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

Data-driven Intelligence & Learning Lab

So far in FoML

- What is ML and the learning paradigms
- Probability refresher
- Maximum Likelihood Principle

Maximum A Posteriori

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

Data-driven Intelligence & Learning Lab

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

$$\mathbf{w}_{ML} = \underset{w}{\arg\max}\; p(D|w)$$

Maximum A Posteriori

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

- 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

$$\mathbf{w}_{MAP} = \underset{w}{\text{argmax}} \quad P(w/D)$$

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

DiL
Data-driven Intelligence
& Learning Lab

MAP - Curve Fitting

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

MAP - Curve Fitting

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

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

$p(w|D)$

MAP - Curve Fitting

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

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

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

MAP - Curve Fitting

$P(w/D) \propto P(D/w)\,P(w)$

- Given data D $D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$

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

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

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

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

MAP - Curve Fitting

- MAP estimate - for convenience apply log

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

MAP - Curve Fitting

$w_i \sim \mathcal{N}(0, \alpha^{-1})$

$i = \{1, 2 \dots m\}$

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

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

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

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

MAP - Curve Fitting

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

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

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

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

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

DiL
Data-driven Intelligence & Learning Lab

MAP - Curve Fitting

- Predictive distribution

Bayesian Prediction

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

DiL
Data-driven Intelligence
& Learning Lab

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

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

Data-driven Intelligence & Learning Lab

Fully Bayesian

- Given a prior belief $p(\mathbf{w}|\alpha)$ over w, and data D
- We are interested in the posterior

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

Fully Bayesian

- The predictive distribution becomes

$$p(x'|D) =$$

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

Fully Bayesian

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

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