Model / Mapper : $y = f(\phi(x), w)$

FEATURES : $\phi(x)$

w : parameters (unknowns)

target : t

OBJECTIVE FUNCTION :

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} (t_i - y_i)^2$$

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

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

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

$\exp(\cdot)$ is monotonic

$$= \arg\max_{w} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(t_i - y_i)^2\right)$$

$$= \arg\max_{w} \prod_{i=1}^{N} \exp\left(-\frac{1}{2}(t_i - y_i)^2\right)$$

$\exp(a+b) = \exp(a)\cdot\exp(b)$

$\sim G(y_i, 1)$

$$= \arg\max_{w} \prod_{i=1}^{N} G(t_i; y_i, 1)$$

$\mathcal{L}^{o} \equiv$ OBSERVED DATA LIKELIHOOD

Reminder

Gaussian R.V.

$$X \sim G(\mu, \sigma^2)$$

mean, variance

PDF: $$f_X(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

OBSERVED DATA LIKELIHOOD :

$$\mathcal{L}^{o} = \prod_{i=1}^{N} G(t_i; y_i, 1)$$

where $y_i = f(w, \phi(x_i))$

DATASET : $\{(x_i, t_i)\}_{i=1}^{N}$ ASSUME x_i ARE i.i.d.

i.i.d. = independent and identically distributed

$$\mathcal{L}^{o} = P(x_1, x_2, x_3, \ldots, x_N | w)$$

$$= P(x_1|w) \cdot P(x_2|w) \cdot \ldots \cdot P(x_N|w)$$

x_i are conditionally independent

$$= \prod_{i=1}^{N} P(x_i|w)$$

$$= \prod_{i=1}^{N} G(t_i; y_i, 1)$$

Now adding a regularizer to obj. fct:

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} (t_i - y_i)^2 + \frac{\lambda}{2} \cdot \sum_{j} w_j^2$$

RIDGE Reg.

$$\arg\min_w J(w) = \arg\max_w \left(-J(w)\right)$$

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

$$= \arg\max_w \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (t_i - y_i)^2 - \frac{\lambda}{2} \sum_{j} w_j^2\right)$$

$$= \arg\max_w \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (t_i - y_i)^2\right) \cdot \exp\left(-\frac{\lambda}{2} \sum_{j} w_j^2\right)$$

$= (*)$

$$(*) = \arg\max_{w} \left[\prod_{i=1}^{N} \exp\left(-\tfrac{1}{2}(t_i - y_i)^2\right) \right] \left[\prod_{j} \exp\left(-\tfrac{\lambda}{2} w_j^2\right) \right]$$

$\sim G(t_i; y_i, 1)$ — OBSERVED DATA

LIKELIHOOD $\equiv \mathcal{L}$

$\equiv P(t|w)$

$\sim G(w_j; 0, 1/\lambda)$ — PRIOR PROBABILITY

I.E. PROBABILISTIC MODEL DESCRIBING THE UNKNOWN PARAMETERS w

$\equiv P(w)$

$$= \arg\max_{w} P(t|w) \cdot P(w)$$

$$= \arg\max_{w} P(w|t) \cdot P(t)$$

↙ BAYES' THEOREM

$$\propto \arg\max_{w} P(w|t)$$

↙ proportional

Reminder:

BAYES' THEOREM:

$$P(w|t) = \frac{P(t|w) \times P(w)}{P(t)}$$

- $P(w|t)$: posterior
- $P(t|w)$: DATA likelihood
- $P(w)$: prior prob
- $P(t)$: evidence prob.

OBJ. FCT with LASSO REGULARIZER

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} (t_i - y_i)^2 + \lambda \sum_j |w_j|$$

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

$$\vdots$$

$$= \arg\max_w \left[\prod_{i=1}^{N} \exp\left(-\frac{1}{2}(t_i - y_i)^2\right) \right] \left[\prod_j \exp\left(-\lambda |w_j|\right) \right]$$

$$G(t_i; y_i, 1) \qquad \mathcal{L}(w_j; 0, 1/\lambda)$$

reminder

Laplacian (μ, b)

$$X \sim \text{Lap.}(\mu, b)$$

$$f_X(x) = \frac{1}{2b} \cdot \exp\left(-\frac{|x-\mu|}{b}\right)$$

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

→ COMPLETELY DATA-DRIVEN

→ it requires a lot more data to make correct estimations for w

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

MAXIMUM A POSTERIORI (MAP)

→ it requires a prior prob.

