^{N} (\text{실제값}_i - \text{예측값}_i)$$

- 편미분 값이 너무 클 수 있기 때문에 보장 계수를 곱하는데 이를 학습률이라고 함

$$w_1 = \text{이전}$$

$$w_1 + \eta\frac{2}{N}\sum_{i=1}^{N} x_i * (\text{실제값}_i - \text{예측값}_i),\ \text{새로운}\ w_0 = \text{이전}\ w_0 + \eta\frac{2}{N}\sum_{i=1}^{N} (\text{실제값}_i - \text{예측값}_i)$$

을 반복적으로 적용하면서 비용 함수가 최소가 되는 값을 찾음

- 경사 하강법의 일반적인 프로세스

- Step 1: w_1, w_0를 임의의 값으로 설정하고 첫 비용 함수의 값을 계산합니다.
- Step 2: w_1을 $w_1 + \eta\frac{2}{N}\sum_{i=1}^{N} x_i * (\text{실제값}_i - \text{예측값}_i)$, w_0을 $w_0 + \eta\frac{2}{N}\sum_{i=1}^{N} (\text{실제값}_i - \text{예측값}_i)$으로 업데이트한 후 다시 비용 함수의 값을 계산합니다.
- Step 3: 비용 함수가 감소하는 방향성으로 주어진 횟수만큼 Step 2를 반복하면서 w_1과 w_0를 계속 업데이트합니다.

04 사이킷런을 이용한 보스턴 주택 가격 예측

LinearRegression 클래스 - Ordinary Least Squares
- 예측 값과 실제 값의 RSS를 최소화해 OLS 추정 방식으로 구현한 클래스
- fit() 메서드로 X,y 배열을 입력받으면 회귀 계수인 W를 coef_ 속성에 저장

- 회귀 계수 계산은 입력 피처의 독립성에 많은 영향을 받음
- 다중 공선성: 피터 간의 상관관계가 매우 높은 경우 분산이 매우 커져, 오류에 민감해짐

회귀 평가 지표
- 실제 값과 회귀 예측 값의 차이 값을 기반으로 한 지표가 중심
- MAE: 실제 값과 예측 값의 차이를 절댓값으로 변환해 평균한 것
- MSE: 실제 값과 예측 값의 차이를 제곱해 평균한 것
- RMSE: MSE에 루트를 씌운 것
- R^: 분산 기반으로 예측 성능을 평가

LinearRegression을 이용해 보스턴 주택 가격 회귀 구현

05 다항 회귀와 과(대)적합/과소적합 이해

다항 회귀 이해
- 회귀가 독립변수의 단항식이 아닌 2차, 3차 방정식과 같은 다항식으로 표현되는 것을 다항 회귀라고 함
- 선형 회귀
- 다항 회귀 곡선형으로 표현한 것이 더 예측 성능이 높음
- 비선형 함수를 선형 모델에 적용시키는 방법을 사용해 구현
- 사이킷런은 polynomialFeatures 클래스를 통해 피처를 다항식 피처로 변환

다항 회귀를 이용한 과소적합 및 과적합 이해
- 다항 회귀의 차수를 높일수록 학습 데이터에만 너무 맞춘 학습이 이루어져 정작 데스트 데이터 환경에서는 오히려 예측 정확도가 떨어짐
- 실습

편향 분산 트레이드 오프
- 저편향 저분산은 예측 결과가 실제 결과에 매우 잘 근접하면서도 예측 변동이 크지 않고 특정 부분에 집중되어 있는 뛰어난 성능을 보여줌
- 저편향 고분산은 예측 결과가 실제 결과에 비교적 근접하지만 예측 결과가 실제 결과를 중심으로 꽤 넓은 부분에 분포되어 있음
- 고편향 저분산은 정확한 결과에서 벗어나면서도 예측이 특정 부분에 집중되어 있음
- 고편향 고분산은 정확한 예측 결과를 벗어나면서도 넓은 부분에 분포되어 있음
- 편향을 낮추고 분산을 높이면서 전체 오류가 가장 낮아지는 골디락스 지점을 통과하면서 분산을 지속적으로 높이면 전체 오류 값이 오히려 증가하면서 예측 성능이 다시 저하됨
- 편향과 분산이 서로 트레이드 오프를 이루면서 오류 cost 값이 최대로 낮아지는 모델을 구축하는 것이 가장 효율적인 머신러닝 예측 모델

