1° Linear Basis Model:

Basic linear model:

$$y(x, w) = w_0 + w_1 x_1 + \cdots + w_D x_D = W^T \cdot X$$

Basis function $\phi(\cdot)$:

Basis function is some sort of "fixed preprocessing" of the input data

By using nonlinear basis functions, the linear model can contain some linear information on the input space X.

⇓

Linear basis model:

$$y(x, w) = w_0 + \sum_{i=1}^{M-1} w_i \phi_i(x) = W^T \Phi(x)$$

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ \vdots & & & & \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{MN} \end{bmatrix} \in \mathbb{R}^{M \times N}$$

let $x_1 := \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{M1} \end{bmatrix}$

$\Rightarrow \Phi(\cdot): \quad X \to \mathbb{R}^J$

$$\Rightarrow \Phi(X) = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{J-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{J-1}(x_2) \\ \vdots & & & \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{J-1}(x_N) \end{bmatrix}$$

Examples of Basis functions:

1° Polynomial basis: $\phi_j(x) = x^j$

2° Fourier basis: periodic data, or data with known boundaries.

$$\phi_0(x)=1 \qquad \phi_j(x)=\cos(\omega_j x+\psi_j)$$

3° Radial basis function.

A function that is symmetric around a center and typically decay to 0 when near the center

Gaussian Radial basis:

$$\phi_j(x)=\exp\left\{-\frac{1}{\ell}\|x-\mu\|_2^2\right\}=\exp\left\{-\frac{1}{2\ell}\sum_{d=1}^{D}(x_d-\mu_{j,d})^2\right\}$$

4° Sigmoid basis: (Parametric Soft Thresholds)

$$\phi_j(x)=\frac{1}{1+\exp\{-x^T w_j\}} \Rightarrow \text{close to zero when } x^T w_j \text{ is large negative.}$$

2° How to solve parameters:

2) Maximum Likelihood:

t: the true target variable

$y(x,w)$: our model.

Assumption 1.

$\Rightarrow t=y(x,w)+\varepsilon$. $\varepsilon \sim N\{0,\sigma^2 I\} \sim N\{0,\beta^{-1}\}$, β is the precision.

$\Rightarrow p\{t\mid x,w,\beta\}=N\{t\mid y(x,w),\beta^{-1}\}$.

Given data $X=\{x_1,\dots,x_N\}$.

Assumption 2: different samples are i.i.d.

$$\Rightarrow p\{t\mid X,w,\beta\}=\prod_{n=1}^{N} N\{t_n\mid w^T\phi(x_n),\beta^{-1}\}$$

$$\Rightarrow \ln p\{t|x,w,\beta\} = \sum_{n=1}^{N} \ln N\{t_n | w^T\phi(x_n), \beta^{-1}\}$$

$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta\cdot\frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$$

$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta\cdot E_D(w) \quad \text{square loss function.}$$

$\Rightarrow$ Maximizing MLE $\Leftrightarrow$ Minimizing square loss function.

Thus: $\nabla \ln p\{t|w,\beta\} = \sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}\phi(x_n)^T = 0$

$$\Rightarrow w_{ML} = (\Phi^T\Phi)^{-1}\Phi^T t.$$

b) Least Squares:

objection: minimize square loss.

$$\Rightarrow E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$$

$$= \frac{1}{2}(t - w^T\Phi)^T(t - w^T\Phi)$$

$$\Rightarrow \nabla_w E_D(w) = -\Phi^T t + \Phi^T\Phi w = 0$$

$$\Rightarrow \hat{w} = (\Phi^T\Phi)^{-1}\Phi^T t.$$

$$\text{var}(\hat{w}) = E\{[\hat{w} - E(w)][\hat{w} - E[w]]^T\}$$

$$= \sigma^2(\Phi^T\Phi)^{-1}$$

2° Regularized LS:

General way, modify the objective function by adding a penalty term

$$E_D(w) + \lambda E_W(w)$$

Collinearity: the event of two (or multiple) covariates being strongly related

⇒ The space spanned by super-collinear covariates is a lower-dimensional subspace of the parameter space.

⇒ The design matrix X is rank deficient.

Why collinearity is a Problem

First consider the case that: two (or multiple) covariates being perfectly dependent (super-collinearity)

⇒ X is column rank deficient, column rank $< p$.

∴ there exists a non-trivial $v \in \mathbb{R}^p$ s.t. $Xv = 0_p$

⇒ $X^T X v = X^T 0_p = 0_p$ ⇒ $\det(X^TX) = 0$ ⇒ one or more of the eigenvalues of A is zero.

