기존의 최소제곱법을 이용한 회귀는 다중공선성과 같은 문제가 발생했을 때, 베타의 해가 없거나 하나 이상이 나오는, 유일해가 나오지 않는 ill-posed problem을 마주할 수 있다. 이런 경우 OLS의 결과는 데이터가 과적합 될 수 있다. 따라서 계수를 줄여주는 정규화와 동시에 비편향성을 일부 희생하여 분산을 줄이는 것을 목표로 다음과 같은 방법들이 나오게 되었다.

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left\{ (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \right\}$$

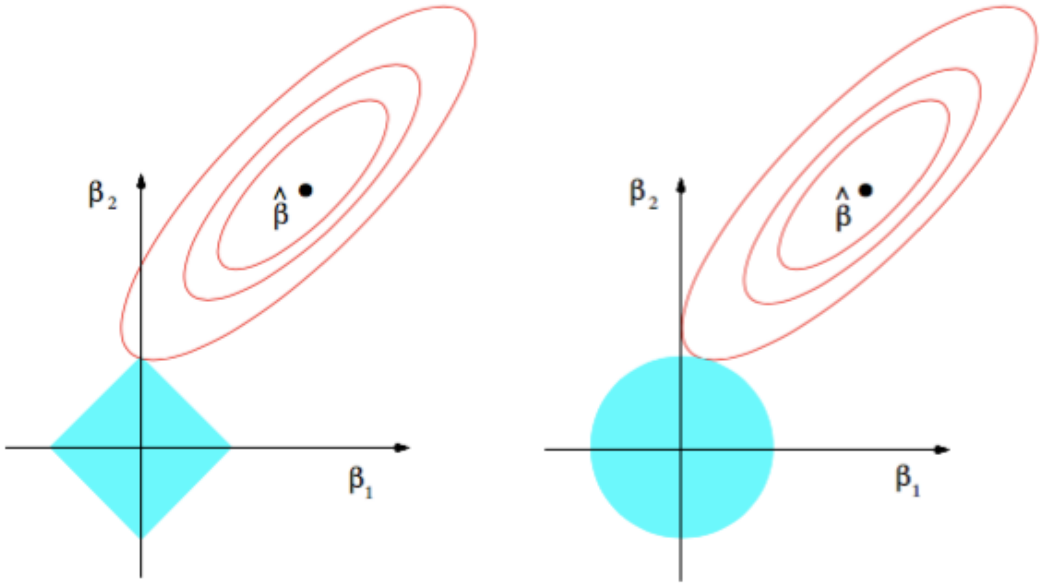
릿지 회귀는 위와 같은 손실함수를 지닌다. 이는 기존의 최소제곱법에 tuning parameter λ와 β^2을 곱한 penalty term을 더한 형태이다. Penalty term은 β의 합에 제약을 걸어 β들의 합이 일정 수준 이하가 되도록 만들고, 이를 통해 값이 매우 커 과적합을 야기하거나, 혹은 중요하지 않은 β들의 값이 줄어들게 만들어 OLS에 비해 overfitting가 해소된 결과를 보인다.

위 식에서 나타난 penalty term은 $\sum \beta^2 < t$ 와 같은 효과를 보이는데, 이처럼 β^2의 합을 이용한 제약을 L2 정규화라고 부른다. 아래의 그래프에서 이 식의 기하학적 해석을 다루도록 하겠다.

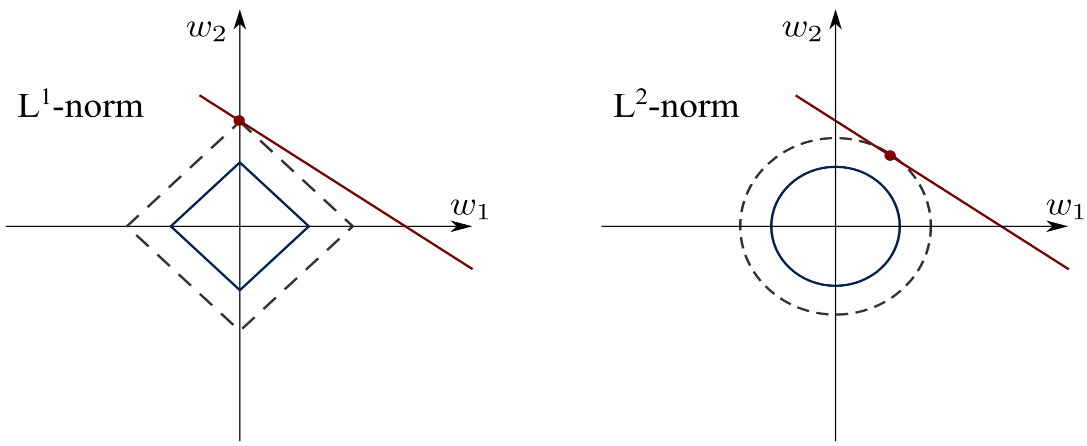
$$\hat{\beta}^{Lasso} = \min_{\beta} \left\{ (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_1 \right\}$$