06 규제 선형 모델 - 릿지, 라쏘, 엘라스틱넷

규제 선형 모델의 개요
- 비용 함수는 학습 데이터의 잔차 오류 값을 최소로 하는 RSS 최소화 방법과 과적합을 방지하기위해 회귀 계수 값이 커지지 않도록 하는 방법이 서로 균형을 이루어야 함

$$\text{비용 함수 목표} = \text{Min}(\text{RSS(W)} + \text{alpha} * \| W \|_2^2)$$

- 알파는 학습 데이터 적합 정도와 회귀 계수 값의 크기 제어를 수행하는 튜닝 파라미터
- 알파를 0에서부터 지속적으로 증가시키면 회귀 계수 값의 크기를 감소시킬 수 있음
- 알파 값으로 페널티를 부여해 회귀 계수 값의 크기를 감소시켜 과적합을 개선하는 방식을 규제라고 함
- 규제는 L1, L2 방식으로 구분되는데 L2 규제를 적용한 회귀를 릿지 회귀, L1 규제를 적용한 회귀를 라쏘 회귀라고 함

릿지 회귀
- 실습

라쏘 회귀
- 실습

엘라스틱넷 회귀
- L2 규제와 L1 규제를 결합한 회귀
- 라쏘 회귀가 서로 상관관계가 높은 피처들의 경우 이들 중 중요 피처만을 셀렉션하고 다른 피처들은 모두 회귀 계수를 0으로 만드는 성향이 강하며 이로 인해 알파값에 따라 회귀 계수의 값이 급격히 변동할 수 있기 때문에 L2 규제를 라쏘 회귀에 추가한 것
- 수행시간이 상대적으로 오래 걸림
- 실습

선형 회귀 모델을 위한 데이터 변환
- 선형 회귀 모델은 일반적으로 피처와 타깃값 간에 선형의 관계가 있다고 가정, 최적의 선형 함수를 찾아내 결괏값을 예측
- 왜곡을 예방하기 위해 데이터에 대한 스케일링/정규화 작업을 수행하는 것이 일반적
- 심하게 왜곡됐을 경우 변환 작업을 수행
- 타깃 값의 경우 일반적으로 로그 변환을 적용: 원복이 어려울 수 있음
- 실습

07 로지스틱 회귀

선형 회귀 방식을 분류에 적용한 알고리즘
- 선형 함수의 회귀 최적선을 찾는 것이 아니라 시그모이드 함수 최적선을 찾고 이 시그모이드 함수의 반환 값을 확률로 간주해 확률에 따라 분류를 결정
- 시그모이드 함수는 x 값이 아무리 커지거나 작어져도 y 값은 항상 0과 1사이 값을 반환
- X 값이 커지면 1에 근사, x값이 작아지면 0에 근사
- 실습

08 회귀 트리
- 트리 기반의 회귀: 회귀 트리를 이용
- 회귀를 위한 트리를 생성, 이를 기반으로 회귀 예측을 진행
- 피처가 단 하나인 x 피처 데이터 세트와 결정값 y가 2차원 평면상에 존재한다고 하면
- X 피처를 결정 트리 기반으로 분할, x값의 균일도를 반영한 지니 계수에 따라 분할 가능
- Split0 기준으로 분할, split1, split2 규칙 노드로 분할이 가능, split3 규칙노드로 변환
- 리프 노드 생성 기준에 부합하는 트리 분할이 완료되었다면 피츠 노드에 소속된 데이터 값의 평균값을 구해 최종적으로 리프 노드에 결정 값으로 할당
- 실습