∵ $v \neq 0_p$

∴ X^TX is not invertible. By spectral decomposition.

$X^TX = \sum_{j=1}^{p} \lambda_j U_j U_j^T$, U_j is the eigenvector and λ_j is the corresponding eigenvalue

⇒ X^TX is singular

⇒ $(X^TX)^{-1} = \sum_{j=1}^{p} \lambda_j^{-1} U_j U_j^T$

⇒ we can't have $w = (X^TX)^{-1}X^Ty$.

⇒ one or more of λ_j must be 0.

Next we consider the case that there isn't an exact linear relationship among the predictors, but the correlations are close to one.

⇒ X^TX is invertible

However:

let $\Sigma \triangleq X^TX$

$\because \Sigma^{-1} = \frac{1}{\det(\Sigma)} \operatorname{adj}(\Sigma)$

Σ is close to singular $\Rightarrow$ $\det(\Sigma) \to 0$

$\therefore$ $\mathrm{Var}(w) = \sigma^2 (X^T X)^{-1}$ is large.

When $p > n$:

the system of equations does not have a unique solution

Conclusion: when X is rank deficient, there are three consequences:

1° w cannot be estimated

2° the variance of the estimation is large

3° the solution of w is not unique.

To solve the above problem, we introduce minimum least squares estimator.

Defn. The minimum least squares estimator of the regression parameter minimizes the sum loss and is of minimum length.

$$\Rightarrow \quad \hat{w}_{MLS} = \arg\min_{w \in \mathbb{R}^p} \| t - \Phi w \|_2^2$$

$$\text{s.t.} \quad \| w \|_{MLS}^2 < \| w \|_2^2 \quad \text{for all } w \text{ that minimizes the sum loss.}$$

$\Rightarrow$ Ridge regression.

Objective function:

$$\min \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2 + \frac{\lambda}{2} w^T \cdot w$$

Ridge estimator: $\hat{w} = (\Phi^T \Phi + \lambda I_p)^{-1} \cdot \Phi^T t$

How can Ridge estimator solve the above problems?

By singular value decomposition:

$$\Phi = U_\phi D_\phi V_\phi^T$$

For OLS:

$$\hat{w} = (\Phi^T \Phi)^{-1} \Phi^T t$$

$$= (V_\phi D_\phi^T D_\phi V_\phi^T)^{-1} V_\phi D_\phi^T U_\phi^T t$$

$$= V_\phi (D_\phi^T D_\phi)^{-1} V_\phi^T V_\phi D_\phi^T U_\phi^T t$$

$$= V_\phi (D_\phi^T D_\phi)^{-1} D_\phi^T U_\phi^T t$$

For Ridge:

$$\hat{w} = (\Phi^T \Phi + \lambda I_{pp})^{-1} \Phi^T t$$

$$= (V_\phi D_\phi^T U_\phi^T U_\phi D_\phi V_\phi^T + \lambda I_{pp})^{-1} V_\phi D_\phi^T U_\phi^T t$$

$$= (V_\phi D_\phi^T D_\phi V_\phi^T + \lambda V_\phi V_\phi^T)^{-1} V_\phi D_\phi^T U_\phi^T t$$

$$= V_\phi (D_\phi^T D_\phi + \lambda I_{pp})^{-1} V_\phi^T V_\phi D_\phi^T U_\phi^T t$$

$$= V_\phi (D_\phi^T D_\phi + \lambda I_{pp})^{-1} D_\phi^T U_\phi^T t$$

For OLS: we have $(D_\phi^T D_\phi)^{-1} D_\phi^T$ in the estimation

$$(D_\phi^T D_\phi)^{-1} D_\phi^T \in \mathbb{R}^{P \times n}$$

$$= \begin{bmatrix} \frac{1}{s_1} & & & 0 \\ & \frac{1}{s_2} & & 0 \\ & & \ddots & \vdots \\ & & \frac{1}{s_p} & 0 \end{bmatrix}$$ ⇒ the reciprocal of the non-zero singular values on the diagonal of the left $p \times p$ matrix

$$(D_\phi^T D_\phi + \lambda I_{pp})^{-1} D_\phi^T \in \mathbb{R}^{P \times n}$$

$$= \begin{bmatrix} \frac{s_1}{s_1^2+\lambda} & & & 0 \\ & \frac{s_2}{s_2^2+\lambda} & & 0 \\ & & \ddots & \vdots \\ & & \frac{s_p}{s_p^2+\lambda} & 0 \end{bmatrix}$$

let d_{jj} denotes the non-zero singular values on the left diagonal:

$$d_{jj}^{-1} = d_{jj} (d_{jj}^2 + \lambda)$$

⇒ Ridge penalty shrinks the singular values

Recall that in the case of super-collinearity, $(\Phi^T\Phi)$ is not invertable because $(\Phi^T\Phi)^{-1} = \sum_{j=1}^{p} (d_{jj}^2)^{-1} \cdot v_j v_j^T$, where one or more of d_{jj} is 0

