Machine learning

Hyeonho Lee

2018년 11월 6일

Contents

1 선형회귀

1. 선형회귀는 양적 반응변수를 예측하는 유용한 도구이다.
2. 중요한 질문들…
    1) X와 Y사이에 상관관계가 있는가
    2) X와 Y사이에 얼마나 강한 상관관계가 있는가
    3) 여러 X들 중 Y에 기여하는 X는?
    4) Y에 대한 각 X 효과를 얼마나 정확하게 추정할 수 있는가
    5) 미래의 Y에 대해 얼마나 정확하게 예측할 수 있는가
    6) 상관관계는 선형인가
    7) X들 사이에 시너지 효과가 있는가(상호작용 항)

1.1 단순선형회귀

1. 단순선형회귀는 매우 간단한 기법으로, 하나의 설명변수 X에 기초하여 양적 반응변수 Y를 예측한다. 이 기법은 X와 Y 사이에 선형적 상관관계가 있다고 가정한다. 수학적으로 선형적 상관관계는 다음과 같이 나타낸다.

$$Y \approx \beta_0 + \beta_1 + \varepsilon$$

2. 계수 추정
    1) 실제로 β_0와 β_1은 알려져 있지 않다. 그러므로 $Y \approx \beta_0 + \beta_1 + \varepsilon$을 사용하여 예측하기 전에 데이터를 이용하여 계수를 추정해야 한다.
    2) n의 데이터 포인트의 개수라고 할 때, n개의 데이터 포인트에 가능한 한 가깝게 되도록 하는 절편 $\hat{\beta_0}$와 기울기 $\hat{\beta_1}$을 찾고자 한다.
    3) 가까움(closeness)을 측정하는 방법은 여러 가지가 있으나, 대표적으로는 최소제곱 기준을 최소화하는 것이다.
    4) $RSS = e_1^2 + e_2^2 + ... + e_n^2$이며, RSS는 잔차제곱합이라고 칭한다. 잔차란, $e_i = y_i - \hat{y_i}$을 칭한다.
    5) $RSS = (y_1 - \hat{\beta_0} - \hat{\beta_1} x_1)^2 + (y_2 - \hat{\beta_0} - \hat{\beta_1} x_2)^2 + ... + (y_n - \hat{\beta_0} - \hat{\beta_1} x_n)^2$으로 다시 나타낼 수 있고, 미적분을 사용하여 수식을 정리하면
    6) $\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$와 $\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$임을 알 수 있다. 추정치 쌍 $(\hat{\beta_0}, \hat{\beta_1})$는 RSS를 최소화하는 값임을 알 수 있다.
3. 계수 추정값의 정확도 평가
    1) X와 Y의 실제 상관관계는 어떤 알려지지 않은 함수 $f$dp eogo $Y = f(x) + \varepsilon$의 형태를 가지며 ε은 평균이 영인 랜덤오차항이다. 만약 f가 선형함수로 근사된다면 이 관계는 $Y = \beta_0 + \beta_1 X + \varepsilon$이라고 할 수 있다.($\beta_0$는 절편이고 즉 X=0일 때 Y의 기대값이고, β_1은 기울기이고 X의 한 유닛 증가에 연관된 Y의 평균 증가임을 알 수 있다.), 오차항의 존재는 단순한 모델로 나타낼 때 수반되는 여러 가지 한계를 위한 것이다.
    2) 오차항의 존재는 매우 중요하다. X와 Y의 실제 관계는 선형적이지 않을 수 있고, Y값의 변화를 초래하는 다른 변수들이 있을 수 있으며, 측정 오차가 있을 수 있다.(오차항은 보통 X와 독립이라고 가정한다.)
    3) 모회귀선과 최소제곱선 사이의 차이는 매우 작고 구별하기 어려울 수 있다. 자료가 하나밖에 없는데 두 개의 다른 직선이 설명변수와 반응변수의 상관관계를 기술하는 것은 무엇을 의미할까…근본적으로 이 두 직선의 개념은 표본의 정보를 사용하여 큰 모집단의 특징을 추정하는 표준통계적 방법의 확장이다. 어떤 확률변수 Y의 모평균 μ를 알고자 한다고 해보면 μ는 알려져 있지 않다. 그러나 우리는 Y의 n개 관측치를 알 수 있고, 이것을 사용하여 μ를 추정할 수 있다. 합리적인 추정값은 $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$이다. 이것을 표본평균이라 부른다.
    4) 선형회귀와 확률변수의 평균값 추정 비유는 bias의 개념에서 보면 적절하다. 표본평균 $\hat{\mu}$를 사용하여 μ를 추정한다면, $\hat{\mu}$은 평균적으로 μ와 동일하다고 기대된다는 점에서 이 추정값은 편향되지 않은 것이다. 이것은 어떤 하나의 특정 관측치셋에서는 과대추정할 수 있고, 또 다른 관측치셋에 대해서는 과소추정할 수 있다는 것을 의미한다. 그러나 아주 많은 관측치셋으로부터 얻은 μ의 추정값들을 평균할 수 있으면 이 평균값은 μ와 정확하게 동일한 값이 될 것이다. 그러므로, 비편향 추정량은 실제 파라미터를 조직적으로 과대추정 또는 과소추정하는 것이 아니다.
    5) 비편향성질 - 이것은 최소제곱계수추정에 대해서도 성립한다. 특정 데이터셋에 대해 β_0와 β_1을 추정하면 그 추정값을 true β_0와 β_1과 일치하지는 않을 것이다. 그러나 아주 많은 수의 데이터셋에 대해 얻은 추정값들을 평균할 수 잇으면 이 추정값들의 평균값을ㄴ 정확하게 일치할 것이다. 다른 데이터셋으로부터 추정된 최소제곱선들의 평균은 실제 모회귀선에 매우 근접한다.

