Probabilistic Interpretation of linear models.

$$\boxed{y = w^T x}$$

w → weight vector

Probabilistic linear regression

y is a random variable

x — is not a random variable,

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

$$\boxed{w, \sigma^2}$$

$$\underline{y = w^T x + \epsilon}$$

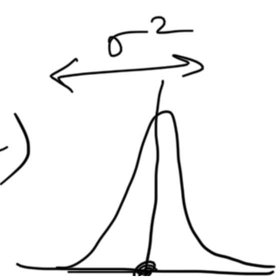
ϵ ← random variable

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

If we have w and σ^2

Given a new x^*,

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



$w^T x^*$

Training data

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

$$y_1 \sim \mathcal{N}(w^T x_1, \sigma^2)$$
$$y_2 \sim \mathcal{N}(w^T x_2, \sigma^2)$$
$$\vdots$$
$$y_N \sim \mathcal{N}(w^T x_N, \sigma^2)$$

Likelihood of the dataset:

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

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

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

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

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

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

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

$i=1$

const

$\underline{w}$

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

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

Squared loss for geometric linear regression.

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

where $X \rightarrow$ data matrix $\underline{N \times d}$

$y \rightarrow$ vector of target values $\underline{N \times 1}$

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

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

<u>Imposing a prior on w</u>

$w \rightarrow$ a d-dimensional vector

MVN or a multi-variate Gaussian

$$\underline{p(w)} \sim \mathcal{N}(w \mid \underbrace{\mu_0}_{d \times 1}, \underbrace{\Sigma_0}_{d \times d})$$

Prior

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

$$\Sigma_0 = \tau^2 \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} = \tau^2 I_{(d\times d)}$$

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

X, Y → Training

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

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

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

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

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

σ^2 - fixed and known

Relationship with ridge regression

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

regularization parameter

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

what is the MAP estimate for w

$$\boxed{\hat{w}_{MAP} = \bar{w}}$$

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

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

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

Generalized linear Models. (GLM)

Replace $\mathcal{N}(\;) \rightarrow \text{Laplace}(\;)$

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

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

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

Least absolute

~~MAD~~ Deviation

$$\sum^{N} \ldots w^T x_i| \quad -$$

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

Logistic Regression is a member of GLM family.

Probabilistic Logistic Regression

1. Generalized linear Models (GLM)

2. Bayesian logistic Regression

3. Handling multiple classes

GLM

$w^T x$

$$\boxed{y \sim F(f(w^T x), --)}$$

Classification $\quad y \in \{0, 1\}$

Bernoulli distribution

$y \sim Ber(y \mid \theta)$

$$\boxed{\theta = \text{sigmoid}(w^T x) = \frac{1}{1 + \exp(-w^T x)}}$$

$0 \le \theta \le 1$

$x^* \rightarrow w^T x^*$

$\text{sigmoid}(w^T x^*)$

If sigmoid $(w^T x^*) \geq 0.5$ $y^* = 1$

else $y^* = 0$

Given training data, D $\langle x_i, y_i \rangle_{i=1}^{N}$

learn w

MLE: $x_i, y_i \rightarrow$ random variable

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

$P(y = y_i)$?

If $y_i = 1$, then $P(Y = y_i) = \theta_i = \dfrac{1}{1 + \exp(-w^T x_i)}$

If $y_i = 0$, then $P(Y = y_i) = 1 - \theta_i$

$$= 1 - \left(\frac{1}{1 + \exp(-w^T x_i)} \right)$$

$$= \frac{1}{1 + \exp(w^T x_i)}$$

In general

$$P(Y = y_i) = \theta_i^{y_i} (1 - \theta_i)^{1 - y_i},$$

$$LL(D) = \prod_{i=1}^{N} \theta_i^{y_i} (1-\theta_i)^{1-y_i}$$

$$LL(D) = \sum_{i=1}^{N} y_i \log \theta_i + (1-y_i) \log (1-\theta_i)$$

where $\theta_i = \frac{1}{1+\exp(-w^T x_i)}$

We can maximize this LL(D) w.r.t w

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

No closed-form expression.

Have to use a gradient based method.

→ Gradient Descent

→ Newton's method

Regularization and prior are some what equivalent.

$p(w)$ $\quad w \sim \mathcal{N}(w \mid 0, \tau^2 I)$

$p(D)$

$LL(D)$

$\hookrightarrow p(w|D)$

I could get the MAP estimate for w by adding a L2-penalty to the $LL(w)$

But it is not easy to get $p(w|D)$

posterior.

Generalize to multi-class classification.

$y \in \{1, 2, 3, \ldots, 10\}$

Bernoulli X

<u>Multinoulli</u> $\rightarrow \theta$

$\leftarrow C \rightarrow$

$$\boxed{P(Y = j) \quad 1 \le j \le C}$$

$$\theta_j = P(Y=j) = \boxed{\frac{\exp(w_j^T x)}{\sum_{k=1}^{C} \exp(w_k^T x)}}$$

C weight vector w_1

w_2

$\vdots$

w_C

wc

Training $\langle x_i, y_i \rangle_{i=1}^{N}$

$$\prod_{i=1}^{N} P(y_i)$$

$w \sim \mathcal{N}(\ )$

Ber $(\ )$

$$\frac{p(w)\,p(D|w)}{\int_{w'} p(w')\,p(D|w')\,dw'}$$

$p(w|D)$ is not easy