BASIC RIDGE

$$\min_{w} \|Xw^T - y\|^2 + \lambda \|w\|^2$$

↳ 1. expand

2. take derivative w.r.t to w
3. set it to 0

⇓

$$w = (X^T X - \lambda I)^{-1} X^T y$$

BASIC RIDGE	DUAL RIDGE
$$\min_{w} \|Xw - y\|^2 + \lambda \|w\|^2$$	- approach from mathematical programming (field of OR/LP)
↳ 1. expand 2. take derivative w.r.t to w 3. set it to 0 ⇓ $w = (X^T X - \lambda I)^{-1} X^T y$	

BASIC RIDGE

$$\min_w \|Xw^T - y\|^2 + \lambda\|w\|^2$$

↳ 1. expand

2. take derivative w.r.t to w

3. set it to 0

⇓

$$w = (X^T X - \lambda I)^{-1} X^T y$$

DUAL RIDGE

- approach from mathematical programming (field of OR/LP)

objective: $\arg\min_{w,e} \|e\|^2 + \lambda\|w\|^2$

s.t. $e = X^T w - y$

↳ - variables e, w

- coefficients X, y

BASIC RIDGE

$$\min_{w} \|Xw^T - y\|^2 + \lambda\|w\|^2$$

↳ 1. expand

2. take derivative w.r.t to w

3. set it to 0

⇓

$$w = (X^TX - \lambda I)^{-1} X^T y$$

DUAL RIDGE

- approach from mathematical programming (field of OR/LP)

objective: $\arg\min_{w,e} \|e\|^2 + \lambda\|w\|^2$

s.t. $e = X^T w - y$

key assumption: X is centred at mean 0, thus **NO** intercept

↳ variables e, w

- coefficients X, y

⇒ reformulate into Lagrangian

BASIC RIDGE	DUAL RIDGE

BASIC RIDGE

$$\min_w \|Xw^T - y\|^2 + \lambda\|w\|^2$$

↳ 1. expand
2. take derivative w.r.t to w
3. set it to 0

$$\Downarrow$$

$$w = (X^TX - \lambda I)^{-1} X^T y$$

DUAL RIDGE

- approach from mathematical programming (field of OR/LP)

objective: $\arg\min_{w,e} \|e\|^2 + \lambda\|w\|^2$

s.t. $e = X^T w - y$ ← key assumption: X is centred at mean 0, thus NO intercept

↳ variables e, w
- coefficients X, y

⇒ reformulate into Lagrangian

lagrange multipliers

$$L(e, w, \alpha) = \|e\|^2 + \lambda\|w\|^2 + \alpha^T(e - X^T w + y)$$

BASIC RIDGE

$$\min_{w} \|Xw^{T} - y\|^2 + \lambda\|w\|^2$$

↳ 1. expand
2. take derivative w.r.t to w
3. set it to 0

⇓

$$w = (X^TX - \lambda I)^{-1}X^Ty$$

DUAL RIDGE

- approach from mathematical programming (field of OR/LP)

objective: $\arg\min_{w,e} \|e\|^2 + \lambda\|w\|^2$

s.t. $e = X^Tw - y$

key assumption: X is centred at mean 0, thus **NO** intercept

↳ - variables e, w
- coefficients X, y

⇒ reformulate into Lagrangian

lagrange multipliers

$$L(e,w,\alpha) = \|e\|^2 + \lambda\|w\|^2 + \alpha^T(e - X^Tw + y)$$

→ derivate!

$$\left.\begin{aligned}\frac{\partial L}{\partial e} &= 2e + \alpha \\ \frac{\partial L}{\partial w} &= 2\lambda w - X\alpha\end{aligned}\right\} \text{set to } \emptyset$$

BASIC RIDGE

$$\min_{w} \|Xw^T - y\|^2 + \lambda\|w\|^2$$

↳ 1. expand
2. take derivative w.r.t to w
3. set it to 0

⇓

$$w = (X^TX - \lambda I)^{-1} X^T y$$

DUAL RIDGE

- approach from mathematical programming (field of OR/LP)

objective: $\arg\min_{w,e} \|e\|^2 + \lambda\|w\|^2$

s.t. $e = X^T w - y$

key assumption: X is centered at mean 0, thus **NO** intercept

↳ variables e, w
- coefficients X, y

⇒ reformulate into Lagrangian

lagrange multipliers

$$L(e, w, \alpha) = \|e\|^2 + \lambda\|w\|^2 + \alpha^T(e - X^T w + y)$$

→ derivate!

$$\left.\begin{aligned} \frac{\partial L}{\partial e} &= 2e + \alpha \\ \frac{\partial L}{\partial w} &= 2\lambda w - X\alpha \end{aligned}\right\} \text{set to } 0$$

$$\boxed{e = -\frac{1}{2}\alpha^T}$$

$$\boxed{w = \frac{1}{2\lambda} X\alpha}$$

⇓

optimal values of w, e

↳ let's substitute them in!

