Probabilistic Interpretation of linear models.

$$\boxed{y = \underbrace{w^T}_{\text{weight vector}} x}$$

Probabilistic linear regression

y is a random variable

x — is not a random variable,

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

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

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

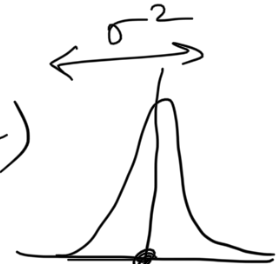
ϵ ← random variable

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

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.

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

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

$\underline{w}$

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

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

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 $\underline{N \times 1}$

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

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

Set a prior distribution for w $\underline{p(w)}$

Use the likelihood $\underline{p(D|w)} \equiv p(X, y|w)$

to calculate the posterior distribution for w: $\underline{p(w|D)} \equiv p(w|X, y)$

---

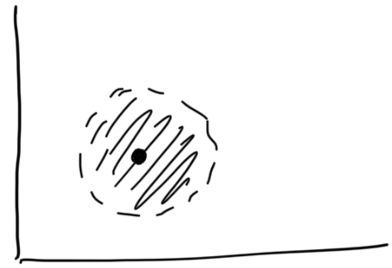
w is a $d \times 1$ vector.

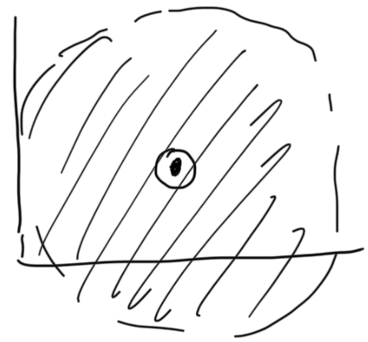
$$\boxed{w \sim \mathcal{N}(\underbrace{\mu_0}_{d \times 1}, \underbrace{\Sigma_0}_{d \times d})}$$

Encodes prior information.

<u>Example</u> let $d = 2$

$$p(w) \sim \mathcal{N}\left(\mu_0 = (3,2), \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$





$$\boxed{p(w \mid D) = p(w \mid x, y) = \frac{p(D \mid w)\, p(w)}{\int_{w'} p(D \mid w')\, p(w')\, dw'}}$$

<u>Simple prior</u> $w \sim \mathcal{N}(0, \tau^2 I)$

+ve number

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

$$p(w \mid X, y) \sim \mathcal{N}(\mu, \Sigma)$$

$$\boxed{\begin{aligned} \mu &= \left(X^T X + \frac{\sigma^2}{\tau^2} I_D\right)^{-1} X^T y \rightarrow \hat{w}_{MAP} \\ \Sigma &= \sigma^2 \left(X^T X + \frac{\sigma^2}{\tau^2} I_D\right)^{-1} \end{aligned}}$$

Looks a little bit like ridge-regression

$$J(w) = \frac{1}{2}(y - Xw)^T (y - Xw) + \frac{\lambda}{2} w^T w$$

$$w = \left(X^T X + \lambda I_D\right)^{-1} X^T y$$

For a general setting, when $\boxed{\begin{matrix} \mu_0 \neq 0 \\ \Sigma_0 \neq \tau^2 I_D \end{matrix}}$

How to predict y^* for a given x^*

① $y^* \sim \mathcal{N}\left(\hat{w}_{MLE}^T X^*, \sigma^2\right)$

② $y^* \sim \mathcal{N}\left(\hat{w}_{MAP}^T x^*, \sigma^2\right)$ $\quad p(w \mid X, y)$

For a Gaussian r.v.
Mode $\equiv$ Mean

③ Bayesian

Things to take away:

① Prob. linear regression $y^* \sim \mathcal{N}(w^T x^*, \sigma^2$

② $\hat{w}_{MLE} \equiv w_{least\ squares}$

③ $\hat{w}_{MAP}$ using a $\mathcal{N}(0, \tau^2 I)$ prior on w, is same as ridge regression estimate.

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

$\exp -\frac{1}{2\sigma^2}(y_i - w^T$

$\mathcal{N}$ can be replaced by other distributions

Generalized linear Models (GLM)

$$\boxed{y|x \sim \text{Laplace}(w^T x, b)}$$

$|y_i - w^T x_i$

↳ Robust regression.