Model fitting and parameter estimation

David J. H. Shih

Intended learning outcomes

- Fit models to data by estimating model parameters

Model fitting

Use data $\boldsymbol{x}$ to estimate parameters $\boldsymbol{\theta}$ of the model.

Maximum likelihoood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right).$$

Maximum *a posteriori* (MAP)

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta}\right).$$

Full Bayesian

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right) = \frac{p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{x}\right)}.$$

Parameter estimates

Point estimate

Posterior for discrete random variable

Posterior for continuous random variable

Generating new samples

Use $\hat{\boldsymbol{\theta}}$ or $p\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right)$ to generate a new data point $\tilde{x}$.

With point estimate from MLE or MAP

$$\tilde{x} \sim p\left(\tilde{x} \mid \hat{\boldsymbol{\theta}}\right).$$

With posterior predictive distribution

$$p\left(\tilde{x} \mid \boldsymbol{x}\right) = \int_{\boldsymbol{\Theta}} p\left(\tilde{x} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right) \, d\boldsymbol{\theta}$$

$$\tilde{x} \sim p\left(\tilde{x} \mid \boldsymbol{x}\right).$$

With posterior distribution

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{x}), \quad \tilde{x} \sim p\left(\tilde{x} \mid \boldsymbol{\theta}\right).$$

Fitting discriminative models

Use data $(\boldsymbol{x}, \boldsymbol{y})$ to estimate parameters $\boldsymbol{\theta}$ of the model and to predict new outcome $\hat{y}$ given new data point $\tilde{x}$.

MLE

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\right).$$

MAP

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x}\right) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta}\right).$$

Full Bayesian

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x}\right) = \frac{p\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{y} \mid \boldsymbol{x}\right)}.$$

Predicting new outcomes

With point estimate from MLE or MAP

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p\left(y \mid \tilde{x}, \hat{\boldsymbol{\theta}}\right).$$

With posterior predictive distribution

$$p\left(\tilde{y} \mid \tilde{x}\right) = \int_{\boldsymbol{\Theta}} p\left(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x}\right) d\boldsymbol{\theta},$$
$$\tilde{y} \sim p\left(\tilde{y} \mid \tilde{x}\right).$$

With posterior distribution

$$\boldsymbol{\theta} \sim p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x}\right), \quad \tilde{y} \sim p\left(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}\right).$$

Full Bayesian

Basic model

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{x}\right) = \frac{p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{x}\right)}.$$

Discriminative model

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x}\right) = \frac{p\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{y} \mid \boldsymbol{x}\right)}.$$

Generative model

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}\right) = \frac{p\left(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}\right)\, p\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)}{p\left(\boldsymbol{x} \mid \boldsymbol{y}\right)}.$$

Maximum likelihood

$$L(\theta) = p(\boldsymbol{x} \mid \theta) \quad \text{or} \quad L(\theta) = p(\boldsymbol{y} \mid \boldsymbol{x}, \theta) \quad \text{or} \quad L(\theta) = p(\boldsymbol{x} \mid \boldsymbol{y}, \theta)$$

$$p(\boldsymbol{x} \mid \theta) = \prod_i p(x_i \mid \theta)$$

$$\ell(\theta) = \log L(\theta)$$

$$\log p(\boldsymbol{x} \mid \theta) = \sum_i \log p(x_i \mid \theta)$$

$$\frac{d\ell(\theta)}{d\theta} = 0$$

Maximum *a posteriori*

$$p(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\ p(\theta)$$

or

$$p(\theta \mid \boldsymbol{y}, \boldsymbol{x}) \propto p(\boldsymbol{y} \mid \boldsymbol{x}, \theta)\ p(\theta)$$

or

$$p(\theta \mid \boldsymbol{x}, \boldsymbol{y}) \propto p(\boldsymbol{x} \mid \boldsymbol{y}, \theta)\ p(\theta \mid y)$$

$$\ell(\theta) = \log p(\theta \mid \boldsymbol{x})$$

$$\log p(\boldsymbol{x} \mid \theta) = \sum_i \log p(x_i \mid \theta)$$

$$\frac{d\ell(\theta)}{d\theta} = 0$$

MLE vs. MAP vs. full Bayesian

MLE is an approximation of MAP with a uniform prior.

MAP is an approximation of the posterior.

Summary

"All models are wrong. Some are useful." - George Box

Blais 2014, chapter 6.

Intended learning outcomes

- Fit models to data by estimating model parameters