Elements of Statistical Learing

Week 2: Linear Regression and Related Methods Part 1

Linear Regression 모델에 대해서 알아본다. penalty가 없는 LR 모델에서 시작하여 이후에는 Ridge Regression 등을 다룬다. notation을 정리해보자. (자세한 증명은 생략. 회귀때 이미 다 배운 내용)

- Input: $X = (X_1, \cdots, X_p)$
- Output: Y
- 관측하는 데이터는 $Y = f(X) + \epsilon = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ 에서 generate 되었다고 가정한다.
- Model: $E[Y \mid X] = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$, 즉 $f(X)$는 실제로는 모르는 unknown function의 형태지만, 'linear'라고 가정하는 것이다.

linear 가정은 가장 간단한 가정이다. 이를 충족시키는 데이터가 많지는 않겠지만, 선형에 데이터를 잘 적합만 한다면 다른 non-linear 모델보다 해석력에서 강력한 우위를 점한다. 또한 데이터가 적을 때, non-linear 모델을 적합한다면, 그 데이터에는 잘 맞겠지만 새로운 데이터에 대해서는 poor prediction을 할 우려가 있다. 즉, Bias-Variance 측면에서 Bias를 줄이는데 너무 초점이 맞추었기 때문에 Variance을 잘 잡아내지 못하는 것이다. 이러한 상황에서는 linear model을 사용하는 것이 더 나은 방법일 수도 있다. 모수 $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_1)$는 MSE(RSS)를 최소화하여 찾는다.

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = ||(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||^2 \tag{1}$$

(1)은 행렬 notation으로 적은 것이며, $\boldsymbol{\beta}$에 대한 convex function이므로, $\boldsymbol{\beta}$에 대한 1차, 2차 partial derivatives을 구하여 $\hat{\boldsymbol{\beta}}$를 찾는다.

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{'}\mathbf{X}\boldsymbol{\beta} + 2\mathbf{X}^{'}\mathbf{y} \ \leftarrow normal\ equation$$

$$\frac{\partial^2 RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{'}} = 2\mathbf{X}^{'}\mathbf{X} \ \leftarrow N.N.$$

$$\therefore \hat{\boldsymbol{\beta}} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y}$$

몇 가지 $\hat{\boldsymbol{\beta}}$와 관련된 성질을 정리한다. (증명은 생략)

- $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\boldsymbol{y}$. 여기서 $\mathbf{H}$를 hat matrix라고 부르고, $\mathbf{H}$는 symmetric, idempotent한 projection matrix이다.

- residual vector는 $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. $\mathbf{I} - \mathbf{H}$도 projection matrix이고, $\mathbf{H}$가 span하는 space와 orthogonal한 공간에 prejection한다.
- $\hat{\boldsymbol{\beta}} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- σ^2에 대한 estimator는 $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N - p - 1)$, 즉 MSE이다. $E[MSE] = \sigma^2$, 즉 unbiased estimator이기도 하다.
- $\hat{\boldsymbol{\beta}}$와 $\hat{\sigma}^2$는 독립이다. 이는 나중에 F 통계량을 유도할 때 사용된다.

Gauss Markov Theorem

- For any given $\mathbf{a}$, let $\hat{\theta}^{LSE} = \mathbf{a}^T\hat{\boldsymbol{\beta}}^{LSE}$ be the LSE of $\mathbf{a}^T\boldsymbol{\beta}$. If we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$, that is unbiased for $\mathbf{a}^T\boldsymbol{\beta}$, that is, $E(\mathbf{c}^T\mathbf{y}) = \mathbf{a}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, then $Var(\hat{\theta}^{LSE}) \leq Var(\tilde{\theta})$
- Gauss Markove Theorem은 우선, unbiased estimator에 한정된 이야기이다. biased estimator는 논의에서 제외한다.
- unbiased 하고 linar한 모든 estimators 중, LSE가 분산이 가장 작다. 그런데, unbiased인 estimator이므로 bias가 0이고, 따라서 분산이 가장 작다는 것은 MSE가 가장 작다는 것이다.
- 하지만 bias를 살짝 허용하고, variance을 크게 줄이는 estimator도 있다. 따라서 unbiased 하다고 무조건 좋은 것은 아니다.