Fundamentals of Machine Learning

MODEL FITTING & PARAMETER ESTIMATION

MAXIMUM A POSTERIORI ESTIMATION

Amit K Roy-Chowdhury

Acknowledgments: Adapted from slides at https://probml.github.io/pml-book/teaching1.html by Prof. Saw Shier Nee
Adapted from slides of Prof. Fowler @ Binghamton University;
Modified by Prof. Roy-Chowdhury @ UC Riverside

Model Fitting / Training

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \boxed{\mathcal{L}(\boldsymbol{\theta})}$$

Loss function / objective function

Maximum likelihood estimation

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta})$$

Estimated Parameter

Probability

Model

Since most optimization algorithms are designed to minimize cost functions, we redefine the objective function to be the (conditional) negative log likelihood or NLL and we minimize NLL

$$\text{NLL}(\boldsymbol{\theta}) \triangleq -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta})$$

Notation and Form for MAP

Notation: $\hat{\theta}_{MAP}$ maximizes the posterior PDF

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \, p(\theta \,|\, \mathbf{x})$$

Equivalent Form (via Bayes' Rule): $\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \, [\, p(\mathbf{x} \,|\, \theta) \; p(\theta)]}$

Proof: Use $p(\theta \,|\, \mathbf{x}) = \dfrac{p(\mathbf{x} \,|\, \theta) p(\theta)}{p(\mathbf{x})}$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \left[\frac{p(\mathbf{x} \,|\, \theta) p(\theta)}{\boxed{p(\mathbf{x})}} \right] = \arg\max_{\theta} \, [\, p(\mathbf{x} \,|\, \theta) \; p(\theta)]$$

MAP Example

Let X be a continuous random variable with the following PDF:

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Also, suppose that

$$Y \mid X = x \quad \sim \quad Geometric(x).$$

Find the MAP estimate of X given $Y = 3$.

We know that $Y \mid X = x \quad \sim \quad Geometric(x)$, so

$$P_{Y|X}(y|x) = x(1-x)^{y-1}, \quad \text{for } y = 1, 2, \cdots.$$

Therefore,

$$P_{Y|X}(3|x) = x(1-x)^2.$$

We need to find the value of $x \in [0, 1]$ that maximizes

$$\begin{aligned} P_{Y|X}(y|x) f_X(x) &= x(1-x)^2 \cdot 2x \\ &= 2x^2(1-x)^2. \end{aligned}$$

We can find the maximizing value by differentiation. We obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2(1-x)^2\right] = 2x(1-x)^2 - 2(1-x)x^2 = 0.$$

Solving for x (and checking for maximization criteria), we obtain the MAP estimate as

$$\hat{x}_{MAP} = \frac{1}{2}.$$

For Bernoulli trials, geometric rv is the number of trials until first success.

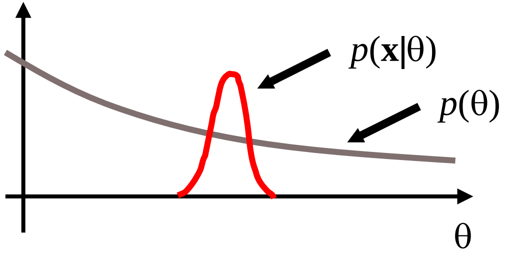
Ack: https://www.probabilitycourse.com/

“Bayesian MLE”

Recall… As we keep getting good data, $p(\boldsymbol{\theta}|\mathbf{x})$ becomes more concentrated as a function of $\boldsymbol{\theta}$. But… since:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta} \,|\, \mathbf{x}) = \arg\max_{\boldsymbol{\theta}} \,[\, p(\mathbf{x} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta})]$$

… $p(\mathbf{x}|\theta)$ should also become more concentrated as a function of $\boldsymbol{\theta}$.



$$\underbrace{\arg\max_{\boldsymbol{\theta}} \,[\, p(\mathbf{x} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta})]}_{\text{MAP}} \;\approx\; \underbrace{\arg\max_{\boldsymbol{\theta}} \; p(\mathbf{x} \,|\, \boldsymbol{\theta})}_{\text{“Bayesian MLE”}}$$