Computational
Data Science LAB

# Foundation of Machine Learning 3주차

정재헌, 우지수 / 2023.02.03

| | |
|---|---|
| 관련문서 | Bloomberg ML EDU |
| 논의사항 및 의문사항 | The Lasso as a Quadratic Program을 사용하는 이유 |
| 기타사항 | |

# Foundation of Machine Learning 3주차

정재헌, 우지수 / 2023.02.03

Computational Data Science LAB

# CONTENTS

1. Tikhonov and Ivanov regularization
2. $\ell_1, \ell_2$ regularization
3. Why does Lasso regression give sparse solution?
4. Finding the Lasso solution
5. Elastic net

# 01 | Tikhonov and Ivanov regularization

## Hypothesis Spaces

- Hypothesis spaces
  - ✓ Hypothesis spaces는 nested sequence를 가지게 됨
  - ✓ $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \cdots \subset \mathcal{F}$
- Complexity measure
  - ✓ Number of variables / features
  - ✓ Depth of a decision tree
  - ✓ Degree of polynomial
- nested hypothesis spaces에서 Complexity measure
  - ✓ Complexity measure $\Omega : \mathcal{F} \to [0, \infty)$
  - ✓ 복잡도가 최대 r인 F의 모든 함수를 고려하는 식: $\mathcal{F}_r = \{f \in \mathcal{F} | \Omega(f) \leq r\}$

# 01 | Tikhonov and Ivanov regularization

## Ivanov regularization vs Tikhonov regularization

- Constrained ERM (Ivanov regularization)
  - ✓ complexity measure가 $\Omega$: $\mathcal{F} \rightarrow [0, \infty)$이고 고정된 complexity $r \geq 0$을 만족할 때:
  - ✓ $\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i) \quad , s.t. \Omega(f) \leq r$
  - ✓ 이 때 r은 validation data또는 cross-validation을 사용해 선택한다(가장 좋은 결과값을 선택)
- Penalized ERM (Tikhonov regularization)
  - ✓ complexity measure가 $\Omega$: $\mathcal{F} \rightarrow [0, \infty)$이고 고정된 $\lambda \geq 0$을 만족할 때:
  - ✓ $\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega f(x)$

# 02 | $\ell_1, \ell_2$ regularization
## Linear Least Squares Regression

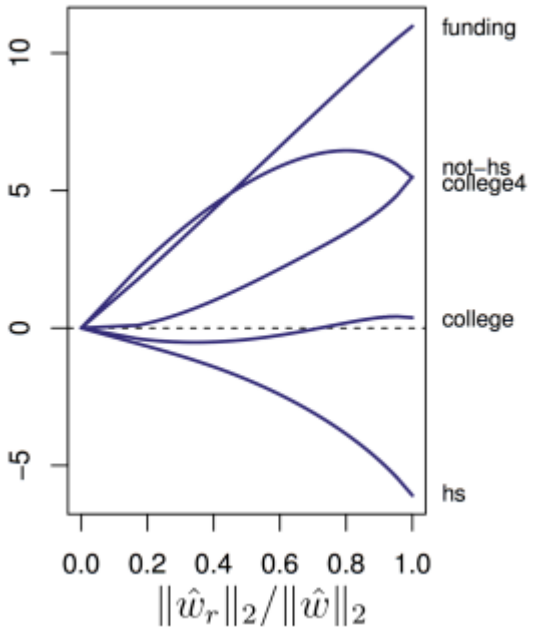
- Linear Least Squares Regression
  - ✓ 선형모델 $\mathcal{F} = \{f: R^d \to R | f(x) = w^T x \; for \; w \in R^d\}$
  - ✓ 이 때 loss: $\ell(\hat{y}, y) = (y - \hat{y})^2$
  - ✓ ∴ Linear least squares regression은 $\ell$에서 F에 대한 ERM:

$$\hat{w} = \underset{w \in R^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{w^T x_i - y_i\}^2$$

# 02 | $\ell_1, \ell_2$ regularization
## Ridge Regression

- Ridge Regression
  - ✓ Tikhonov Form: $\hat{w} = \underset{w \in R^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2 \qquad s.t. \lambda \geq 0$
  - ✓ Ivanov Form: $\hat{w} = \underset{\|w\|_2^2 \leq r^2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \qquad s.t. r \geq 0$
  - ✓ $\ell_2\ norm$: $\|w\|_2^2 = w_1^2 + \cdots + w_d^2$
  - ✓ Ridge는 전체적인 변수에 대한 가중치도 줄이고 불필요한 변수에 대한 가중치를 0에 **가깝게** 줄임

