Probabilistic Interpretation of linear models.

$$y = w^T x$$

w → weight vector

Probabilistic linear regression

y is a random variable

x — is not a random variable.

$$y \sim \mathcal{N}(w^T x, \sigma^2)$$

w, σ^2

$$y = w^T x + \epsilon$$

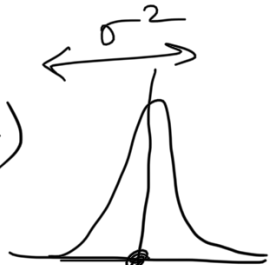
ϵ ← random variable

$\epsilon \sim \mathcal{N}(0, \sigma^2)$

If we have w and σ^2

Given a new x^*,

$$y^* \sim \mathcal{N}(w^T x^*, \sigma^2)$$



$w^T x^*$

Training data

$$D = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \\ x_N & y_N \end{bmatrix}$$

$y_1 \sim N(w^T x_1, \sigma^2)$

$y_2 \sim N(w^T x_2, \sigma^2)$

$\vdots$

$y_N \sim N(w^T x_N, \sigma^2)$

Likelihood of the dataset:

$$L(D) = \prod_{i=1}^{N} p(y_i)$$

$$ll(D) = \sum_{i=1}^{N} \log p(y_i)$$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right]\right)$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - w^T x_i)^2$$

MLE estimate $\equiv$ the values of w & σ^2 at which $ll(D)$ is max.

$$\arg\max_{w,\sigma^2} ll(D)$$

$$ll(D) = \text{const} - N\log\sigma - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{N}(y_i - w^T x_i)^2$$

$\sum_{i=1}^{N}$

const /

w

Equivalent to maximizing: $-\frac{1}{2}\sum (y_i - w^T x_i)^2$

which is equiv. to minimizing $\frac{1}{2}\sum (y_i - w^T x_i)^2$

Squared loss for geometric linear regression.

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

where $X \rightarrow$ data matrix $N \times d$

$y \rightarrow$ vector of target values $N \times 1$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum (y_j - w^T x_j)^2$$

$$= \frac{1}{N} (y - Xw)^T (y - Xw)$$

Imposing a prior on w

$w \rightarrow$ a d-dimensional vector

MVN or a multivariate Gaussian

$$p(w) \sim \mathcal{N}(w \mid \mu_0, \Sigma_0)$$

$d \times 1$ $d \times d$

Prior

Simple case: $\mu_0 = \underset{d\times 1}{0}$

$$\Sigma_0 = \tau^2 \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

$$= \underset{(d\times d)}{\tau^2 I}$$

$X, Y \rightarrow$ Training

$$p(w|X,Y) = \frac{p(X,Y|w)\,p(w)}{\int_{w'} p(X,Y|w')\,p(w')\,dw'}$$

$$p(X,Y|w) = \prod_{i=1}^{N} \mathcal{N}(y_i | w^T x_i, \sigma^2)$$

Posterior is also Gaussian $p(w|X,Y) \sim \mathcal{N}(w|\bar{w}, \bar{\Sigma})$

$$\bar{w} = \underset{d\times 1}{\left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T y}$$

$$\bar{\Sigma} = \underset{d\times d}{\sigma^2 \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1}}$$

$$y \sim \mathcal{N}(y | w^T x, \sigma^2)$$

σ^2 - fixed and known

Relationship with ridge regression

$$W_{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$

λ: regularization parameter

$$\bar{W} = (X^T X + \frac{\sigma^2}{\tau^2} I)^{-1} X^T y$$

What is the MAP estimate for w

$$\hat{W}_{MAP} = \bar{W}$$

Inference or prediction task:

$$x^* \longrightarrow y^*$$

$$y^* = \hat{W}_{MLE}^T x^*$$

$$y^* = \hat{W}_{MAP}^T x^*$$

Full Bayesian Treatment:

CSE610 - Fall 2020

Non-parametric Bayesian Methods

$$y \sim \mathcal{N}(y \mid w^T x, \sigma^2)$$

$$y \sim \mathcal{F}(y \mid \theta = f(w^T x))$$

Generalized linear Models. (GLM)

Replace $\mathcal{N}(\,) \rightarrow$ Laplace( )

Robust Regression

$$p(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}(y_i - w^T x_i)^2\right]$$

Impacted by outliers

$$p(y) = \text{Const} \exp\left[-\text{Cst}\,|y_i - w^T x|\right]$$

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - w^T x_i)^2$$ least squares

least absolute MAD Deviation

$\frac{1}{2} \sum_{i=1} (y_i - w \, x_i)$

Logistic Regression is a member of GLM family.