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

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

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

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

MAP - Curve Fitting

- Given data D

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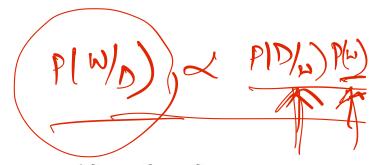
MAP - Curve Fitting

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

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

MAP - Curve Fitting

$P(w/D)$

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

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

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

MAP - Curve Fitting

$P(W/D) \propto \dfrac{P(D/_w) P(w)}{}$

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

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

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

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

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

MAP - Curve Fitting

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

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

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

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

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

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

MAP - Curve Fitting

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

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

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

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

MAP - Curve Fitting

- Predictive distribution

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

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

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

भारतीय सांकेतिक विज्ञान संस्था हैदराबाद్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
**Data-driven Intelligence
& Learning Lab**

Rough work

Next
Linear Models - Regression