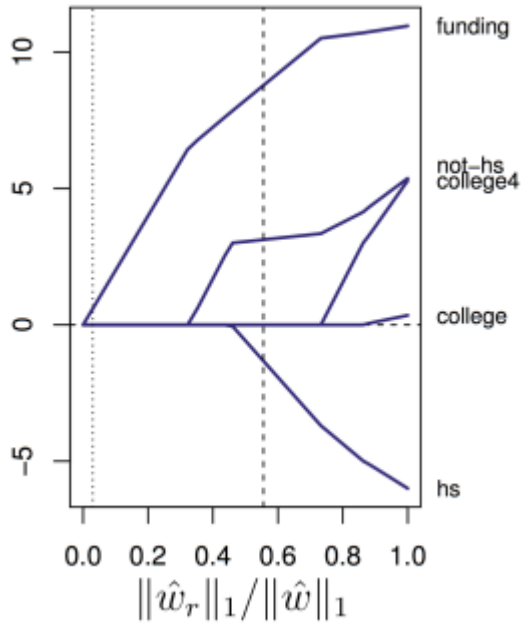
[Regularization Path]



# 02 | $\ell_1, \ell_2$ regularization
## Lasso Regression

- Lasso Regression
  - ✓ Tikhonov Form: $\hat{w} = \underset{w \in R^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1 \; s.t. \lambda \geq 0$
  - ✓ Ivanov Form: $\hat{w} = \underset{\|w\|_1 \leq r}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \; s.t. r \geq 0$
  - ✓ $\ell_1 \; norm$: $\|w\|_1 = |w_1| + \cdots + |w_d|$
  - ✓ LASSO는 불필요한 변수에 대한 가중치를 완전히 0으로 억압하여 변수선택 기능 제공

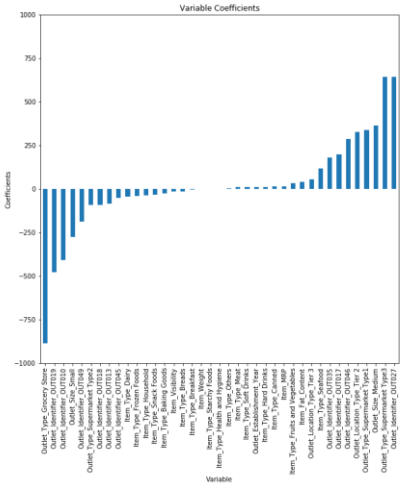
[Regularization Path]



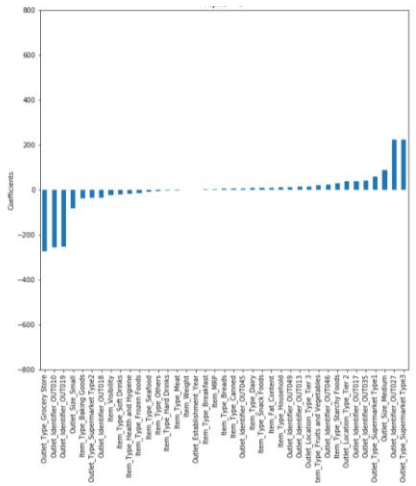
# 02 | $\ell_1, \ell_2$ regularization

## Ridge vs Lasso

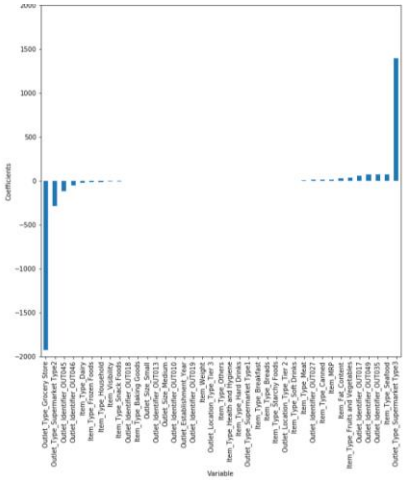
- Ridge Regression vs Lasso Regression




[Original]




[Ridge]



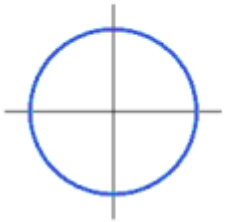

[Lasso]

# 03 | Why does Lasso regression give sparse solution?

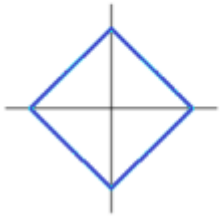
## The $\ell_1$ and $\ell_2$ Norm Constraints

- The $\ell_1$ and $\ell_2$ Norm Constraints
  - ✓ 2차원의 $input\ space$에서 $\mathcal{F} = \{f(x) = w_1\, x_1 + w_2\, x_2\}$인 매개변수 $linear\ hypothesis\ space$ 가 존재한다고 가정

