Bayesian inference and prediction in finite regression models

Carl Edward Rasmussen

October 13th, 2016
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: \( x, y \)
- model \( \mathcal{M}: y = f_w(x) + \varepsilon \)

Gaussian likelihood:

\[
p(y|x, w, \mathcal{M}) \propto \prod_{n=1}^{N} \exp\left(-\frac{1}{2}(y_n - f_w(x_n))^2/\sigma_{\text{noise}}^2\right).
\]

Maximize the likelihood:

\[
w_{\text{ML}} = \arg\max_w p(y|x, w, \mathcal{M}).
\]

Make predictions, by plugging in the ML estimate:

\[
p(y_*|x_*, w_{\text{ML}}, \mathcal{M})
\]
Bayesian inference, parametric model

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

\[
p(w|x, y, M) = \frac{p(w|M)p(y|x, w, M)}{p(y|x, M)}
\]

Making predictions (marginalizing out the parameters):

\[
p(y_*|x_*, x, y, M) = \int p(y_*, w|x, y, x_*, M)dw
\]

\[
= \int p(y_*|w, x_*, M)p(w|x, y, M)dw.
\]

Marginal likelihood:

\[
p(y|x, M) = \int p(w|x, M)p(y|x, w, M)dw
\]
Posterior and predictive distribution in detail

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

• Gaussian prior on the weights: \( p(w|M) = \mathcal{N}(w; 0, \sigma_w^2 I) \)

• Gaussian likelihood of the weights: \( p(y|x, w, M) = \mathcal{N}(y; \Phi w, \sigma_{\text{noise}}^2 I) \)

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

\[
p(w|x, y, M) = \frac{p(w|M)p(y|x, w, M)}{p(y|x, M)} = \mathcal{N}(w; \mu, \Sigma)
\]

\[
\Sigma = \left( \sigma_{\text{noise}}^{-2} \Phi^\top \Phi + \sigma_w^{-2} I \right)^{-1} \quad \text{and} \quad \mu = \left( \Phi^\top \Phi + \frac{\sigma_{\text{noise}}^2}{\sigma_w^2} I \right)^{-1} \Phi^\top y
\]

The predictive distribution is given by:

\[
p(y_*|x_*, x, y, M) = \int p(y_*|w, x_*, M)p(w|x, y, M)dw
\]

\[
= \mathcal{N}(y_*; \Phi(x_*)^\top \mu, \Phi(x_*)^\top \Sigma \Phi(x_*) + \sigma_{\text{noise}}^2).
\]
Remember that a finite linear model $f(x_n) = \Phi(x_n)^\top w$ with prior on the weights $p(w) = \mathcal{N}(w; 0, \sigma_w^2 I)$ has a posterior distribution

$$p(w|x, y, M) = \mathcal{N}(w; \mu, \Sigma) \quad \text{with} \quad \Sigma = \left( \sigma_{\text{noise}}^{-2} \Phi^\top \Phi + \sigma_w^{-2} \right)^{-1}$$

$$\mu = \left( \Phi^\top \Phi + \frac{\sigma_{\text{noise}}^2}{\sigma_w^2} I \right)^{-1} \Phi^\top y$$

and predictive distribution

$$p(y_*|x_*, x, y, M) = \mathcal{N}(y_*; \Phi(x_*)^\top \mu, \Phi(x_*)^\top \Sigma \Phi(x_*) + \sigma_{\text{noise}}^2 I)$$
Marginal likelihood (Evidence) of our polynomials

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

\[
p(y|x, M) = \int p(w|x, M)p(y|x, w, M)dw
\]

\[
= \mathcal{N}(y; 0, \sigma_w^2 \Phi \Phi^\top + \sigma_{\text{noise}}^2 I).
\]

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.