→ prior will inject beliefs as to what values are most likely for w.

→ if prior is selected correctly then w_{MAP} is better than w_{MLE}.

→ if not, w_{MAP} will typically require a lot more data to compensate.

$$\arg\max_w P(t|w) \cdot P(w)$$

Example: flip a coin 3 times

and we observe $H_1 \cap H_2 \cap H_3 = E$

outcome : $E = H_1 \cap H_2 \cap H_3$

MLE (Frequentist approach):

$$P(\text{Heads}) = \frac{\#\,\text{Heads}}{\#\,\text{flips}} = \frac{3}{3} = 1$$

MAP (Bayesian approach):

1st start with a set of prior beliefs. Beliefs: fair coin or 2-headed coin

$P(\mu = \frac{1}{2} \equiv \text{fair} \mid E)$

$P(\mu = 1 \mid E)$

$$P(\mu = \tfrac{1}{2} \mid E) \underset{\mu=1}{\overset{\mu=\frac{1}{2}}{\gtrless}} P(\mu = 1 \mid E)$$

$$P(\mu \mid E) = \frac{P(E \mid \mu) \times P(\mu)}{P(E)}$$

$E = H_1 \cap H_2 \cap H_3$

$E = H_1$

$$P(\mu \mid H) = \frac{P(H \mid \mu) \times P(\mu)}{P(H)}$$

$\mu = \frac{1}{2}$: $= \dfrac{1/2 \times 1/2}{P(H \mid \mu = 1/2) \times P(\mu = 1/2) + P(H \mid \mu = 1) \times P(\mu = 1)}$

Law of Total Probability $= \dfrac{1/2 \times 1/2}{1/2 \times 1/2 + 1 \times 1/2}$

$$= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{1}{3}$$

Similarly, $P(\mu = 1 \mid H) = \frac{2}{3}$

Since $P(\mu = 1 \mid H) > P(\mu = \frac{1}{2} \mid H)$

then Estimate is $\mu = 1$ i.e. a 2-headed coin is most likely to have generated this outcome.

HEADS $\rightarrow$ 1

TAILS $\rightarrow$ 0

$S = \{0, 1\}$

$\mu \equiv$ UNKNOWN parameter
prob. of flipping heads

$x \equiv$ input data (0 or 1)

$P(x = 1 | \mu) = \mu$

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

BERNOULLI R.V.

$$P(x|\mu) = \begin{cases} 1-\mu, & x = 0 \\ \mu, & x = 1 \\ 0, & \text{otherwise} \end{cases}$$

$x \sim BER(\mu)$

TRAINING DATASET : $\{x_i\}_{i=1}^{N}$, $x_i \in \{0,1\}$

STARTING with MLE

STEP 1

OBSERVED DATA LIKELIHOOD:

$$\mathcal{L}^{o} = P(x_1 \cap x_2 \cap \dots \cap x_N | \mu)$$

$$= P_1(x_1|\mu) \cdot P_2(x_2|\mu) \cdot \dots \cdot P_N(x_N|\mu)$$

x_i's are conditionally independent

$$= \prod_{i=1}^{N} P(x_i|\mu)$$

identically distributed

$$= \prod_{i=1}^{N} \mu^{x_i} \cdot (1-\mu)^{1-x_i}$$

assuming $P(\cdot)$ is a Bernoulli

Example : $x = \{1, 0, 1\}$

$$\mathcal{L}^0 = \mu \cdot (1-\mu) \cdot \mu = \mu^2 \cdot (1-\mu)$$

$$\arg\max_{\mu} \mathcal{L}^0$$

$$= \arg\max_{\mu} \left[\prod_{i=1}^{N} \mu^{x_i} \cdot (1-\mu)^{1-x_i} \right]$$

STEP 2

Apply the log to $\mathcal{L}^{\circ}$.