라소 회귀는 위과 같은 손실함수를 가진다. 릿지 회귀처럼 기존의 최소제곱법에 penalty term을 더했지만 β^2이 아닌 |β|를 더하게 된다. 마찬가지로 β 합의 제약에 의해 β값은 줄어들게 된다. 이 때 비주요 변수는 0까지 줄어들게 된다. 이 과정을 통해 data fitting 문제를 해결하는 동시에 비주요 변수를 0으로 만들어 variable selection까지 동시에 실행하게 된다.

위 식에서 나타난 penalty term은 $\sum |\beta| < t$ 와 같은 효과를 보이는데, 이처럼 |β| 이용한 제약을 L1 정규화라고 부른다. 마찬가지로 아래의 그래프에서 다루도록 하겠다.



1



2

위 그림들을 보면 L1 과 L2의 차이를 알 수 있다. 1번 그림을 보면, 타원의 중심에 찍힌 β hat이 OLS를 통해 구한 최소제곱법의 해가 된다. 이 해는 비편향성을 만족시킨다. 그러나 L1 L2 제약으로 인해 β들의 합은 일정 값 이하여야 한다는 조건을 만족해야 한다. 이를 만족하는 범위가 파란색 원 마름모와 원이다. L1 정규화은 절대값을 사용하기 때문에 마름모의 형태로 나타났고, L2 정규화는 제곱을 사용하기 때문에 원으로 나타난다. 이 범위를 만족하면서, 동시에 편향을 최소로 유지하기 위해 기존의 β hat과 가장 가까운 점이 새로운 제약조건을 만족하는 해가 된다. 이 때 L1은 꼭지점에서 만나기 때문에 β가 0까지 줄어들 수 있는 것이고, L2는 β 범위가 원으로 나타나기 때문에 0까지 줄어들지는 못한다. 이를 p차원까지 확장시키면 된다.

비교하자면,

릿지와 라쏘 모두 penalty term을 이용하여 정규화하며, 분산을 줄여 과적합을 줄여준다. λ값이 클수록 제약이 강해지고, 0에 가까울수록 제약이 줄어들어 OLS와 비슷하게 변화한다.

릿지는 L2 정규화을 사용하며, 모든 변수가 남아있게 되어 모든 정보를 가지고 있지만, 모델 복잡성을 줄여주지 못해 설명에 문제가 생길 수 있다.

라쏘는 L1정규화를 사용하며, 계수가 0까지 줄어들기 때문에 변수를 줄여주는 변수선택까지 진행해준다. 이를 통해 모델의 복잡성까지 줄여줄 수 있다. 그러나 원래의 정보가 손실된다는 단점도 동시에 존재한다.

일반적으로 위 두 방법들 중 하나의 성능이 다른 하나를 압도하는 경우는 쉽게 관찰되지는 않지만(Frank and Friedman, 1993), 변수들 그룹 간의 상관관계가 클 경우 라쏘는 그룹을 무시하고 그룹의 변수들 중 하나만 남기고 나머지의 계수를 0으로 만드는 경향성이 있다. p>n일 경우, 라쏘는 n개의 변수만을 고르는 경향이 있고(Hui Zou, 2005), n>p이면서 변수들 간의 상관관계가 높다면, ridge의 성능이 lasso를 뛰어넘는 경우가 주로 관찰된다(Tibshirani, 1996).

$$\hat{\beta}^{enet} = \min_{\beta} \left\{ (y - X\beta)^T (y - X\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^T \beta \right\}$$

