Linear Regression with basis functions $\phi_i(x)$

$\phi(x)$ could be a
- polynomial basis function
- RBF
- FOURIER

$$y = \sum_{i=0}^{M} w_i \boxed{\phi_i(x)}$$

$$= w_0 \cdot \phi_0(x) + \dots + w_M \cdot \phi_M(x)$$

An example of non linear fct:

$$y = \mathrm{EXP}\left(\sum_i w_i \phi_i(x)\right)$$

$$\boxed{J(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - y_n)^2 + \frac{1}{2}\sum_{i=0}^{M} w_i^2}$$

Start with labeled data $\{x_n, t_n\}_{n=1}^{N}$

$$\arg\min_{w} J(w)$$

$$= \arg\max_{w} -J(w)$$

y_n

$$= \arg\max_{w} \exp(-J(w))$$

exp. fct. is monotonic

x_n are independent and identically distributed (i.i.d.)

$$\arg\max_{w}\left[\exp\left(-\frac{1}{2}\sum_n (t_n-y_n)^2 - \frac{\lambda}{2}\sum_i w_i^2\right)\right]$$

$$=\arg\max_{w}\left[\underbrace{\exp\left(-\frac{1}{2}\sum_n (t_n-y_n)^2\right)}\cdot \underbrace{\exp\left(-\frac{\lambda}{2}\sum_i w_i^2\right)}\right]$$

$$=\arg\max_{w}\left[\underbrace{\left(\prod_{n=1}^{N}\exp\left(-\frac{1}{2}(t_n-y_n)^2\right)\right)}\cdot \underbrace{\prod_{i=0}^{M}\exp\left(-\frac{\lambda}{2}w_i^2\right)}\right]$$

$$\propto \arg\max_{w}\ \underbrace{N(t|y,1)}_{\text{likelihood of my model}}\ \underbrace{N(w|0,1/\lambda)}_{\text{prior on my parameters } w}$$

samples are i.i.d.

$$= \underset{w}{\arg\max}\; p(t|w) \cdot p(w)$$

$$= \underset{w}{\arg\max}\; p(w|t) \cdot p(t)$$

Bayes' Rule

$$\propto \underset{w}{\arg\max}\; p(w|t)$$

$p(t)$ is constant

$$\underbrace{p(w|t)}_{\text{posterior prob.}} = \frac{\overbrace{p(t|w)}^{\text{likelihood}} \cdot \overbrace{p(w)}^{\text{prior prob.}}}{\underbrace{p(t)}_{\text{Evidence prob.} = \text{constant value for a given data set}}}$$

When using Least Squares obj. fct. with Ridge Regularizer, we see that is equivalent to maximizing

$$\boxed{\underset{\text{of data}}{\text{Likelihood}} \times \underset{\text{on parameter}}{\text{prior}} \propto \underset{\text{on parameter}}{\text{posterior}}}$$

$\mathcal{N}(t|y, 1)$ $\quad$ $\mathcal{N}(w|0, 1/\lambda)$

Least squares without REG.	Least squares with Regularization
$\arg\min_w J(w)$	$\arg\min_w J(w)$
$\Leftrightarrow \arg\max_w \mathcal{N}(t \mid y, 1)$	$\Leftrightarrow \arg\max_w \mathcal{N}(t \mid y, 1) \times \mathcal{N}(w \mid 0, 1/\lambda)$
In general, for any cost fct.	In general, for any cost fct.
$\arg\max_w p(t \mid w)$	$\arg\max_w p(t \mid w) \cdot p(w) \propto \arg\max_w p(w \mid t)$
Maximum Likelihood Estimation (MLE)	Maximum A Posteriori (MAP)

Least Squares with Lasso Regularization

$$J(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - y_n)^2 + \underbrace{\frac{1}{2}\lambda\sum_{i=0}^{M}|w_i|}$$

$$\underset{w}{\arg\min}\, J(w)$$

$$\iff \underset{w}{\arg\max}\, \mathcal{N}(t\,|\,y, 1)\cdot \mathcal{L}(w\,|\,0, 2/\lambda)$$

$\uparrow$ Laplacian

Example: flip a coin 3 times.

$S = \{H, T\}$

$H_i \equiv$ event of heads comes up on flip i

$$\boxed{E = H_1 \cap H_2 \cap H_3}$$

① Classical / frequentist

$$P(A) = \frac{|H|}{|S|} = \frac{3}{3} = 1$$

② Bayesian inference:

Hidden state : what coin was used?

$$p(E \mid fair) = P(H_1 \mid fair) \cdot P(H_2 \mid fair) \cdot p(H_3 \mid f)$$

$$= \left(\frac{1}{2}\right)^3$$

H_i's are conditionally independent

Heads $= 1$ $\quad S = \{1, 0\}$

Tails $= 0$

$P(x=1 \mid \mu) = \mu$, e.g. if coin is fair $\mu = \frac{1}{2}$

$P(x=0 \mid \mu) = 1-\mu$

Data Likelihood:

$$\boxed{P(x \mid \mu) = \mu^{x} \cdot (1-\mu)^{1-x}}$$

Bernoulli distribution

Ex. Suppose $x = \{1, 0, 1\}$, $p(x \mid \mu) = \mu \cdot (1-\mu) \cdot \mu$

$= \mu^2 \cdot (1-\mu)$

MLE

$$E = \{x_1, x_2, \dots, x_n\}$$

$$\begin{aligned}
P(E|\mu) &= P(x_1 \cap x_2 \cap \dots \cap x_n | \mu) \\
&= P(x_1|\mu) \cdot \dots \cdot P(x_n|\mu) \\
&= \prod_{n=1}^{N} P(x_i|\mu) \\
&= \prod_{i=1}^{N} \mu^{x_i} (1-\mu)^{1-x_i} \leftarrow \text{data Likelihood}
\end{aligned}$$

$$\begin{aligned}
&\arg\max_{\mu} P(E|\mu) \\
&= \arg\max_{\mu} \ln(P(E|\mu)) \\
&= \arg\max_{\mu} \sum_{n=1}^{N} x_n \cdot \ln(\mu) + (1-x_n) \cdot \ln(1-\mu)
\end{aligned}$$

$$\ln(P(E|\mu)) = \sum_{n=1}^{N} \left[\underline{\underline{x_n \ln \mu}} + (1-x_n) \cdot \underline{\underline{\ln(1-\mu)}} \right]$$

$$\frac{\partial \ln(P(E|\mu))}{\partial \mu} = 0$$

$$\Leftrightarrow \underbrace{\sum_{n=1}^{N} \left[x_n \cdot \frac{1}{\mu}\right.} + \underbrace{\left.(1-x_n) \cdot \left(-\frac{1}{1-\mu}\right)\right]} = 0$$

$$\Leftrightarrow \frac{1}{\mu} \cdot \sum_n x_n - \frac{1}{1-\mu} \cdot \sum_n (1-x_n) = 0$$

$$\Leftrightarrow \boxed{\mu = \frac{1}{N} \cdot \sum_n \underline{\underline{x_n}}}$$

$x_n \in \{0, 1\}$

tails heads

MLE is purely data driven!

$$p\left(\{x_i\}_{i=1}^{N}\right) \underset{\text{i.i.d.}}{=} p(x_1)\cdot p(x_2)\cdots p(x_n)$$

Gaussian dist : $f(x \mid \mu, \sigma) = \dfrac{1}{\underbrace{\sqrt{2\pi\sigma^2}}_{X}} \cdot \exp\left(-\dfrac{1}{2}\dfrac{(x-\mu)^2}{\sigma^2}\right)$