Dual Ridge (continued)

$$L(e(\alpha), w(\alpha), \alpha) =$$

$$= \frac{1}{4}\|\alpha\|^2 + \frac{\lambda}{4\lambda^2}\|X\alpha\|^2 + \alpha^T\left(-\frac{1}{2}\alpha - \frac{1}{2\lambda}X^TX\alpha + y\right)$$

$$= -\frac{1}{4}\|\alpha\|^2 - \frac{1}{4\lambda}\|X\alpha\|^2 + \alpha^T y$$

Dual Ridge (continued)

$L(e(\alpha), w(\alpha), \alpha) =$

$$= \frac{1}{4}\|\alpha\|^2 + \frac{\lambda}{4\lambda^2}\|X\alpha\|^2 + \alpha^T\left(-\frac{1}{2}\alpha - \frac{1}{2\lambda}X^TX\alpha + y\right)$$

$$= -\frac{1}{4}\|\alpha\|^2 - \frac{1}{4\lambda}\|X\alpha\|^2 + \alpha^T y$$

↳ X: known
y: known
λ: we choose
α: <u>unknown</u>!

↳ <u>DUAL</u>: $\arg\min_{\alpha}\left[-\frac{1}{4}\alpha^T\left(I + \frac{1}{\lambda}X^TX\right)\alpha + \alpha^T y\right]$ $/\cdot(-\lambda)$

$$g(\alpha) = \arg\min_{\alpha}\left[\frac{1}{4}\lambda\,\alpha^T(X^TX - \lambda I)\alpha - \lambda\alpha^T y\right]$$

Dual Ridge (continued)

$$L(e(\alpha), w(\alpha), \alpha) =$$

$$= \frac{1}{4}\|\alpha\|^2 + \frac{\lambda}{4\lambda^2}\|X\alpha\|^2 + \alpha^T\left(-\frac{1}{2}\alpha - \frac{1}{2\lambda}X^TX\alpha + y\right)$$

$$= -\frac{1}{4}\|\alpha\|^2 - \frac{1}{4\lambda}\|X\alpha\|^2 + \alpha^T y$$

$\hookrightarrow$ X: known

y: known

λ: we choose

α: unknown!

$\Rightarrow$ DUAL: $\arg\min_{\alpha}\left[-\frac{1}{4}\alpha^T\left(I + \frac{1}{\lambda}X^TX\right)\alpha + \alpha^T y\right]$ $\quad / \times(-\lambda)$

$$g(\alpha) = \arg\min_{\alpha}\left[\frac{1}{4}\lambda\,\alpha^T(X^TX - \lambda I)\alpha - \lambda\alpha^T y\right]$$

$$\boxed{\begin{aligned} &\frac{\partial g}{\partial \alpha} = \frac{1}{2}\alpha^T(X^TX - I\lambda) - \lambda y = 0 \\ &\alpha = 2\lambda(X^TX - I\lambda)^{-1}y \\ &\Rightarrow w = X(X^TX - I\lambda)^{-1}y \end{aligned}}$$

! EQUIVALENT TO PRIMAL !

Dual Ridge (continued)

Key point of comparison:
- requires only X^TX !
- we don't need X by itself

$$L(e(\alpha), w(\alpha), \alpha) =$$

$$= \frac{1}{4}\|\alpha\|^2 + \frac{\lambda}{4\lambda^2}\|X\alpha\|^2 + \alpha^T\left(-\frac{1}{2}\alpha - \frac{1}{2\lambda}X^TX\alpha + y\right)$$

$$= -\frac{1}{4}\|\alpha\|^2 - \frac{1}{4\lambda}\|X\alpha\|^2 + \alpha^T y$$

↳ X: known
y: known
λ: we choose
α: unknown!

$$\frac{\partial g}{\partial \alpha} = \frac{1}{2}\alpha^T(X^TX - I\lambda) - \lambda y = 0$$

$$\alpha = 2\lambda(X^TX - I\lambda)^{-1}y$$

↳ $w = X(X^TX - I\lambda)^{-1}y$

! EQUIVALENT TO PRIMAL !