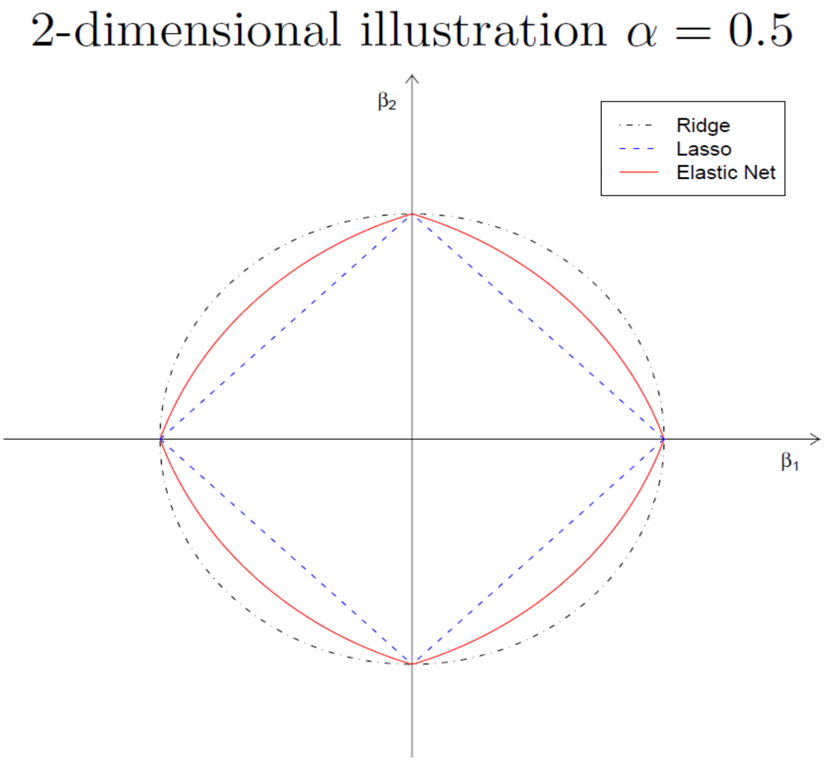
엘라스틱 넷은 위와 같은 손실함수를 지닌다. 이는 L1과 L2 정규화을 모두 사용한다. 앞서 나타난 여러 가지 제약들을 해결하기 위해 고안되었다. 따라서 라쏘를 보완하여 p>n일 경우에도 성능에 제약이 걸리지 않고, 변수들 간의 상관관계가 클 경우에 변수들끼리 그룹을 지어서 그룹이 모델에 동시에 포함되게 하고, 계수들을 거의 비슷하게 맞춘다. 이를 통해 L1 정규화를 보완해준다.

λ1과 λ2의 비율에 따라 L1과 L2 정규화에 각각 가까워지도록 조절할 수 있으며 λ1이 1이고 λ2가 0일 때는 L1, λ1이 0이고 λ2가 1일 때는 L2와 동일해진다.

일반적으로 엘라스틱 넷은 자료의 크기가 클 때 더 좋은 성능을 보여주는 것이 알려져 있다

.

아래 그림은 $\lambda 1$과 $\lambda 2$의 비율이 0.5와 0.5로 나누어져 있을 때의 제약조건의 범위를 보여준다.



tuning parameter λ는 cross validation을 이용하여 구하는 것이 일반적이고 잘 알려진 방법이다. 주로 K-fold cross validation이 사용되는데, 원래의 데이터를 k개의 데이터로 분할한 뒤 각각을 test set으로 사용하고, 선택되지 않은 나머지를 training set으로 사용하여 오차의 평균을 계산한다. Elastic net의 경우 구해야 하는 parameter가 2개 이기 때문에 주로 $\lambda 2$를 고정한 뒤(예시로, 0, 0.01, 0.1, 1, 10, 100) 분산을 최소로 하는 $\lambda 1$을 구해준다.

참고자료

"How Correlations Influence Lasso Prediction" Mohamed Hebiri and Johannes C. Lederer

Regularization and variable selection via the elastic net (Zui,2005)
https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

https://brunch.co.kr/@itschloe1/11

https://en.wikipedia.org/wiki/Lasso_(statistics)

https://en.wikipedia.org/wiki/Tikhonov_regularization

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

https://web.stanford.edu/~hastie/TALKS/enet_talk.pdf

https://en.wikipedia.org/wiki/Elastic_net_regularization