Ridge Regression
(a.k.a. Bayesian Linear Regression)

We have already seen:

$$Y \mid X = x \sim \mathcal{N}(w^T x, \sigma^2) \quad \text{] Likelihood.}$$

Now let's put a prior on $\vec{w}$:

$$\vec{w} \sim \mathcal{N}(0, t^2 I) \quad \text{] prior}$$

$$\Big\downarrow \text{ univariate form}$$

$$w_j \sim \mathcal{N}(0, t^2) \quad \text{(iid samples)}$$

$$P(\vec{w}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi t^2}} \exp\left[-\frac{(w_i - 0)^2}{2t^2}\right]$$

$$\hat{w}_{MAP} = \arg\max_{\vec{w}} \log P(\vec{w} \mid D)$$

$$= \arg\max_{\vec{w}} \log \left[\frac{P(D \mid \vec{w})\, P(\vec{w})}{P(D)}\right]$$

$$= \arg\max_{\vec{w}} \left[\log P(D \mid \vec{w}) + \log P(\vec{w})\right]$$

$$= \arg\max_{\vec{w}} \left[-\sum_{i=1}^{n} \frac{(y_i - \vec{w}^T x_i)^2}{2\sigma^2} - \underbrace{\frac{n}{2} \log(2\pi\sigma^2)}_{\text{const. wrt } w} - \sum_{j=0}^{d} \frac{w_j^2}{2t^2} - \underbrace{\frac{d}{2} \log(2\pi t^2)}_{\text{const. wrt } w}\right]$$

$$= \arg\min_{\vec{w}} \left[\sum_{i=1}^{n} (y_i - w^T x_i)^2 + \frac{2\sigma^2}{2t^2} \sum_{j=0}^{d} w_j^2\right] \quad \text{(why?)}$$

$$= \arg\min_{\vec{w}} \left[\underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2}_{\substack{\text{Average error} \\ \text{(least sq. fitting error)}}} + \frac{1}{n} \underbrace{\frac{2\sigma^2}{2t^2}}_{\lambda} \underbrace{\sum_{j=0}^{d} w_j^2}_{\|\vec{w}\|_2^2}\right]$$

$$= \arg\min_{\vec{w}} \left[\underbrace{\frac{1}{n} \sum_{i} (y_i - w^T x_i)^2}_{\text{Avg. loss}} + \underbrace{\frac{\lambda}{n}}_{\text{Trade-off}} \underbrace{\|w\|_2^2}_{\text{Regularization!!}}\right]$$

$$\boxed{\text{loss} + \lambda \cdot \text{prior}}$$

So, Bayesian loss = Avg. fitting error + λ . Norm square of $\vec{w}$

The norm of $\vec{w}$ should be small (why?)

(12)

As $\lambda \to 0$, $t^2 \to \infty \Rightarrow$ Prior with a Gaussian having very large variance $\to$ nearly uniform

$$\Rightarrow \hat{W}_{MAP} = \hat{W}_{MLE}$$

As $\lambda \to \infty$, $\vec{w} \to 0$

$$\vec{w}^T\vec{x} = w_0 + \underbrace{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}_{\to 0}$$

With w_0, we don't have any information about the data. We have only the y's. The best that can be done is to take an average. So w_0 becomes the mean of y. So, $\hat{y}$ becomes y_{mean}

In some sense, the regularizer encourages the optimization to pick a simple model - here pushes to mean.

Now let's get back to solving the minimization problem:

$$\hat{w}_{MAP} = \arg\min_w \frac{1}{n}\left[(y - Xw)(y - Xw) + \lambda w^T w\right]$$

$$\frac{\partial}{\partial w}\left[\frac{1}{n}\left((y-Xw)(y-Xw) + \lambda w^T w\right)\right] = 0$$

$$\Rightarrow \frac{\partial}{\partial w}\left[\frac{1}{n}\left((y^T y + w^T X^T X w - 2y^T X w) + \lambda w^T w\right)\right] = 0$$

$$\Rightarrow 2w^T X^T X - 2y^T X + 2\lambda w^T = 0$$

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

$$\Rightarrow \boxed{\hat{w}_{MAP} = (X^T X + \lambda I)^{-1} X^T y}$$

Compare with $\hat{w}_{MLE} = (X^T X)^{-1} X^T y$

Modification of pseudo-inverse: $(X^T X)^{-1} \to (X^T X + \lambda I)^{-1}$

What adding an identity matrix does to matrix eigenvalues??

$$Ax = \lambda x$$

$$(A + I)x = Ax + Ix = (\lambda + 1)x$$

↳ Shifts eigenvalues up!!

So now the noninvertible problem becomes solvable.

Interpret the effect of regularizer.

Likelihood	Prior	Name
Gaussian	Uniform	Least square
"	Gaussian	Ridge Regression
"	Laplace	Lasso
Laplace	Uniform	Robust Regression
Student	Uniform	Robust Regression.