↳ DUAL: $\arg\min_\alpha \left[-\frac{1}{4}\alpha^T\left(I + \frac{1}{\lambda}X^TX\right)\alpha + \alpha^Ty\right]$ $/ \times (-\lambda)$

$$g(\alpha) = \arg\min_\alpha \left[\frac{1}{4}\lambda\,\alpha^T(X^TX - \lambda I)\alpha - \lambda\alpha^Ty\right]$$

Kernel Regression

↳ what is a kernel?

Kernel Regression

↳ what is a kernel?

↳ implicit mapping into a higher dimensional space that gives us pairwise distances of the mapped data points but NOT their actual coordinates.

Kernel Regression

↳ what is a kernel?

↳ implicit mapping into a higher dimensional space that gives us pairwise distances of the mapped data points but NOT their actual coordinates.

$X^T X$ -> cov. matrix
aka
dot product matrix
aka
matrix of pair-wise distances

Kernel Regression

↳ what is a kernel?

↳ implicit mapping into a higher dimensional space that gives us pairwise distances of the mapped data points but NOT their actual coordinates.

X^TX → cov. matrix
aka
dot product matrix
aka
matrix of pair-wise distances

↳ since dual ridge only needs X^TX it provides a perfect opportunity to replace $X^TX = K = \Phi(x) \cdot \phi(x)$

Understanding why we need $\bar{X} = 0$

↳ usual ridge: $f(w_0, w) = \sum_{i=1}^{n} (x_i^T w + w_0 - y_i)^2 + \lambda \|w\|^2$

↳ $\frac{\partial f}{\partial w_0} = \cancel{2} \sum_{i=1}^{n} x_i^T w + w_0 - y_i = 0$

Understanding why we need $\bar{X} = 0$

↳ usual ridge: $f(w_0, w) = \sum_{i=1}^{n} (x_i^T w + w_0 - y_i)^2 + \lambda \|w\|^2$

↳ $\frac{\partial f}{\partial w_0} = \cancel{2} \sum_{i=1}^{n} x_i^T w + w_0 - y_i = 0$

$$n w_0 = \sum_i y_i - \sum_i x_i^T w$$

$$w_0 = \bar{y} - \bar{x}^T w$$

Understanding why we need $\bar{X} = 0$

↳ usual ridge: $f(w_0, w) = \sum_{i=1}^{n} (x_i^T w + w_0 - y_i)^2 + \lambda \|w\|^2$

↳ $\frac{\partial f}{\partial w_0} = \cancel{2} \sum_{i=1}^{n} x_i^T w + w_0 - y_i = 0$

$n w_0 = \sum_i y_i - \sum_i x_i^T w$

$f(w, w_0(w)) = \sum_{i=1}^{n} \left[(x_i - \bar{x})^T w - (y_i - \bar{y}) \right]^2 + \lambda \|w\|^2 \leftarrow w_0 = \bar{y} - \bar{x}^T w$

Understanding why we need $\bar{X} = 0$

↳ usual ridge: $f(w_0, w) = \sum_{i=1}^{n} (x_i^T w + w_0 - y_i)^2 + \lambda \|w\|^2$

↳ $\frac{\partial f}{\partial w_0} = \cancel{2} \sum_{i=1}^{n} x_i^T w + w_0 - y_i = 0$

$n w_0 = \sum_i y_i - \sum_i x_i^T w$

$w_0 = \bar{y} - \bar{x}^T w$

$$\underbrace{f(w, w_0(w)) = \sum_{i=1}^{n} [(x_i - \bar{x})^T w - (y_i - \bar{y})]^2 + \lambda \|w\|^2}_{\text{if } \bar{x} \text{ \& } \bar{y} \text{ had 0 mean life would be simpler}}$$

Understanding Why we need $\bar{X} = 0$

↳ usual ridge: $f(w_0, w) = \sum_{i=1}^{n} (x_i^T w + w_0 - y_i)^2 + \lambda \|w\|^2$

↳ $\frac{\partial f}{\partial w_0} = \cancel{2} \sum_{i=1}^{n} x_i^T w + w_0 - y_i = 0$

$n w_0 = \sum_i y_i - \sum_i x_i^T w$

$f(w, w_0(w)) = \sum_{i=1}^{n} \left[(x_i - \bar{x})^T w - (y_i - \bar{y}) \right]^2 + \lambda \|w\|^2$ <— ↳ $w_0 = \bar{y} - \bar{x}^T w$

if $\bar{x}$ & $\bar{y}$ had 0 mean life would be simpler $\Rightarrow$

thus we center using

$JX = (I - \frac{1}{n} \mathbb{1}\mathbb{1}^T) X$

↳ $x_i - \frac{1}{n} \bar{x} \quad \forall i$