- $\ell_2$ contour: $w_1^2 + w_2^2 = r$



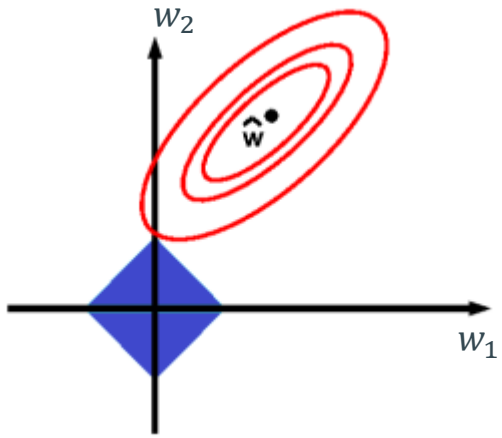
- $\ell_1$ contour: $|w_1| + |w_2| = r$



# 03 | Why does Lasso regression give sparse solution?

## The $\ell_1$ and $\ell_2$ Norm Constraints

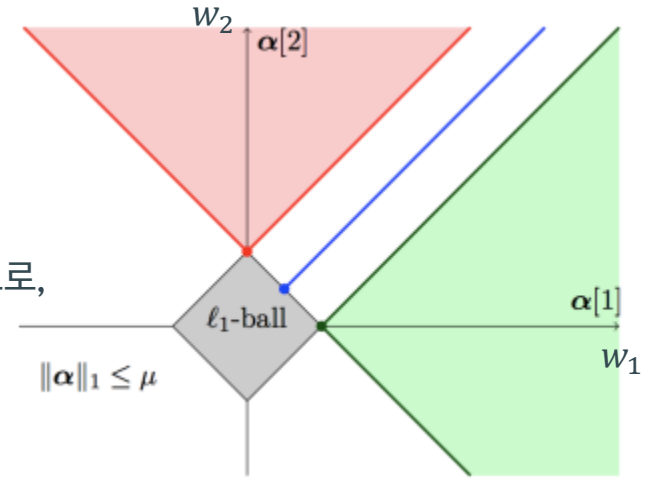
- The $\ell_1$ Regularization
  - ✓ 이 상황에서 L1 contour line에 대해 Lasso regularization을 적용한다고 하면 그림과 같은 상황이 나옴
  - ✓ 파란색 부분은 복잡성 제약 조건(complexity constraint)을 충족하는 영역: $|w_1| + |w_2| \leq r$
  - ✓ 빨간 선은 $\hat{R}_n(w) = \Sigma_{i=1}^{n} \left(w^T x_i - y_i\right)^2$의 등고선
  - ✓ $w$를 찾아 나가면서, constraint를 만족할 때까지 반복을 한 모습을 보여줍니다.
  - ✓ (같은 빨간 라인에 있을 경우 같은 Loss 값, 바깥쪽의 빨간 라인이 더 큰 Loss 값)



# 03 | Why does Lasso regression give sparse solution?

## The $\ell_1$ and $\ell_2$ Norm Constraints

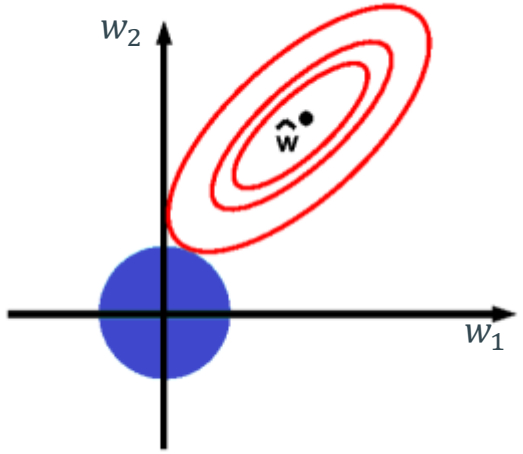
- Why are Lasso Solutions Often Sparse?
  - ✓ 이 그림은 결국 빨간 area나 초록 area에서 constrained를 만족할 때 $w_1, w_2$ 둘 중 하나는 0이 되므로, Least square을 만족하면서도 sparsity 가 나오기 쉽다는 점을 보여줌
  - ✓ 빨간 영역과 초록 영역은 모서리에서 constrained solution이 되는 OLS(최소자승법)의 해의 영역



