Foundations of Machine Learning AI2000 and AI5000

FoML-05
Maximum A Posteriori
Fully Bayesian treatment

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

So far in FoML

- What is ML and the learning paradigms

- Probability refresher

- Maximum Likelihood Principle

Maximum A Posteriori

Maximum A Posteriori

- Given - Dataset of N independent observations D

Maximum A Posteriori

- Given - Dataset of N independent observations D

- ML estimate - w that maximizes the data likelihood

$$\mathbf{w}_{ML} = \arg\max_{w} \; P(D|w)$$

Maximum A Posteriori

- Given - Dataset of N independent observations D $= \{x_1, x_2 \dots x_N\}$

- MAP estimate - choose most probable w given data

Maximum A Posteriori

- Given - Dataset of N independent observations D

- MAP estimate - choose most probable w given data

$$\mathbf{w}_{MAP} = \arg\max_{w} P\left(w / D\right)$$

MAP - Curve Fitting

- Given data D

$$D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$$

MAP - Curve Fitting

- Given data D $\qquad D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Model $\qquad p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\dfrac{\beta}{2\pi}} \; e^{-\frac{\beta}{2}(t - y(x,w))^2}$

भारतीय सांकेतिक विज्ञान संस्थ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

MAP - Curve Fitting
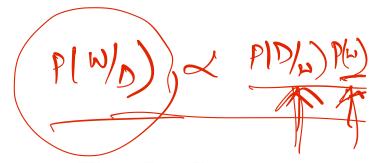
$P(w/D)$

- Given data D
$$D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$$

- Model
$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} \quad p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \beta)$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

MAP - Curve Fitting

$$P(w/D) \propto \overline{P(D/w) P(w)}$$

- Given data D $\quad D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Model $\quad p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} \ p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \beta)$$

Given a prior $P(\omega|\alpha)$ the posterior distribution becomes

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \beta, \alpha) = \frac{P(t|w, x, \beta) \ P(w|\alpha)}{P(t|x, \beta, \alpha)}$$

$$\propto P(t/x, w, \beta) \ P(w/\alpha)$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

MAP - Curve Fitting

- MAP estimate - for convenience apply log

$$\mathbf{w}_{MAP} = \operatorname*{argmax}_{w} \left[\log P\left(t \middle/ x, w, \beta\right) + \log P(w|\alpha) - \underbrace{\log P(t|x, \beta, \alpha)}_{\text{independent of } w} \right]$$

MAP - Curve Fitting

- Assuming Gaussian Prior and independence on parameters $\mathbf{w} \in \mathbb{R}^{\mathbf{M}}$

$$w_i \sim \mathcal{N}(0, \alpha^{-1})$$
$$i = \{1, 2 \ldots m\}$$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{\mathbf{M}} \mathcal{N}(\mathbf{w_i}|\mathbf{0}, \alpha^{-1})$$

$$= \prod_{i=1}^{m} \sqrt{\frac{\alpha}{2\pi}} \, e^{-\frac{\alpha}{2} w_i^2} = \left(\frac{\alpha}{2\pi}\right)^{m/2} \prod_{i=1}^{m} e^{-\frac{\alpha}{2} w_i^2}$$

$$= \left(\frac{\alpha}{2\pi}\right)^{m/2} e^{-\frac{\alpha}{2} \sum_{i=1}^{M} w_i^2} = \left(\frac{\alpha}{2\pi}\right)^{m/2} e^{-\frac{\alpha}{2} w^T w}$$

MAP - Curve Fitting

$$\mathbf{w_{MAP}} = \arg\min -\log \mathbf{p(w|x,t,}\beta,\alpha) = \arg\min -\log \mathbf{p(t|x,w,}\beta) -\log \mathbf{p(w|}\alpha)$$

$$= \arg\min_{w} \quad -\log p(t|x,w,\beta) - \log\left(\left(\frac{\alpha}{2\pi}\right)^{m/2} e^{-\frac{\alpha}{2}w^T w}\right)$$

$$= \arg\min_{w} \quad -\log \underbrace{p(t|x,w,\beta)}_{\prod_{i=1}^{N}\sqrt{\frac{\beta}{2\pi}} \, e^{-\frac{\beta}{2}\left[t_i - y(x_i,w)\right]^2}} - \underbrace{\frac{m}{2}\log\frac{\alpha}{2\pi}}_{\text{Independent of } w} + \frac{\alpha}{2}w^T w$$

$$w_{MAP} = \arg\min_{w}\left[\frac{\beta}{2}\sum_{i=1}^{N}\left[t_i - y(x_i,w)\right]^2 + \frac{\alpha}{2}w^T w\right]$$

MAP - Curve Fitting

- Predictive distribution

$$p(t' \mid x', \beta) = \mathcal{N}\left(t' \mid y(w_{MAP}, x'), \beta\right)$$

point estimate $\quad t' = y(w_{MAP}, x')$

Bayesian Prediction

So far

- Our estimates for w have been point estimates
  - ML and MAP

So far

- Our estimates for w have been point estimates
  - ML and MAP
  - Regarded as frequentist because they discard 'uncertainty' about the w

Fully Bayesian

- An approach that relies on consistent application of sum and product rules of probability at all levels of modeling

Fully Bayesian

- Given a prior belief $p(\mathbf{w}|\alpha)$ over w, and data D

Fully Bayesian

- Given a prior belief $p(\mathbf{w}|\alpha)$ over w, and data D

- We are interested in the posterior

$$p(\mathbf{w}|\mathbf{D}) = \frac{P(D|w)\, P(w)}{P(D)}$$

Fully Bayesian

- The predictive distribution becomes

$$p(x'|D) = \int P(x',w|D) \, dw = \int P(x'|w,D) \, P(w|D) \, dw$$

Fully Bayesian

- Curve fitting example

Fully Bayesian

- Curve fitting example

- Given training data (x, t)

Fully Bayesian

- Curve fitting example

- Given training data (**x, t**) and a test sample x

Fully Bayesian

- Curve fitting example

- Given training data (**x**, **t**) and a test sample x

- Goal - predict the value of t

Fully Bayesian

- Curve fitting example

- Given training data (**x**, **t**) and a test sample x

- Goal - predict the value of t

We wish to evaluate the predictive distribution

$$\mathcal{N}\left(t \mid y(w_{ML}, \tilde{x}), \beta^{-1}\right)$$

$$p(t|x, \mathbf{x}, \mathbf{t})$$

Fully Bayesian

$$w \sim N(0, \Sigma)$$

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, d\mathbf{w}.$$

$$= \int P(t, w / x, \underline{x}, \underline{t}) \, dw$$

$$= \int P(t / x \underline{x}, \underline{t}, w) \; P\left(\underline{w} / x, \underline{x}, \underline{t}\right) \, dw$$

$$= \int P(t / x, \underline{w}) \cdot P(\underline{w} / \underline{x}, \underline{t}) \, dw$$

Fully Bayesian

- Advantages
  - Inclusion of the prior knowledge
  - Represents uncertainty in t' due to the target noise and uncertainty over w

Fully Bayesian

- Advantages
  - Inclusion of the prior knowledge
  - Represents uncertainty in t' due to the target noise and uncertainty over w

- Disadvantages
  - Posterior is hard to compute analytically
  - Prior is often a mathematical convenience

Rough work

Next
Linear Models - Regression