By Ridge penalty:

$$(\Phi^T\Phi + \lambda I_{pp})^{-1} = \sum_{j=1}^{p} (d_{jj}^2 + \lambda)^{-1} v_j v_j^T$$

If $d_{jj} = 0 \Rightarrow d_{jj}^2 + \lambda > 0$

$\Rightarrow \Phi^T\Phi + \lambda I_{pp}$ is invertable.

we could transform the Ridge regression into an optimization problem:

$$w^* = \underset{\|w\|_2^2 \le \lambda}{\operatorname{argmin}} \|t - \Phi w\|_2^2$$

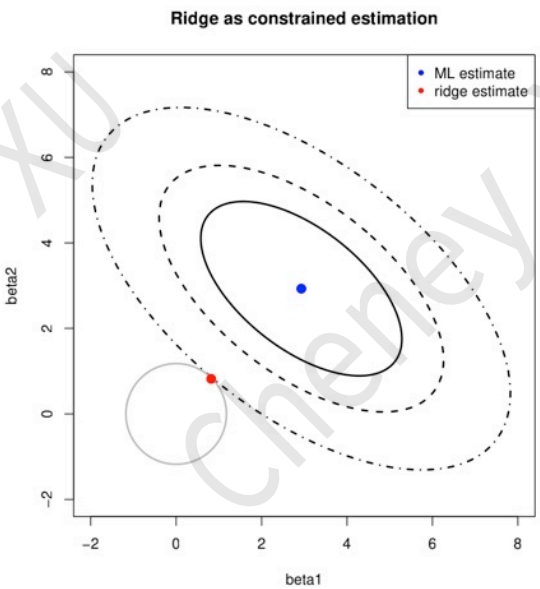
This could be solved by KKT condition:

Lagrangian: $L = \|Y - Xw\|_2^2 + \nu(\|w\|_2^2 - \lambda)$

1st KKT: $\nabla_w L = 0$

2nd KKT: $\nu(\|w(\nu)\|_2^2 - \lambda) = 0$

$\Rightarrow$ Both have solution when $\nu = \|w(\nu)\|_2^2$



Lasso

objective function:

$\|Y - Xw\|_2^2 + \lambda\|w\|_1$

Proposition. Lasso does not have a closed form solution, except when Φ is an orthogonal matrix.

proof: let $w_{OLS} \equiv (\Phi^T\Phi)^{-1}\Phi^T t = \Phi^T t$ (Φ is orthogonal)

$$\min_w \|t - \Phi w\|_2^2 + \lambda \|w\|_1 = \min_w t^T t - t^T\Phi w - w^T\Phi^T t + w^T\Phi^T\Phi w + \lambda\sum_{j=1}^p |w_j|$$

$$\propto \min_w - w_{OLS}^T w - w^T w_{OLS} + w^T w + \lambda|w|$$

$$= \min_{w_1,\dots,w_p} \sum_{j=1}^p \left(-2 w_{j,OLS}\, w_j + w_j^2 + \lambda|w_j|\right)$$

$$= \sum_{j=1}^p \min_{w_1,\dots,w_p} \left(-2 w_{j,OLS}\, w_j + w_j^2 + \lambda|w_j|\right)$$

$\Rightarrow$ The question becomes find w_j for each:

$$\min -2w_{j,OLS}\, w_j + w_j^2 + \lambda|w_j|$$

$$= \min \begin{cases} -2 w_{j,OLS} w_j + w_j^2 + \lambda w_j & \text{if } w_j > 0 \\ -2 w_{j,OLS} w_j + w_j^2 - \lambda w_j & \text{if } w_j < 0 \end{cases}$$

$$\Rightarrow w_j^*(\lambda) = \begin{cases} w_{j,OLS} - \frac{1}{2}\lambda & \text{if } w_j^*(\lambda) > 0 \Rightarrow \text{if } w_{j,OLS} > \frac{1}{2}\lambda \\ w_{j,OLS} + \frac{1}{2}\lambda & \text{if } w_j^*(\lambda) < 0 \Rightarrow \text{if } w_{j,OLS} < -\frac{1}{2}\lambda. \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow w_j^*(\lambda) = \operatorname{sign}(w_{j,OLS})\left(|w_{j,OLS}| - \tfrac{1}{2}\lambda\right)_+$$