# 03 | Why does Lasso regression give sparse solution?

## The $\ell_1$ and $\ell_2$ Norm Constraints

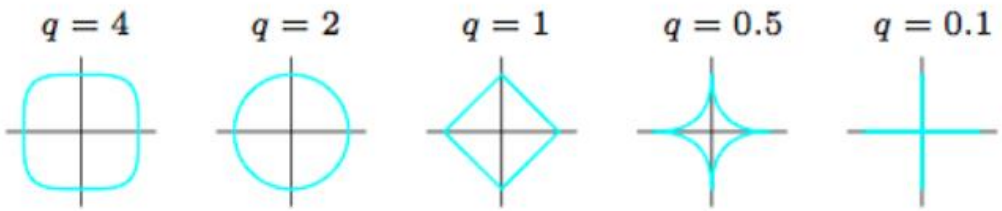
- The $\ell_2$ Regularization
  - ✓ Ridge의 경우, 그림에서 볼 수 있듯이, feature 둘 중 하나가 0이 되기 매우 힘든 구조



# 03 | Why does Lasso regression give sparse solution?

## The $(\ell_q)^q$ Constraint

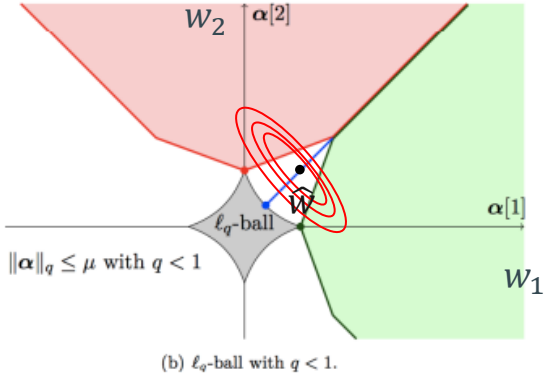
- The $(\ell_q)^q$ Constraint
  - ✓ $\ell_q$를 일반화한 식은 : $\left(||w||_q\right)^q = |w_1|^q + |w_2|^q$이다
  - ✓ $\mathcal{F} = \{f(x) = w_1 x_1 + w_2 x_2\}$
  - ✓ $||w||_q^q = |w_1|^q + |w_2|^q$의 등고선들은 다음과 같다.



# 03 | Why does Lasso regression give sparse solution?

## The $(\ell_q)^q$ Constraint

- Q가 0.5인 경우(1보다 작을 때)
  - ✓ Sparsity에 대해서는 만족할지 모르지만($w_1, w_2$ 둘중 하나가 0이 될 확률이 더 높음, red/green region이 더 넓음)
  - ✓ 오목한 부분에서 $w$의 loss가 더욱 증가하므로 최적화가 잘 안됨



(b) $\ell_q$-ball with $q < 1$.

# 04 | Finding the Lasso Solution

## How to find the Lasso solution?

- Lasso function의 objective function
  - ✓ $\min_{w \in R^d} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1$
  - ✓ $\|w\|_1 = |w_1| + |w_2|$ 은 미분 불가

# 04 | Finding the Lasso Solution

## Splitting a Number into Positive and Negative Parts

- Splitting a Number into Positive and Negative Parts
  - ✓ Consider any number $a \in \mathrm{R}$
  - ✓ Let the positive part of a be: $a^+ = a1(a \geq 0)$
  - ✓ Let the negative part of a be: $a^- = -a1(a \leq 0)$
  - ✓ $a = a^+ - a^-$
  - ✓ $|a| = a^+ + a^-$

| $a$ | $a^+$ | $a^-$ |
|---|---|---|
| 1 | 1 | 0 |
| -5 | 0 | 5 |
| 0 | 0 | 0 |
| 3 | 3 | 0 |
| -4 | 0 | 4 |

# 04 | Finding the Lasso Solution

## How to find the Lasso solution?

- The Lasso problem
  - ✓ $\min_{w \in R^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \left|\left|w\right|\right|_1$
  - ✓ $w$벡터는 두개의 벡터로 나누어지게 됨
  - ✓ $w^+ = (w_1^+, \ldots, w_d^+)$, $w^- = (w_1^-, \ldots, w_d^-)$,

# 04 | Finding the Lasso Solution

## How to find the Lasso solution?

- The Lasso problem
  - ✓ $\min_{w^+,w^-} \sum_{i=1}^{n} \left((w^+ - w^-)^T x_i - y_i\right)^2 + \lambda_1^T (w^+ + w^-)$
    $s.t. w_i^+ \geq 0 \; for \; all \; i \; , \qquad w_i^- \geq 0 \; for \; all \; i$
  - ✓ Objective is differentiable
  - ✓ 하지만 w+와 w-는 양수와 음수의 개념이 있기 때문에 이를 각각 a, b로 치환해서 quadratic program을 진행

