Least Squares & Friends

Generative Models

- Think about how the model is generated (prior) and how the data are generated (likelihood) from the model.

- Bayes Rule: $\underbrace{P(\text{model} \mid \text{data})}_{\text{posterior}} \propto \underbrace{P(\text{data} \mid \text{model})}_{\text{likelihood}} \underbrace{P(\text{model})}_{\text{prior}}$

$$\log \text{posterior} \propto \log \underbrace{\text{likelihood}}_{\text{comes from data}} + \log \underbrace{\text{prior}}_{\text{comes from prior}}$$

- MAP: $\max_{\text{model}} \log p(\text{model} \mid \text{data}) \rightsquigarrow \min_{\text{model}} \underbrace{-\log \text{likelihood}}_{\text{loss function}} \underbrace{-\log \text{prior}}_{\text{regularization}}$

supervised learning

- Maximum likelihood is a special case (where we don't use a prior, or use a uniform prior so log prior is constant.)

- Belief? Not necessary

Ridge Regression

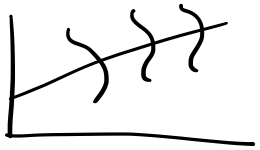
$$\min_{\lambda} \frac{1}{n} \sum_i (y_i - f_\lambda(x_i))^2 + C \|\lambda\|_2^2$$

where $f_\lambda(\bar{x}_i) = \sum_j \lambda_j x_{ij}$

A Bayesian version of it:

$$\vec{\beta} \sim N(\vec{0}, \tau^2 \bar{\bar{I}}) \qquad \tau^2 \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$

$$\vec{Y} \sim N(\bar{\bar{X}}\vec{\beta}, \sigma^2 \bar{\bar{I}})$$



$$\text{posterior} := p(\beta \mid \bar{Y}, \bar{\bar{X}}) = P(\bar{Y} \mid \vec{\beta}, \bar{\bar{X}})\, P(\vec{\beta}) \cdot \frac{1}{Z}$$

"evidence" doesn't depend on β $\rightarrow$ $\int P(\bar{Y} \mid \vec{\beta} \bar{\bar{X}})\, p(\vec{\beta})\, d\vec{\beta}$

$$= \underbrace{\frac{1}{Z} \cdot \frac{1}{(2\pi)^{p/2} \sigma^p}}_{} \exp\left(-\frac{1}{2\sigma^2} \|\bar{Y} - \bar{\bar{X}}\vec{\beta}\|_2^2\right) \cdot \underbrace{\frac{1}{(2\pi)^{p/2} \tau^p}}_{} \exp\left(-\frac{1}{2\tau^2} \|\beta\|^2\right)$$

$$-\log p(\vec{\beta} \mid \bar{Y}, \bar{\bar{X}}) = -\log \text{stuff} + \underbrace{\frac{1}{2\sigma^2} \|Y - \bar{\bar{X}}\vec{\beta}\|_2^2 + \frac{1}{2\tau^2} \|\vec{\beta}\|^2}_{\text{Ridge Regression}}$$

Fact 1: Ridge regression has a generative & frequentist interpretation.

Fact 2 Least squares regression (no regularization) has a closed form solution.

$$\min_{\bar{\lambda}} F(\bar{\lambda}) \quad \text{where} \quad F(\bar{\lambda}) = \|\bar{Y} - \bar{\bar{X}}\bar{\lambda}\|^2 = \sum_i (y_i - \bar{x}_i \bar{\lambda})^2$$

$$\frac{dF(\bar{\lambda})}{d\lambda_j} = -\sum_i 2(y_i - \bar{x}_i \bar{\lambda}) x_{ij} = -2 \bar{X}_{\cdot j}^T (\bar{Y} - \bar{\bar{X}}\bar{\lambda})$$

