# 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 $\beta_0$와 $\beta_1$은 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 $\varepsilon$은 평균이 영인 랜덤오차항이다. 만약 $f$가 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, $\beta_1$은 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 $\mu$를 알고자 한다고 해보면 $\mu$는 알려져 있지 않다. 그러나 우리는 Y의 n개 관측치를 알 수 있고, 이것을 사용하여 $\mu$를 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 $\mu$를 추정한다면, $\hat{\mu}$은 평균적으로 $\mu$와 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 $\mu$의 추정값들을 평균할 수 있으면 이 평균값은 $\mu$와 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 $\beta_0$와 $\beta_1$을 추정하면 그 추정값을 true $\beta_0$와 $\beta_1$과 일치하지는 않을 것이다. 그러나 아주 많은 수의 데이터셋에 대해 얻은 추정값들을 평균할 수 있으면 이 추정값들의 평균값은 정확하게 일치할 것이다. 다른 데이터셋으로부터 추정된 최소제곱선들의 평균은 실제 모회귀선에 매우 근접한다.

6) 하나의 $\hat{\mu}$는 $\mu$를 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 $\beta$에 대한 표준오차 식들이 유효하려면 각 관측치에 대한 오차 $\varepsilon_i$가 곧옹의 분산 $\sigma^2$과 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) 귀무가설을 기각하고 대립가설을 채택했다면, 모델이 데이터에 적합한 정도를 수량화하고자 할 것이다. 선형회귀적합의 질은 보통 잔차표준오차(RSE)와 $R^2$를 사용하여 평가한다.
2) 잔차표준오차(RSE) = $\sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i}^2)}$이며, 각 관측치에 오차항 $\varepsilon$이 관련되어 있다. 이러한 오차항 때문에 실제 회귀선을 알아도 X로부터 Y를 정확하게 예측할 수 없을 것이다. RSE는 $\varepsilon$의 표기준편차에 대한 추정값으로, 대략 반응변수 값이 실제 회귀선으로부터 벗어나게 될 평균값을 의미한다.
3) $R^2$ 통계량 = $\frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$ 이며, RSE의 데이터에 대한 모델의 적합성결여를 나타내주는 절대적 측도가 되는 것과는 다르게, Y의 단위로 측정되므로 적정한 RSE가 무엇인지 항상 명확환 것은 아니다. 적합도에 대한 다른 측도를 제공하며, 설명된 분산의 비율형태를 나타낸다(0과 1사이의 값만을 가진다.)
4) $R^2$는 RSE에 비해 해석이 쉽다는 장점이 있다. 왜냐하면, RSE와는 달리 그 값이 항상 0과 1사이에 있기 때문이다. 좋은 $R^2$값이 무엇인지에 대한 결정은 어렵지만, 일반적으로 응용에 따라 다르다. 또한, $R^2$은 X와 Y 사이의 선형상관관계에 대한 측도이다. 다음과 같이 정의되는 상관계수도 X와 Y 사이의 선형상관관계의 측도이다. $Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}^2)}}$ 이다. 이것은 선형모델의 적합성을 평가하기 위해 $R^2$ 대신 $r = Cor(x,Y)$를 사용할 수 도 있음을 의미한다. 단순선형회귀에서 $R^2 = r^2$임을 보여줄 수 있다. 다중선형회귀에서는 통용되지 않는 개념이나, 변수쌍 사이의 연관성을 수량화 하기 때문에 $R^2$을 다르게 접근한다.

## 1.2 다중선형회귀

1. 단순선형회귀는 단일 설명변수를 기반으로 반응변수를 예측하는 유용한 기법이다. 하지만 실제로는 보통 하나보다 많은 설명변수가 관련된다. 두 개의 추가적인 설명변수를 포함하기 위해 Y에 대한 분석을 어떻게 확장할 것인가. 한가지 방법은 매우 단순한 방법이다. 각각의 X에 대해 단순선형 회귀를 사용하는 것이다. 하지만 이 방법은 만족할만한 방식이 아니다. 우선 X들에 대해 Y를 예측하는 것이 어떻게 예측하는지 명확하지 않다. 왜냐하면 서로 다른 회귀방정식에 연관되어 잇기 때문이다. 두번째로 각각의 회귀계수를 추정하는 데 다른 X를 고려하지 않는다. 만약 여러개의 X들 중 $X_1$과$X_2$가 상관되어 있으면 Y에게 영향을 미치는 것이 다르기 때문이다. 그러므로 단순선형회귀를 확장하여 다중선형회귀를 사용한다. 이것은 하나의 모델에서 각 설명변수에 다른 기울기 계수를 할당하면 된다. $Y = \beta_0 + \beta_1 X_1 ... + \beta_p X_p + \varepsilon$는 다중선형회귀모델이며, $\beta_j$는 다른 설명변수들은 변동되지 않을 때 $X_j$의 한 유닛 증가가가 Y에 미치는 평균 효과로 해석된다.

