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

$$\mathbf{w}_{ML} =$$

Maximum A Posteriori

- Given - Dataset of N independent observations D

- 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

$$\mathbf{w}_{MAP} =$$

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

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

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

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

MAP - Curve Fitting

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

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

MAP - Curve Fitting

- MAP estimate - for convenience apply log

$$\mathbf{w}_{MAP} =$$

MAP - Curve Fitting

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

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

MAP - Curve Fitting

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

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

Rough work

Next
Linear Models - Regression