Bayesian inference and prediction in finite regression models

Carl Edward Rasmussen

October 10th, 2023

Key concepts

Bayesian inference in finite, parametric models

- we contrast maximum likelihood with Bayesian inference
- when both prior and likelihood are Gaussian, all calculations are tractable
  - the posterior on the parameters is Gaussian
  - the predictive distribution is Gaussian
  - the marginal likelihood is tractable
- we observe the contrast
  - in maximum likelihood the data fit gets better with larger models (overfitting)
  - the marginal likelihood prefers an intermediate model size (Occam's Razor)

Maximum likelihood, parametric model

Supervised parametric learning:

- data: $\mathbf{x}, \mathbf{y}$
- model $\mathcal{M}$: $y = f_{\boldsymbol{w}}(x) + \varepsilon$

Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{w}, \mathcal{M}) \;\propto\; \prod_{n=1}^{N} \exp(-\tfrac{1}{2}(y_n - f_w(x_n))^2/\sigma^2_{\mathrm{noise}}).$$

Maximize the likelihood:

$$\boldsymbol{w}_{\mathrm{ML}} \;=\; \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}, \boldsymbol{w}, \mathcal{M}).$$

Make predictions, by plugging in the ML estimate:

$$p(y_*|x_*, \boldsymbol{w}_{\mathrm{ML}}, \mathcal{M})$$

Bayesian inference, parametric model

Posterior parameter distribution by Bayes rule ($p(a|b)p(b) = p(a)p(b|a)$):

$$p(\mathbf{w}|\mathbf{x},\mathbf{y},\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathcal{M}) \;=\; p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M})$$

Making predictions (marginalizing out the parameters):

$$\begin{aligned} p(y_*|x_*,\mathbf{x},\mathbf{y},\mathcal{M}) \;&=\; \int p(y_*,\mathbf{w}|\mathbf{x},\mathbf{y},x_*,\mathcal{M})d\mathbf{w} \\ &=\; \int p(y_*|\mathbf{w},x_*,\mathcal{M})p(\mathbf{w}|\mathbf{x},\mathbf{y},\mathcal{M})d\mathbf{w}. \end{aligned}$$

Marginal likelihood:

$$p(\mathbf{y}|\mathbf{x},\mathcal{M}) \;=\; \int p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M})d\mathbf{w}.$$

Posterior and predictive distribution in detail

For a linear-in-the-parameters model with Gaussian priors and Gaussian noise:

- Gaussian *prior* on the weights: $p(\mathbf{w}|\mathcal{M}) = \mathcal{N}(\mathbf{w};\ 0,\ \sigma_{\mathbf{w}}^2\,\mathbf{I})$
- Gaussian *likelihood* of the weights: $p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M}) = \mathcal{N}(\mathbf{y};\ \mathbf{\Phi}\,\mathbf{w},\ \sigma_{\text{noise}}^2\,\mathbf{I})$

Posterior parameter distribution by Bayes rule $p(a|b) = p(a)p(b|a)/p(b)$:

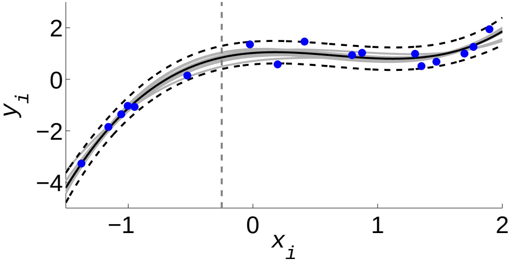
$$p(\mathbf{w}|\mathbf{x},\mathbf{y},\mathcal{M}) \;=\; \frac{p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M})}{p(\mathbf{y}|\mathbf{x},\mathcal{M})} \;=\; \mathcal{N}(\mathbf{w};\ \boldsymbol{\mu},\ \mathbf{\Sigma})$$

$$\mathbf{\Sigma} \;=\; \big(\sigma_{\text{noise}}^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma_{\mathbf{w}}^{-2}\,\mathbf{I}\big)^{-1} \quad \text{and} \quad \boldsymbol{\mu} \;=\; \Big(\mathbf{\Phi}^\top\mathbf{\Phi} + \frac{\sigma_{\text{noise}}^2}{\sigma_{\mathbf{w}}^2}\,\mathbf{I}\Big)^{-1}\mathbf{\Phi}^\top\mathbf{y}$$

The predictive distribution is given by:

$$\begin{aligned} p(y_*|x_*,\mathbf{x},\mathbf{y},\mathcal{M}) \;&=\; \int p(y_*|\mathbf{w},x_*,\mathcal{M})p(\mathbf{w}|\mathbf{x},\mathbf{y},\mathcal{M})d\mathbf{w} \\ &=\; \mathcal{N}(y_*;\ \boldsymbol{\phi}(x_*)^\top\boldsymbol{\mu},\ \boldsymbol{\phi}(x_*)^\top\mathbf{\Sigma}\boldsymbol{\phi}(x_*) + \sigma_{\text{noise}}^2). \end{aligned}$$

Multiple explanations of the data



Remember that a finite linear model $f(x_n) = \boldsymbol{\phi}(x_n)^\top \boldsymbol{w}$ with prior on the weights $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w};\ 0, \sigma_{\boldsymbol{w}}^2 \mathbf{I})$ has a posterior distribution

$$p(\boldsymbol{w}|\mathbf{x}, \mathbf{y}, \mathcal{M}) \ = \ \mathcal{N}(\boldsymbol{w};\ \boldsymbol{\mu}, \ \boldsymbol{\Sigma}) \quad \text{with} \quad \begin{aligned} \boldsymbol{\Sigma} \ &= \ \big(\sigma_{\text{noise}}^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \sigma_{\boldsymbol{w}}^{-2}\big)^{-1} \\ \boldsymbol{\mu} \ &= \ \Big(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \frac{\sigma_{\text{noise}}^2}{\sigma_{\boldsymbol{w}}^2}\,\mathrm{I}\Big)^{-1} \boldsymbol{\Phi}^\top \mathbf{y} \end{aligned}$$

and predictive distribution

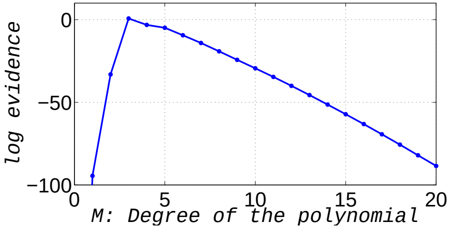
$$p(y_*|x_*, \mathbf{x}, \mathbf{y}, \mathcal{M}) \ = \ \mathcal{N}(y_*;\ \boldsymbol{\phi}(x_*)^\top \boldsymbol{\mu}, \ \boldsymbol{\phi}(x_*)^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(x_*) + \sigma_{\text{noise}}^2\,\mathbf{I})$$

Marginal likelihood (Evidence) of our polynomials

Marginal likelihood, or "evidence" of a finite linear model:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x},\mathcal{M}) &= \int p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M})\,\mathrm{dw} \\ &= \mathcal{N}(\mathbf{y};\ 0, \sigma_{\mathbf{w}}^2\, \mathbf{\Phi}\, \mathbf{\Phi}^\top + \sigma_{\text{noise}}^2\, \mathbf{I}). \end{aligned}$$

Luckily for Gaussian noise there is a closed-form analytical solution!



- The evidence prefers $M = 3$, not simpler, not more complex.
- Too simple models consistently miss most data.
- Too complex models frequently miss some data.