Ridge regression

Let $\underline{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}$, $x_i \in \mathbb{R}^n$, m = # samples

Let $y \in \mathbb{R}^m$ represent a set of targets. Want to solve

$$(*) \qquad \underline{X} w \approx y$$

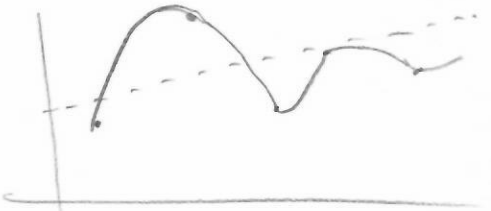
Ex:

Consider polynomial regression on $\mathbb{R}$ w/ a degree $n-1$ polynomial. ~~Have $\underline{X} \in \mathbb{R}$~~

Let $m = n$.

Have $\underline{X} \in \mathbb{R}^{n \times n}$ is square

$w^* := \underline{X}^{-1} y$ solves $(*)$. Unless (x_i, y_i) lie on a line, the resulting polynomial is very wiggly. Suppose also, that $y_i = x_i + \eta$, $\eta \sim N(0, \varepsilon)$



Then our polynomial will be way too wiggly & will generalize poorly.

Thought:
A simpler hypothesis will place less energy on coeffs of high degree monomials. In practice, we see that often a lot of energy gets placed in these coeffs (they "cancel each other out"?)

Maybe if we try to find something that balances between solving $(*)$ & keeping $\|w\|^2$ small, we can ~~obtain a~~ focus on a simpler hypothesis class.

Ridge regression prob.:

- Fix $\lambda \geq 0$

$$\min_w \quad \|Xw - y\|^2 + \lambda \|w\|^2 \qquad (P)$$

Note: (P) is convex, can be efficiently solved by an optimizer. But can also solve directly w/ linear algebra.

Let

$$f_\lambda(w) = \|Xw - y\|^2 + \tfrac{1}{2}\lambda \|w\|^2$$

$w^* \in \operatorname{argmin}_w f_\lambda(w) \iff \nabla f_\lambda(w) = 0$ (convex)

Let's compute $\nabla f_\lambda(w)$.

Prelim 1: Suppose $g(x) = Ax$, $g: \mathbb{R}^n \to \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$

$$Dg(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial x_n} & & \frac{\partial g_m}{\partial x_n} \end{pmatrix}$$

Note $\frac{\partial g_i}{\partial x_j} = a_{ij} \leftarrow g_i$ focuses on ith row, x_j on jth col.

$\Rightarrow Dg(x) = A$.

Recall: chain rule: $D_w f(g(w)) = Df \circ Dg$

Prelim 2: $D_y \|y\|^2 = 2y$

pf 1. $\|y\|^2 = y^T y$. Apply product rule. Note $\frac{\partial}{\partial y_j} y^T z = z = \frac{\partial}{\partial y_j} z^T y$.

Thus $D_y \|y\|^2 = y + y$

pf 2 $f(y) = \|y\|^2 = \sum_{i=1}^n |y_i|^2$. $\frac{\partial f}{\partial y_i} = 2y_i \Rightarrow Df = 2\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = 2y$ ∎

Now, compute $\nabla f_\lambda(w)$:

$$\nabla_w \|Xw - y\|^2 = (Xw - y) \cdot X = (Xw - y)^T X$$

$$0 = \nabla f_\lambda(w) \iff (X^T w - y)^T X + \lambda w = 0$$

Nota: $X : m \times n$

$$\Rightarrow w^T X X^T - y^T X + \lambda w = 0$$

$$\Rightarrow X X^T w - y^T X + \lambda w = 0$$

$$\Rightarrow (X X^T + \lambda I_n) w = y^T X$$

$$\Rightarrow w = (X X^T + \lambda I_m)^{-1} y^T X$$

$m \times m$ $\quad \mathbb{R}^m$ ✓

Note: How do we know $X X^T + \lambda I_m$ is invertible?

$X X^T$ is positive semi-definite by construction ($\forall z \in \mathbb{R}^m$ $z^T X X^T z = \|z^T X\|^2$)

Moreover,

$z^T (X X^T + \lambda I_m) z = \|z^T X\|^2 + \lambda \|z\|^2 \Rightarrow$ it's pos. def.

Question: What happens numerically when m or n is large. When is it better to solve via optimization vs normal eqn?

m	dim
4	2
	4
1	20
10	4
	20
	4
400	4

← when $d < m$, problem what if I do wrong?