Parameter estimation for linear regression

$$y_i = \vec{w}^T \vec{x}_i + \epsilon \quad \text{where } \epsilon \sim N(0, \sigma^2)$$

Here, y_i is a real scalar and $\vec{w}$ and $\vec{x}_i$ are vectors of dimensions $d \times 1$

Given n pairs of (x_i, y_i), we can estimate $\vec{w}$ that best fits the model.

These notes cover the following methods of parameter estimation:

(1) least squares estimation } Similar or equivalent

(2) MCLE

(3) Regularized least squares estimation } Similar or equivalent

(4) MAP estimate

(1) Least Squares Estimation

We consider the sum of squared errors as our loss function:

$$f(\vec{w}) = (\vec{y} - X\vec{w})^T (\vec{y} - X\vec{w})$$

$\vec{y}$: $n \times 1$, X: $n \times d$, $\vec{w}$: $d \times 1$

$$= (\vec{y}^T - \vec{w}^T X^T)(\vec{y} - X\vec{w})$$

$$= (\vec{y}^T\vec{y} - \vec{w}^T X^T \vec{y} - \vec{y}^T X\vec{w} + \vec{w}^T X^T X\vec{w})$$

$\vec{w}^T X^T \vec{y}$ and $\vec{y}^T X\vec{w}$: same scalar value

$$= (\vec{y}^T\vec{y} - 2\vec{y}^T X\vec{w} + \vec{w}^T X^T X\vec{w})$$

$$\nabla_{\vec{w}} f = 0 - 2X^T\vec{y} + 2X^T X\vec{w}$$

Put to zero:

$$2X^T X\vec{w} = 2X^T\vec{y}$$

$$\therefore \vec{w} = (X^T X)^{-1} X^T \vec{y}$$

(2) Maximum Conditional likelihood Estimation

$$y_i = \vec{w}^T \vec{x}_i + C, \quad \text{where } \epsilon \sim N(0, \sigma^2)$$

$$\therefore y \sim N(w^T x, \sigma^2)$$

$$MCLE \equiv \arg\max_{\vec{w}} \prod_{i=1}^{n} P(y_i \mid x_i, \vec{w})$$

$$\log l(\vec{w}) = \log \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{1}{2}\left(\frac{y_i - \vec{w}^T \vec{x}_i}{\sigma}\right)^2} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{-1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{(y_i - \vec{w}^T \vec{x}_i)^2}{\sigma^2} \right]$$

Ignoring the 1st term because it does not involve $\vec{w}$

$$\log l(\vec{w}) = \sum_{i=1}^{n} \left[\frac{-1}{2\sigma^2} \times (y_i - \vec{w}^T \vec{x}_i)^2 \right]$$

$$= \frac{-1}{2\sigma^2} \times (\vec{y} - X\vec{w})^T (\vec{y} - X\vec{w})$$

$\therefore$ We are maximizing a function proportional to the negative of sum of squared errors

$\therefore$ MCLE view is equivalent to sum of squared errors estimation.

(3) Regularized least squares estimation

We now consider our loss function to be sum of squared errors PLUS square of l_2 norm of the parameter vector $\vec{w}$:

$$f(\vec{w}) = (\vec{y} - X\vec{w})^T (\vec{y} - X\vec{w}) + \frac{\lambda}{2} \|w\|_2^2$$

Here, λ is a parameter that helps decide the importance attached to minimized the squares of the weight vector $\vec{w}$

$$\nabla_{\vec{w}} f = 0 - X^T\vec{y} - X^T\vec{y} + 2X^TX\vec{w} + \lambda\vec{w}$$

Put to zero:

$$2X^TX\vec{w} + \lambda\vec{w} = 2X^T\vec{y}$$

$$\therefore \vec{w} = \left(X^TX + \frac{\lambda}{2}I\right)^{-1} X^Ty$$

(4) MAP estimate for linear regression

$$P(\vec{w}|D) = \frac{P(D|\vec{w}) \times P(\vec{w})}{P(D)}$$

$P(D|\vec{w})$ → likelihood

Here, D is data

$$\vec{w}_{MAP} = \arg\max_{\vec{w}} \left[\ln\left[P(D|\vec{w})\right] + \ln\left[P(\vec{w})\right] - \ln\left[P(D)\right] \right]$$

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

Let us assume that $\vec{w}$ is drawn from a

Gaussian i.e.

$$\vec{w} \sim \mathcal{N}\left(\vec{0}, \frac{1}{\lambda} \times I\right)$$ where I is identity matrix

$$\therefore \vec{w}_{MAP} = \arg\max_{\vec{w}} \left[\ln\left[\frac{1}{(2\pi)^{d/2} \left|I \times \frac{1}{\lambda}\right|^{1/2}} \times e^{-\frac{1}{2}\left(\vec{w}^T(\lambda I)w\right)} \right] + \ln\left[P(D|\vec{w})\right] \right]$$

$$= \arg\max_{\vec{w}} \left[\ln\left[\frac{1}{(2\pi)^{d/2} \left|I \times \frac{1}{\lambda}\right|^{1/2}} \right] - \frac{\lambda}{2} \vec{w}^T w + \ln\left[P(D|\vec{w})\right] \right]$$

↓ Ignore

$$= \arg\max_{\vec{w}} \left[-\frac{\lambda}{2} \|w\|^2 + \text{log-likelihood} \right]$$

$\equiv L_2$ regularization for linear regression

For $\sigma =$ some constant, this is equivalent to L_2 regularizatio