Recap of Linear Regression / Ordinary Least Squares (OLS)

Problem: Given input / predictor variables $(x_i)_{i=1}^{N}$ and output / response variables $(y_i)_{i=1}^{N}$ (training set)

can we construct a linear model that

- ▷ explains the relationship between $\{x_i\}$ and $\{y_i\}$, and
- ▷ predict the output $\tilde{y}_i$ of new input variables $\tilde{x}_i$? (test set: $(\tilde{x}_i, \tilde{y}_i)_{i=1}^{\tilde{N}}$)

Note: $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^K$ (multiple predictor variables per data point)

OLS: Cost function $C = \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} x_{ik} \beta_k \right)^2 = \| y - X\beta \|_2^2$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad X = \begin{pmatrix} - & x_1 & - \\ & \vdots & \\ - & x_N & - \end{pmatrix} \in \mathbb{R}^{N \times K}, \quad \text{we call } \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} \text{ coefficient vector}$$

We saw: $\hat{\beta} = \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \| y - X\beta \|_2^2 = (X^T X)^{-1} X^T y$

How can we make this more flexible?

Preprocessing of input: Input $x' \in \mathbb{R}^d \longmapsto x := \Phi(x') \in \mathbb{R}^K$, $K > d$

Ex: $d = 10$ predictor variables in diabetes $\longrightarrow$ $K = 55$ predictor variables

Important notions: R^2 score: $R^2 = 1 - \frac{\|y - X\hat{\beta}\|_2^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$, Mean Squared Error: $MSE = \frac{1}{N}\|y - X\hat{\beta}\|_2^2$

Ridge Regression

Fix "regularization parameter" $\alpha > 0$.

Set $$\hat{\beta} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \{ \underbrace{\|y - X\beta\|_2^2 + \alpha \|\beta\|_2^2}_{=: J(\beta)} \} =$$

Compute derivative of $J(\beta)$:

$$\nabla J(\hat{\beta}) = 2X^T(X\hat{\beta} - y) + 2\alpha\hat{\beta} \overset{!}{=} 0$$

$$\Leftrightarrow (2X^TX + 2\alpha \cdot I)\hat{\beta} = 2X^Ty$$

$$\Leftrightarrow \hat{\beta} = \boxed{(X^TX + \alpha I)^{-1} X^T y}$$ (closed form solution)

- If $\alpha = 0$: Linear Regression
- If α large: Coefficients of $\hat{\beta}$ "shrinked to 0" ("less complex model", avoid overfitting)

Sparse Regression

Fix again $\alpha > 0$.

- Useful if interpretability is desired: Which of the K predictor variables "explain" response variables best?

Lasso: $$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^K} \{ \|y - X\beta\|_2^2 + \alpha \underbrace{\|\beta\|_1}_{\text{"} \sum_{i=1}^K |\beta_i| \text{"}} \} \quad (*)$$ "Least absolute shrinkage and selection operator"

⚠ $(*)$ has no closed form solution.

However, can be solved by convex optimization.

- ▷ Behavior for varying α is similar to ridge regression
- ▷ Unlike ridge regression, lasso performs variable selection if α is properly chosen.