5장

1. 회귀 소개

회귀는 여러 개의 독립변수와 한 개의 족속변수 간의 상관관계를 모델링하는 기법

- 회귀 분석은 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 기법

회귀 유형 – 선형/비선형 여부, 독립변수의 개수, 종속변수의 개수에 따라 나뉨

선형 회귀는 실제 값과 예측값의 차이를 최소화하는 직선형 회귀선을 최적화하는 방식

2. 단순 선형 회귀를 통한 회귀 이해

최적의 회귀 모델을 만든다는 것은 전체 데이터의 잔차 합이 최소가 되는 모델을 만드는 것. 동시에 오류 값 합이 최소가 될 수 있는 최적의 회귀 계수를 찾는 것.

오류값을 계산하는 방법: MAE – 절댓값을 취해서 더하기, RSS – 오류 값의 제곱을 구해서 더하기

RSS는 비용이며 w 변수로 구성되는 RSS는 비용 함수

3. 비용 최소화하기 – 경사 하강법

경사 하강법은 고차원 방정식에 대한 문제를 해결해 주면서 비용 함수 RSS를 최소화하는 방법을 직관적으로 제공한다

- 점진적으로 반복적인 계산을 통해 W 파라미터 값을 업데이트하면서 오류 값이 최소가 되는 W 파라미터를 구하는 방식

경사 하강법의 프로세스

1. w1, w0를 임의의 값으로 설정하고 첫 비용 함수의 값을 계산
2. w1를 업데이트한 후 다시 계산
3. 비용 함수의 값이 감소했으면 2번 반복. 감소하지 않으면 그때의 w1, w0 값을 구한다.

4. 사이킷런 LinearRegression을 이용한 보스턴 주택 가격 예측

5. 다항 회귀와 과(대)적합/과소적합 이해

다항 회귀: 회귀가 독립변수의 단항식이 아닌 2차, 3차 방정식과 같은 다항식으로 표현되는 것

- 다항 회귀는 선형 회귀이다.

편향-분산 트레이드오프

- 매우 단순화되어 지나치게 한 방향성으로 치우이면 고편향성, 학습 데이터 하나하나의 특성을 반영하면서 매우 복잡한 모델이 되고 지나치게 높은 변동성을 가지면 고분산
- 편향과 분산은 한 쪽이 높으면 한 쪽이 낮아진다.
- 높은 편향/낮은 분산에서 과소적합되기 쉽고, 낮은 편향/높은 분산에서 과적합되기 쉽다.
- 편향과 분산이 서로 트레이드오프를 이루면서 오류 cost 값이 최대로 낮아지는 모델을 구축하는 것이 가장 효율적인 머신러닝 예측 모델을 만드는 방법

6. 규제 선형 모델 – 릿지, 라쏘 엘라스틱넷

릿지 클래스의 주요 생성 파라미터는 alpha이며, 이는 릿지 회귀의 alpha L2 규제 계수에 해당된다.

라쏘는 W의 절댓값에 페널티를 부여하는 L1 규제를 선형 회귀에 적용한 것이다.

- L2 규제가 회귀 계수의 크기를 감소시키는 데 반해 L1 규제는 불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거한다.

엘라스틱넷 회귀는 L2 규제와 L1 규제를 결합한 회귀. 그래서 수행시간이 상대적으로 오래 걸린다.

7. 로지스틱 회귀
8. 회귀 트리