($\bar{X}_{\cdot j}$: j^{th} feature vector)

in vector notation:

minimize

$$\nabla F(\bar{\lambda}) = -2\bar{\bar{X}}^T(\bar{Y} - \bar{\bar{X}}\bar{\lambda}) = -2\left[\bar{\bar{X}}^T Y - \bar{\bar{X}}^T \bar{\bar{X}} \bar{\lambda}\right] \overset{!}{=} 0$$

$$\bar{\bar{X}}^T Y = \bar{\bar{X}}^T \bar{\bar{X}} \bar{\lambda}^*$$

$$\boxed{(\bar{\bar{X}}^T \bar{\bar{X}})^{-1} \bar{\bar{X}}^T Y = \bar{\lambda}^*}$$

$$\hat{Y} = \bar{\bar{X}}\lambda^* = \underbrace{\bar{\bar{X}} (\bar{\bar{X}}^T \bar{\bar{X}})^{-1} \bar{\bar{X}}^T}_{\text{"hat matrix"}} \bar{Y} = H\bar{Y} = \text{"smoother" of } Y$$

Fact 3 Ridge Regression has a closed form solution.

$$F(\bar{\lambda}) = \|\bar{Y} - \bar{\bar{X}}\lambda\|_2^2 + C\|\bar{\lambda}\|_2^2$$

$$\frac{dF(\bar{\lambda})}{d\lambda_j} = -2\bar{X}_{\cdot j}^T (\bar{Y} - \bar{\bar{X}}\bar{\lambda}) + 2C\lambda_j$$

$$\nabla F(\bar{\lambda}) = -2\bar{\bar{X}}^T(\bar{Y} - \bar{\bar{X}}\bar{\lambda}) + 2C\bar{\lambda}$$

$$\nabla F(\bar{\lambda}) = -2\bar{\bar{X}}^T(\bar{Y} - \bar{\bar{X}}\bar{\lambda}) + 2C\bar{\lambda}$$

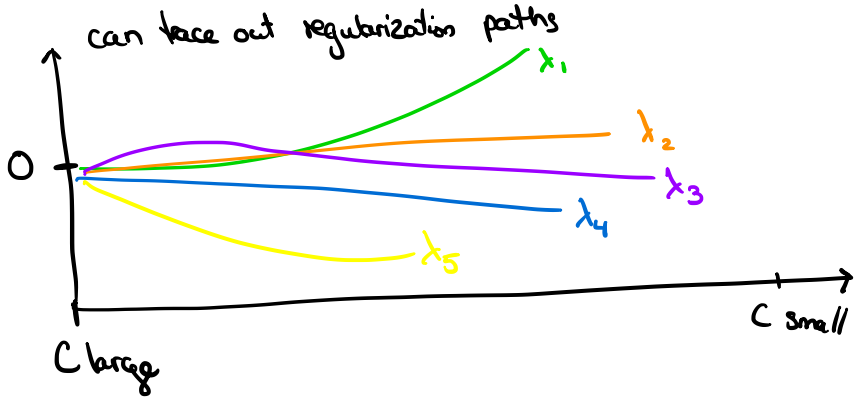
$$= 2(-\bar{\bar{X}}^T\bar{Y} + \bar{\bar{X}}^T\bar{\bar{X}}\bar{\lambda} + C\bar{\lambda}) \overset{\text{minimize}}{=} 0$$

$$\bar{\bar{X}}^T\bar{Y} = (\bar{\bar{X}}^T\bar{\bar{X}} + C\bar{\bar{I}})\bar{\lambda}^*$$

$$\boxed{(\bar{\bar{X}}^T\bar{\bar{X}} + C\bar{\bar{I}})^{-1}\bar{\bar{X}}^T\bar{Y} = \bar{\lambda}^*}$$

Hoerl & Kennard 1970

Moore-Penrose inverse $\rightarrow \bar{\bar{X}}^+ = \lim_{\delta \to 0} (\bar{\bar{X}}^T\bar{\bar{X}} + \delta I)^{-1}\bar{\bar{X}}^T$



Computation issues

$$\lambda^* = (\bar{\bar{X}}^T\bar{\bar{X}} + C\bar{\bar{I}})^{-1}\bar{\bar{X}}^T\bar{Y}$$

↑ expensive $O(p^3)$

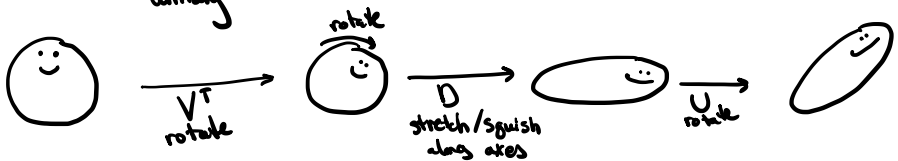
use SVD: $\bar{\bar{X}} = \bar{\bar{U}}\bar{\bar{D}}\bar{\bar{V}}^T$ ← $O(np^2)$