6) 하나의 $\hat{\mu}$는 μ를 상당히 과소추정 과대추정한다는 것을 알 수 있다. 그렇다면 얼마나 다를 것인가? 일반적으로 이 질문에 대한 답은 $SE(\hat{\mu})$으로 표현하는 $\hat{\mu}$의 표준오차를 계산하는 것이다. 표준오차의 식은 대체로 $Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$ 이다.(평균에 대한 표준오차)
7) 그렇다면 $\hat{\beta_0}$와 $\hat{\beta_1}$의 표준오차는 어떻게 계산할까? $SE(\hat{\beta_0})^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]$, $SE(\hat{\beta_1})^2 = [\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]$으로 구할 수 있다.
8) 위 β에 대한 표준오차 식들이 유효하려면 각 관측치에 대한 오차 ε_i가 곱옹의 분산 σ^2과 무상관이라는 가정이 필요하다.
9) 표준오차는 주로 계수들에 대한 가설검정을 하는 데 사용될 수 있다. (H0 : X와 Y 사이에 상관관계가 없다. H1 : X와 Y사이에 어떤 상관관계가 있다.) 수학적으로 이 가설은 $\beta_1 = 0$과 $\beta_1 \neq 0$ 인지를 검정하는 것과 같다.

4. 모델의 정확도 평가

1) 잔차표준오차(RSE)
2) R^2 통계량

1.2 다중선형회귀

1.3 회귀모델에서 다른 고려할 사항

1.4 마케팅 플랜

1.5 선형회귀와 KNN의 비교

2 선형모델 선택 및 Regularization

subset(부분집합) 선택

Shrinkage 방법

차원축소 방법

고차원의 고려

3 선형성을 넘어서(비선형성)

다항식회귀

계단함수

기저함수

회귀 스플라인

평활 스플라인

국소회귀

일반화가법모델

4 트리 기반의 방법

의사결정트리의 기초

배깅, 랜덤 포레스트, 부스팅