$$\mathcal{L} = \ln(\mathcal{L}^{\circ})$$

$$= \ln\left[\prod_{i=1}^{N} \mu^{x_i} \cdot (1-\mu)^{1-x_i} \right]$$

$$= \sum_{i=1}^{N} \ln(\mu^{x_i} \cdot (1-\mu)^{1-x_i})]$$

$$= \sum_{i=1}^{N} \left(\ln(\mu^{x_i}) + \ln((1-\mu)^{1-x_i}) \right)$$

$$= \sum_{i=1}^{N} \left(x_i \cdot \ln(\mu) + (1-x_i) \cdot \ln(1-\mu) \right)$$

$$\mathcal{L} = \sum_{i=1}^{N} [x_i \cdot \ln(\mu) + (1-x_i) \cdot \ln(1-\mu)]$$

STEP 3

Solve for μ: $\frac{\partial \mathcal{L}}{\partial \mu} = 0$

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

$$\Leftrightarrow \sum_{i=1}^{N} \left(x_i (1-\mu) + \mu \cdot (x_i - 1) \right) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} \left(x_i - x_i \mu + \mu x_i - \mu \right) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} x_i - \underbrace{\sum_{i=1}^{N} \mu}_{= N \cdot \mu} = 0$$

$$\Leftrightarrow \boxed{\mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

$\equiv \frac{\#\text{heads}}{\#\text{flips}}$

$\equiv$ relative freq. for outcome Heads

MAP APPROACH

1st assume a prior probability

E.g., BETA DISTRIBUTION as prior

$$\text{Beta}(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)} \cdot x^{\alpha-1} \cdot (1-x)^{\beta-1}$$

where $\Gamma(x) = (x-1)!$

and $\alpha, \beta > 0$

STEP 1 OBSERVED DATA LIKELIHOOD

$$\mathcal{L}^{o} = P(x|\mu) \times P(\mu|\alpha,\beta)$$

$$= \left(\prod_{i=1}^{N} \mu^{x_i} (1-\mu)^{1-x_i} \right) \cdot \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \mu^{\alpha-1} \cdot (1-\mu)^{\beta-1} \right)$$

$$= \mu^{\sum_{i=1}^{N} x_i} \cdot (1-\mu)^{\sum_{i=1}^{N}(1-x_i)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \mu^{\alpha-1} \cdot (1-\mu)^{\beta-1}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \mu^{\sum_{i=1}^{N} x_i + \alpha - 1} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i + \beta - 1}$$

$$\propto \mu^{\sum_{i=1}^{N} x_i + \alpha - 1} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i + \beta - 1}$$

STEP 2

$$\mathcal{L} = \ln(\mathcal{L}^{\circ})$$

$$= \left(\sum_{i=1}^{N} x_i + \alpha - 1\right) \cdot \ln(\mu) + \left(N - \sum_{i=1}^{N} x_i + \beta - 1\right) \cdot \ln(1-\mu)$$

STEP 3

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \iff \left(\sum_{i=1}^{N} x_i + \alpha - 1\right) \cdot \frac{1}{\mu}$$

$$- \left(N - \sum_{i=1}^{N} x_i + \beta - 1\right) \cdot \frac{1}{1-\mu} = 0$$

$$\vdots$$

$$\Leftrightarrow \quad \mu_{MAP} = \frac{\sum_{i=1}^{N} x_i + \alpha - 1}{N + \alpha + \beta - 2}$$

$N \equiv$ # samples

In MAP

BETA: $P(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \mu^{\alpha-1}(1-\mu)^{\beta-1}$

PRIOR

POSTERIOR $P(\mu|x) \propto \mu^{\sum_{i=1}^{N} x_i + \alpha - 1}(1-\mu)^{N - \sum_{i=1}^{N} x_i + \beta - 1}$

When the prior and the posterior have the same parametric form (minus constants) they are said to have a Conjugate Prior Relationship

This means that we can use the posterior as the new prior in an online fashion:

Other conjugate prior relationships:

	Likelihood	Prior	Posterior
(1)	Gaussian	Gaussian	Gaussian
(2)	Bernoulli	Beta	Beta
⋮			

many more.

For a new batch of data

iteration (t) :

$$\alpha^{(t+1)} \longleftarrow \alpha^{(t)} + \sum_{i=1}^{N} x_i$$

$$\beta^{(t+1)} \longleftarrow \beta^{(t)} + N - \sum_{i=1}^{N} x_i$$

Review:

Multivariate Gaussian

Bivariate Gaussian

↳ covariance