Note:
$U^TU = 1$
$V^TV = 1$

$\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_p \\ | & | & & | \end{bmatrix}$ ↑↑ evectors of XX^T nxp orthonormal unitary

$\begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \end{bmatrix}$

$\begin{bmatrix} — & v_1^T & — \\ — & v_2^T & — \\ & \vdots & \\ — & v_p^T & — \end{bmatrix}$ } evectors of X^TX pxp orthogonal ↑ $A^TA = AA^T = 1$



Simplify:

$$\bar{X}^T\bar{X} = (UDV^T)^T(UDV^T) \overset{\text{property of transpose}}{=} V D U^T U D V^T$$

$U^T U = 1$

$$= VDDV^T = V \begin{pmatrix} d_1^2 & & \\ & \ddots & \\ & & d_p^2 \end{pmatrix} V^T = VD^2V^T$$

Sub into λ^*:

$$\lambda^* = (VD^2V^T + \underbrace{CI})^{-1} X^T Y$$

$$CV\begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}V^T = V\begin{pmatrix} c & & \\ & \ddots & \\ & & c \end{pmatrix}V^T$$

$$= (VD^2V^T + V\begin{pmatrix} c & & \\ & \ddots & \\ & & c \end{pmatrix}V^T)^{-1}\bar{X}^T\bar{Y}$$

$$= \left(V \begin{pmatrix} d_1^2+c & & \\ & d_2^2+c & \\ & & \ddots \end{pmatrix} V^T\right)^{-1} X^T Y$$

properties of inverse

$$= (V^T)^{-1} \begin{pmatrix} d_1^2+c & & \\ & d_2^2+c & \\ & & \ddots \end{pmatrix}^{-1} V^{-1} X^T Y$$

$$= V \begin{pmatrix} 1/(d_1^2+c) & & \\ & 1/(d_2^2+c) & \\ & & \ddots \end{pmatrix} V^{-1} \underbrace{X^T Y}$$

aside: $(UDV^T)^T = V^{TT} D^T U^T = V D U^T$

$$= V \begin{pmatrix} 1/(d_1^2+c) & & \\ & \ddots & \\ & & \ddots \end{pmatrix} \cancel{V^{-1}V} D U^T Y$$

$$= V \begin{pmatrix} 1/(d_1^2+c) & & \\ & \ddots & \\ & & \ddots \end{pmatrix} \begin{pmatrix} d_1 & & \\ & d_2 & \\ & & \ddots \end{pmatrix} U^T Y$$

$$= V \begin{pmatrix} d_1/(d_1^2+c) & & \\ & \ddots & \\ & & \ddots \end{pmatrix} U^T Y = V \operatorname{diag}\left(\frac{d_j}{d_j^2+c}\right) U^T Y$$

Since we have λ^*:

$$\hat{Y}^{ridge} = X\lambda^* = \cancel{UDV^T}\cancel{V}\,\mathrm{diag}\left(\frac{d_j}{d_j^2+c}\right)U^TY$$

$$= U\,\mathrm{diag}\left(\frac{d_j^2}{d_j^2+c}\right)U^TY$$

$$= \sum_j \underbrace{\bar{v}_j}_{n\times 1\ \text{feature}_j}\ \underbrace{\frac{d_j^2}{d_j^2+c}}_{\text{damping effect}}\ \underbrace{\bar{u}_j^T}_{1\times n}\cdot\underbrace{Y}_{n\times 1}$$

Fact 4 This is not too difficult to compute.

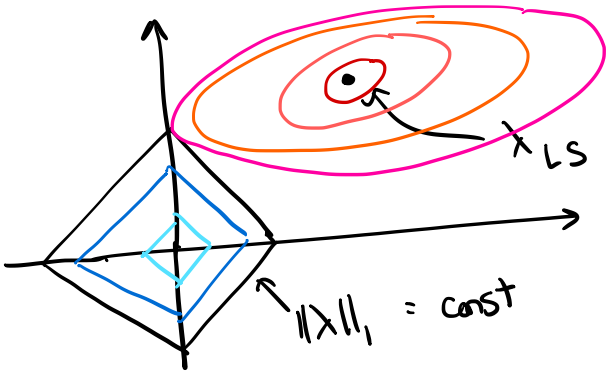
Computational issue averted.