# 04 | Finding the Lasso Solution

## Coordinate Descent Method

- Coordinate Descent Method(좌표 하강 방법)
  - ✓ 우리의 목적은 $L(w) = L(w_1, \dots, w_d)$를 최소화 시키는 것
  - ✓ 일반적인 Gradient Descent 나 Stochastic Gradient descent 에서는, 매 step 마다 모든 w의 값이 변화,
    하지만 Coordinate Descent 는 한 step마다 한가지 feature만을 변화시킴
  - ✓ 각각의 step에서는 $w_i^{new} = \underset{w_i}{\operatorname{argmin}} L(w_1, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_d)$를 풀어나감
  - ✓ 특정 coordinate(feature)을 정하고, iteration을 반복하여 , 그 feature에 대한 최적의 coefficient를 찾는 알고리즘
  - ✓ Feature을 random 하게 고르면, stochastic coordinate descent라고 하고, 그냥 순차적으로 고르게 되면,
    Cyclic Coordinate descent라고 함

# 04 | Finding the Lasso Solution

## Closed Form Solution of Lasso

- Closed Form Solution of Lasso regression

✓ $\widehat{w_j} = \underset{w_j \in R}{\operatorname{argmin}} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda |w|_1$

✓ $then, \quad w_j \begin{cases} \frac{c_j + \lambda}{a_j}, & if\ c_j < -\lambda \\ 0, & if\ c_j \in [-\lambda,\ \lambda] \\ \frac{c_j - \lambda}{a_j}, & if\ c_j > \lambda \end{cases}$

✓ $a_j = 2 \sum_{i=1}^{n} x_{i,j}^2 \qquad c_j = 2 \sum_{i=1}^{n} x_{i,j} (y_i - w_{-j}^T x_{i,-j})$

✓ $w_{-j}$는 j번째 데이터를 제외한 모든 $w$를 나타냄

✓ $c_j$항에 $(y_i - w_{-j}^T x_{i,-j})$ 부분은 $y_i$를 j번째 항을 제외한 채 예측을 하고, 그 값을 feature의 값과 곱해줌

✓ 따라서 $c_j$는 j번째 feature를 넣는 것이 다른 coefficient에 얼마나 영향을 줄 지에 대한 지표

# 05 | Elastic-net
## Elastic-net Regression

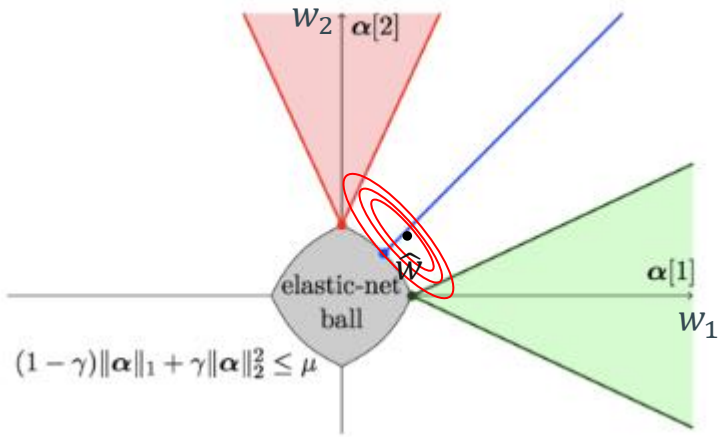
- Elastic-net Regression
  - ✓ $\hat{w} = \underset{w \in R^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$
  - ✓ Lasso와 Ridge의 penalty를 결합한 하이브리드 모델
  - ✓ Elastic-net regression은 변수 선택기능을 가지지 못하는 Ridge regression과 다중공선성이 높으면 좋은 성능을 가지지 못하는 Lasso regression의 단점들을 절충한 모델이기 때문에 변수 선택과 다중공선성이 높은 변수에 대해 좋은 성능을 갖는 장점을 보임
  - ✓ 즉, 변수간 상관관계가 크더라도 중요하지 않은 변수들을 모두 버리면서 중요한 변수들만 잘 골라내어 중요도와 상관관계에 따라 적합한 가중치를 적용할 수 있음

# 05 | Elastic-net

## Elastic-net Regression

- Elastic-net Regression

  ✓ $\hat{w} = \underset{w \in R^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$




  ✓ Elastic-net에서 기하학적 해석을 진행해보면 Lasso regression보다는

  Red/green region이 좁지만 볼록하기에 loss가 적고

  ✓ Ridge regression보다는 볼록한 정도가 적지만 $w_1, w_2$에서 0이 나오는 sparsity를 가질 수 있음

# Q&A

감사합니다.