

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

$\phi(x)$ could be a

- | polynomial basis func,
- | RBF
- | FOURIER

$$y = \sum_{i=0}^M w_i \phi_i(x) \\ = w_0 \cdot \phi_0(x) + \dots + w_M \cdot \phi_M(x)$$

An example of non linear fct:

$$y = \text{Exp} \left(\sum_i w_i \phi_i(x) \right)$$

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

$$\arg \min_w J(w)$$

$$= \arg \max_w -J(w)$$

$$= \arg \max_w \exp(-J(w))$$

$\nearrow w$
exp.fct. is
monotonic

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

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

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

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

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

$\rightarrow \arg \max_w N(t|y, 1) N(w|0, 1/\lambda)$

samples
are i.i.d.

likelihood
of my
model

prior
on my
parameters w

$$\begin{aligned}
 &= \arg \max_{w_0} p(t|w) \cdot p(w) \\
 &= \arg \max_{w_0} p(w|t) \cdot p(t) \\
 &\xrightarrow{\text{Bayes' Rule}} \propto \arg \max_w p(w|t) \\
 &p(w|t) = \frac{\text{likelihood} \cdot \text{prior prob.}}{\text{Evidence prob.}} \\
 &\quad \left[\begin{array}{l} \text{posterior prob.} \\ \text{Evidence prob.} = \text{constant value for a given data set} \end{array} \right] \quad p(t) \text{ is constant}
 \end{aligned}$$

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

$$\left[\begin{array}{l} \text{Likelihood of data} \times \text{prior on parameter} \\ N(t|y, 1) \quad N(w|0, 1/\lambda) \end{array} \right] \propto \text{posterior on parameter}$$

Least Squares
without $\frac{1}{2}\|w\|^2$

Least Squares
with regularization

$$\arg \min_w J(w)$$

$$\Leftrightarrow \arg \max_w N(t|y, 1)$$

In general, for
any cost fct.

$$\arg \max_w p(t|w)$$

Maximum Likelihood
Estimation (MLE)

$$\arg \min_w J(w)$$

$$\Leftrightarrow \arg \max_w N(t|y, 1) \times N(w|\sigma_1^2/\lambda)$$

In general, for any cost.
fct.

$$\arg \max_w p(t|w) \cdot p(w)$$
$$\propto \arg \max_w p(w|t)$$

Maximum A Posteriori
(MAP)

Least Squares with

Lasso Regularization

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

$$\arg \min_w J(w)$$

$$\Leftrightarrow \arg \max_w N(t | y, 1) \cdot \mathcal{L}(w | 0, \lambda)$$

↑
Laplacian

Example: flip a coin 3 times.

$$S = \{H, T\}$$

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

H_i = event of heads
comes up on
flip i

① Classical / frequentist

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

② Bayesian inference:

Hidden state : What coin was used?

$$P(E | \text{fair}) = P(H_1 | \text{fair}).P(H_2 | \text{fair}).P(H_3 | \text{fair})$$

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

H_i 's are conditionally independent

Heads = 1
Tails = 0

$$S = \underline{\{1, 0\}}$$

$$P(X=1|\mu) = \mu \quad , \text{ e.g. if } \mu = \frac{1}{2}$$

$$P(X=0|\mu) = 1-\mu \quad \text{coin is fair}$$

Data Likelihood:

$$P(X|\mu) = \mu^x \cdot (1-\mu)^{1-x}$$

Bernoulli distribution

Ex. Suppose $X = \{1, 0, 1\}$, $P(X|\mu) = \mu \cdot (1-\mu) \cdot \mu$
 $= \mu^2 \cdot (1-\mu)$

MLE

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

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

$$\arg \max_{\mu} P(E|\mu)$$

$$= \arg \max_{\mu} \ln(P(E|\mu))$$

$$= \arg \max_{\mu} \sum_{n=1}^N x_n \cdot \ln(\mu) + (1-x_n) \cdot \ln(1-\mu)$$

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

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

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

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

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

$x_n \in \{0, 1\}$
 ↑ ↑
 tails heads

MLE is purely data driven!

$$p(\{x_i\}_{i=1}^n) = \overbrace{p(x_1) \cdot p(x_2) \cdots p(x_n)}^{i.i.d.}$$

Gaussian : $f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$

alist

~~$\sqrt{2\pi\sigma^2}$~~