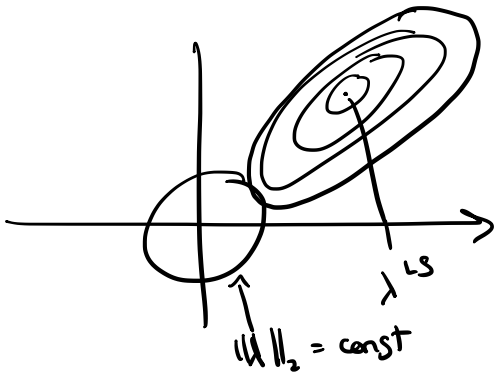
linear in n, linear in p after the svd

Fact 5 None of this works for l_1 penalties (lasso).

$$\sum_i (y_i - f_\lambda(x_i))^2 + c\|\lambda\|_1$$

No closed form solution. Leads to sparser solutions though.





$$\sum_i (y_i - f_\lambda(x_i))^2 \text{ s.t. } \|\lambda\|_1 \le a$$

$$\sum_i (y_i - f(x_i))^2 \text{ s.t. } \|\lambda\|_2^2 \le a$$

Kernel Least Squares

regular ridge

$$F(\bar{\lambda}) = \|\bar{Y} - \bar{\bar{X}}\bar{\lambda}\|_2^2 + C\|\lambda\|_2^2$$

replace $\bar{\lambda} \sim \bar{\bar{X}}^T\bar{r}$ ← If I can get r, then I can get λ

$$F(\bar{r}) = \|Y - \bar{\bar{X}}\bar{\bar{X}}^T\bar{r}\|_2^2 + C\|X^T r\|_2^2$$

$\bar{\bar{X}}$: $n \times p$, $\bar{\bar{X}}^T$: $p \times n$, $\bar{r}$: $n \times 1$, product: $n \times n$

$$F(r) = \|Y - Kr\|_2^2 + C\langle X^T r, X^T r\rangle_2$$

$$\langle X^T r, X^T r\rangle = (X^T r)^T (X^T r)$$

$$r^T X X^T r$$

$$\langle r, Kr\rangle$$

$$F(r) = \|Y - Kr\|_2^2 + C\langle r, Kr\rangle$$

K: $n \times n$, r: $n \times 1$

Shorthand ↓

$$\nabla F(r) = -2K(Y - Kr) + C2Kr = 0 \quad \text{at } r^*$$

$$-Y + Kr^* + Cr^* = 0$$

$$(K + CI)r^* = Y$$

$$\boxed{r^* = (K + cI)^{-1} Y}$$

and remember, $\boxed{\lambda^* = X^T r^*} = \sum_i x_i^T r_i^*$

Prediction at a new test point $\tilde{x}$:

$$f(\tilde{x}) = \tilde{x}\lambda^* = \tilde{x}\sum_i x_i^T r_i^* = \sum_i \tilde{x} x_i^T r_i^*$$

$$= \sum_i \underbrace{k(\tilde{x}, x_i)}_{K_{\tilde{x},i}} r_i^* = K_{\tilde{x}}^T r^* = K_{\tilde{x}}^T (K + cI)^{-1} y$$

representer theorem told us this would happen

Trick: $(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (BPB^T + R)^{-1}$

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} (BPB^T + R) = PB^T$$

$$B^T R^{-1} (BPB^T + R) = (P^{-1} + B^T R^{-1} B) PB^T$$

$$B^T R^{-1} BPB^T + B^T = B^T + B^T R^{-1} BPB^T$$

$X^T r^*$

$X^T(XX^T + cI)^{-1}$ are they equal? $(X^TX + cI)^{-1}X^T$

$B^T(BB^T + R) = (P^{-1} + B^TR^{-1}B)^{-1}X^TR^{-1}$

$P = I$, $R = cI$

$= c(I^{-1} + cX^TX)^{-1}X^TI^{-1}$

$= c\frac{1}{\gamma_c}\left(I + \frac{1}{c}X^TX\right)^{-1}X^TI^{-1}$

$= (cI + X^TX)^{-1}X^T$

$\therefore$ equal

Kernel ridge = ridge in linear case