How to estimate W in the general case

Proposition 1: Lasso estimator may not be unique.

proof: Considering the case of super-collinearity.

$$t_i = \Phi_{i,*} W + \varepsilon_i$$

If Φ only has two colums, with

$$\Phi_{*,1} = \Phi_{*,2}$$

$u := w_1 + w_2$, $v := w_1 - w_2$

$$\Rightarrow \|t - \Phi w\|_2^2 + \lambda \|w\|_1 = \|t - \Phi_{*1} \cdot u\|_2^2 + \frac{1}{2}\lambda \left(|u+v| + |u-v| \right)$$

$\because$ $|u+v| + |u-v|$ is minimized with respect to v for any v such that $|v| < |u|$

$\Rightarrow$ if $|v| < |u|$:

$$|u+v| + |u-v| = 2|u|$$

$$\Rightarrow \perp = \| t - \Phi_{*,1} u \|_2^2 + \lambda |u|$$

If $\lambda \to 0$, $\perp \to \| t - \Phi_{*,1} u \|_2^2$

$\Rightarrow$ For any $v = w_1 - w_2$, $\perp$ will be the same.

$\Rightarrow$ $w^*(\lambda)$ is not uniquely defined.

Since $w_1^* = \frac{1}{2}\left(\hat{u}(\lambda) + \hat{v}(\lambda) \right)$

$w_2^* = \frac{1}{2}\left(\hat{u}(\lambda) - \hat{v}(\lambda) \right)$

Proposition 2: Though $w^*(\lambda)$ is not uniquely defined, $\Phi \cdot w^*$ is uniquely defined.

proof: Suppose, for contradiction,

$$\exists\ w_\alpha^* , w_\beta^* \quad \text{s.t.}$$

$$\Phi w_\alpha^* \neq \Phi w_\beta^*$$

$$\| t - \Phi w_\alpha^* \|_2^2 + \lambda \| w_\alpha^* \|_1 = \perp = \| t - \Phi w_\beta^* \|_2^2 + \lambda \| w_\beta^* \|_1$$

$\Rightarrow \exists\, \theta \in (0,1)$:

$$\|t - \Phi[(1-\theta)w_\alpha^* + \theta w_\beta^*]\|_2^2 + \lambda \|(1-\theta)w_\alpha^* + \theta w_\beta^*\|_1$$

$$= \|(1-\theta)[t - \Phi w_\alpha^*] + \theta[t - \Phi w_\beta^*]\|_2^2 + \lambda \|(1-\theta)w_\alpha^* + \theta w_\beta^*\|_1$$

$$< (1-\theta)\|t - \Phi w_\alpha^*\|_2^2 + \theta\|t - \Phi w_\beta^*\|_2^2 + \lambda(1-\theta)\|w_\alpha^*\|_1 + \theta\|w_\beta^*\|_1$$

$$= (1-\theta)\mathcal{L} + \theta\mathcal{L} = \mathcal{L}$$

which means, $\exists\, \theta \in (0,1)$, s.t. the cost function of $w = (1-\theta)w_\alpha^* + \theta w_\beta^*$ is lower than $\mathcal{L}$

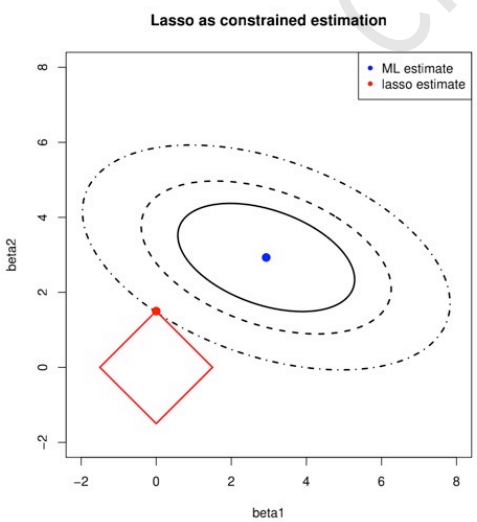
$\Rightarrow$ contradiction!

$\Rightarrow$ $\Phi w_\alpha^* = \Phi w_\beta^*$

$\Rightarrow$ Φw^* is unique.

Lasso as an optimization problem:

$$w^* = \underset{\|w\|_1 < \lambda}{\operatorname{argmin}} \|t - \Phi w\|$$



$\Rightarrow$ the lasso estimation is sparse.

How to estimate?