2. 회귀계수의 추정은 단순선형회귀와 같이 최소제곱법을 사용하여 추정할 수 있다. $RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n}(y_i = \hat{\beta_0} - \hat{\beta_1}x_{i1} - ... - \hat{\beta_p}x_{ip})$로 RSS를 최소화 하도록 $\beta_0, \beta_1, \beta_2, ..., \beta_p$를 선택한다. 또한 단순선형회귀와는 다르게 다중선형회귀추정값은 다소 복잡한 형태를 가지며 가장 쉬운 표현방식은 행렬대수를 사용하는 것이다. 또한 단순선형회귀는 상관관계가 있음을 나타낼수도 있다. 그러나 다중회귀는 그 반대결과를 보일 수도 있다.(해변에서 파는 아이스크림과 상어의 공격 그리고 온도에 대한 문제)

3. 몇가지의 중요한 것들

1) 설명변수들 $X_1, X_2, .., X_p$ 중 적어도 하나는 반응변수를 예측하는 데 유용한가 : 단순선형회귀에서는 단순히 $\beta_1 = 0$ 인지 검사하면 결정 할 수 있다. 그러나 다중회귀에서는 H0 : $\beta_1 = \beta_2 = ... = \beta_p = 0$, H1 : 적어도 하나의 $\beta_j$는 영이 아니다. 로 이루어진다. 이 가설은 F통계량을 계산하면서 이루어진다. ($F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)}$). p value와 F통계량에 대한 이슈가 있다. p value는 변수들과 반응변수 사이에 어떤 상관관계가 있는지 잘못 결론 내릴 가능성이 매우 높다. 하지만 F통계량은 설명변수의 개수를 조정하므로 이런 문제가 없다. 따라서, 만약 H0이 참이면, 설명변수의 개수 또는 관측횟수에 상관없이 F통계량의 p value가 0.05보다 작아지게될 가능성은 단지 5%이다.

2) Y를 설명하는 데 모든 설명변수들이 도움이 되는가? 또는 설명변수들의 일부만이 유용한가(중요 변수의 결정) : 위의 내용처럼 다중회귀분석의 첫 번째 단계는 F-통계량을 계산하여 관련된 p-값을 살펴보는 것이다. 만약 p value에 근거하여 적어도 하나의 설명변수는 반으변수와 상관성이 있다는 결론에 도달한다면 그 설명변수가 어느 것인지 궁금할 것이다. 그러나 p가 크다면 잘못된 결론에 도달할 가능성이 높다. 이 때 어느변수가 반응변수와 상관성이 있는지 결정하는 것을 변수선택이라고 한다. 변수선택과 더불어 어느 모델이 최고인지 계산하는 지표는 여러가지 지표가 있다. Mallows $C_p$, AIC, BIC, Adjusted $R^2$ 가 포함된다. 모든 모델을 계산하는 방법은 $2^p$ 의 계산량이지만 모든 걸 계산할 수 없기에 효율적이고 고전적인 방법 3가지가 있다. 전진선택법, 후진소거법, 단계별방법이 있다.
   (1) 전진선택법
   (2) 후진소거법
   (3) 단계별방법
3) 모델은 데이터에 얼마나 잘 맞는가(모델 적합)
4) 주어진 설명변수 값들에 대해 어떤 반응변수 값을 예측해야 하고 그 예측은 얼마나 정확한가

4.

## 1.3 선형다중회귀의 기본 가정(중요)

1) 회귀모형은 모수에 대해 선형인 모형이다. $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\varepsilon_i$
2) 독립변수 $X_1i, X_2i$는 비확률이다.(nonstochastic)
3) 오차항의 평균은 영이다. $E(\epsilon_i)=0$
4) 오차항의 분산은 모든 관찰지에 대해 $\sigma^2$의 일정한 분산을 갖는다.(등분산성 : homoskedasity = Var($\epsilon_i$)
5) 서로 다른 오차항은 상관이 없다. : $Cov(\epsilon_i, \epsilon_j) = 0$, 오차항은 서로 독립적이며, 그들의 공분산은 0이다
6) 오차항은 각 독립변수와 독립적이다. : $E(X_i, \epsilon_i) = 0$
7) 오차항이 정규분포를 따른다.
8) 여기부터는 다중회귀의 가정이다.
9) 독립변수간에는 정확한 선형관계가 없다.
10) 관측된 값들의 수는 독립변수의 수보다 최소한 2는 커야한다.

## 1.4 회귀모델에서 다른 고려할 사항

1. 질적 설명변수
2. 선형모델의 확장
3. 잠재적 문제

## 1.5 마케팅 플랜

## 1.6 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