1° Quadratic Planning.

$$\min_{Rw \geq 0} \frac{1}{2}(t - \Phi w)^T (t - \Phi w)$$

$R \in \mathbb{R}^{q \times p}$: the matrix that specifies the linear constraints on the parameter w

constraints: $\{ w \in \mathbb{R}^p : \|w\|_1 < \lambda_1 \}$

i.e. for $p=2$:

$$\{ w \in \mathbb{R}^2 : w_1 + w_2 < \lambda_1 \} \cap \{ w \in \mathbb{R}^2 : w_1 + w_2 > -\lambda_1 \} \cap \{ w \in \mathbb{R}^2 : w_1 - w_2 > -\lambda_1 \} \cap \{ w \in \mathbb{R}^2 : w_1 - w_2 < \lambda_1 \}$$

$$\Rightarrow R = \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix} \Rightarrow Rw \geq -\lambda_1$$

By Lagrangian:

$$L(w, v) = \frac{1}{2}(t - \Phi w)^T (t - \Phi w) + v^T R w$$

$$\Rightarrow w^* = (\Phi^T \Phi)^{-1} \Phi^T t + (\Phi^T \Phi)^{-1} R^T v$$

Substituting w^* to w

$$\min_{v} \frac{1}{2} v^T R (\Phi^T \Phi)^{-1} R^T v + v^T R (\Phi^T \Phi)^{-1} \Phi^T t + t^T \Phi (\Phi^T \Phi)^{-1} \Phi^T t$$

$$s.t. \quad v \geq 0$$

We could find the optimum v^* of this quadratic planning problem and thus find the optimum w.

2° Gradient Descent.

For the function $f: \mathbb{R}^p \to \mathbb{R}$ at $x \in \mathbb{R}^p$ in the direction of v:

$$f'(x) = \lim_{\tau \to 0} \frac{[f(x+\tau v) - f(x)]}{\tau}$$

for $x \in \mathbb{R}$, $f(x) = |x|$; $\forall u \in \mathbb{R}\setminus\{0\}$, define:

$$f'(x) = \begin{cases} u \cdot \frac{x}{|x|} & \text{if } x \neq 0 \\ u & \text{if } x = 0 \end{cases}$$

limit $u \in \mathbb{R}^p$ to be: $\|u\| = 1$ and u is the direction of the deepest descent:

$$\nabla f(x) = \begin{cases} f'(x) \cdot u_{opt} & \text{if } f(x) \neq 0 \\ 0_p & \text{if } f(x) = 0 \end{cases}$$

$$u_{opt} \in \mathbb{R}^p = \underset{\{u : \|u\|=1\}}{\operatorname{argmax}} f(x)$$

Define: $L_{OLS} = (t - \Phi w)^T (t - \Phi w)$

$$\Rightarrow \frac{\partial}{\partial w_j} L(t, \Phi, w) = \begin{cases} \frac{\partial E}{\partial w}^T (t, \Phi, w) - \lambda \operatorname{sign}(w_j) & \text{if } w_j \neq 0 \\ \frac{\partial E}{\partial w}^T (t, \Phi, w) - \lambda \operatorname{sign}\left[\frac{\partial E}{\partial w}^T (t, \Phi, w)\right] & \text{if } w_j = 0 \text{ and } \left|\frac{\partial E}{\partial w}\right| > \lambda \\ 0 & \text{otherwise.} \end{cases}$$

How to select the Hyperparameter λ?

1° The degree of freedom (dof)

Degrees of freedom denotes how many independent things are there.

The dof of regression:

OLS: $\hat{t} = \Phi (\Phi^T \Phi)^{-1} \Phi^T t$

Hat matrix $H = \Phi (\Phi^T \Phi)^{-1} \Phi^T$

The dof of regression is: $tr(H)$

For Ridge regression:

$$H(\lambda) = \Phi(\Phi^T\Phi + \lambda I_{PP})^{-1}\Phi^T$$

let $D_\phi := diag(\Phi)$

$$\Rightarrow tr(H(\lambda)) = tr\left[\Phi(\Phi^T\Phi + \lambda I_{PP})^{-1}\Phi^T\right]$$

$$= \sum_{j=1}^{P} (D_\phi^T D_\phi)_{jj}\left[(D_\phi^T D_\phi)_{jj} + \lambda\right]^{-1}$$

1° The range of the value of λ can be determined by the expected dof

To avoid overfitting, one could limit the dof of the model.

If we want the maximum dof of the model to be v:

$$tr(H(\lambda)) \le v$$

$$\Rightarrow \sum_{j=1}^{P} (D_\phi^T D_\phi)_{jj}\left[(D_\phi^T D_\phi)_{jj} + \lambda\right]^{-1} \le v$$

2° Find the optimum λ by Cross Validation.