Marginal Likelihood

Carl Edward Rasmussen

July 1st, 2016

Key concepts

Marginal likelihood

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

Marginal likelihood:

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

Second level inference: model comparison and Bayes' rule again

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

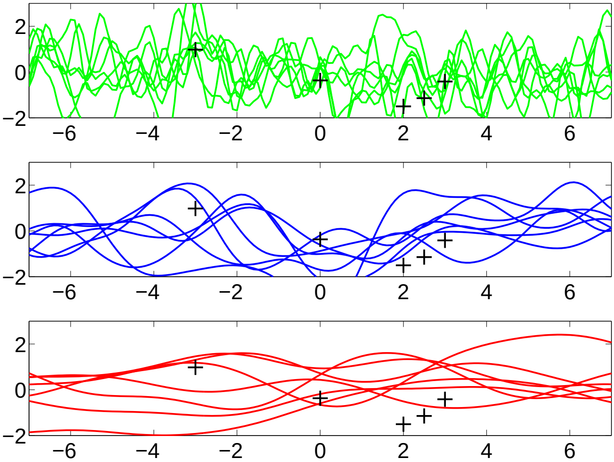
The *marginal likelihood* is used to select between models.
For linear in the parameter models with Gaussian priors and noise:

$$p(\mathbf{y}|\mathbf{x},\mathcal{M}) \;=\; \int p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M})d\mathbf{w} \;=\; \mathcal{N}(\mathbf{y};\;\mathbf{0},\sigma_{\mathbf{w}}^2\,\boldsymbol{\Phi}\,\boldsymbol{\Phi}^\top+\sigma_{\text{noise}}^2\,\mathbf{I})$$

Understanding the marginal likelihood (1). Models

Consider 3 models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. Given our data:

- We want to compute the *marginal likelihood* for each model.
- We want to obtain the predictive distribution for each model.



Understanding the marginal likelihood (2). Noise

Consider a very simple noise model for $y_n = f(x_n) + \epsilon_n$

- $\epsilon_n \sim \text{Uniform}(-0.2, 0.2)$ and all noise terms are independent.

$$p(y_n|f(x_n)) = \begin{cases} 0 & \text{if } |y_n - f(x_n)| > 0.2 \\ 1/0.4 = 2.5 & \text{otherwise} \end{cases}$$

- The likelihood of a given function from the prior is

$$p(\mathbf{y}|\mathbf{f}) = \prod_{n=1}^{N} p(y_n|f(x_n)) = \begin{cases} 0 & \text{if for any } n,\ |y_n - f(x_n)| > 0.2 \\ 2.5^N & \text{otherwise} \end{cases}$$

We will approximate the marginal likelihood by *Monte Carlo* sampling:

$$p(\mathbf{y}|\mathcal{M}_i) = \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|\mathcal{M}_i)\, d\,\mathbf{f} \approx \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}|\mathbf{f}_s) = \frac{S_a}{S} \cdot 2.5^N$$

- A total of S functions are sampled from the prior $p(\mathbf{f}|\mathcal{M}_i)$.
- $\mathbf{f}_s$ is the s^{th} function sampled from the prior.
- S_a is the number of samples with non-zero likelihood: these are accepted. The remaining $S - S_a$ samples are rejected.

Simple Monte Carlo

We can approximate integrals of the form

$$z = \int f(x)p(x)dx,$$

where $p(x)$ is a probability distribution, using a sum

$$z \simeq \frac{1}{T}\sum_{t=1}^{T} f(x^{(t)}), \quad \text{where} \quad x^{(t)} \sim p(x).$$

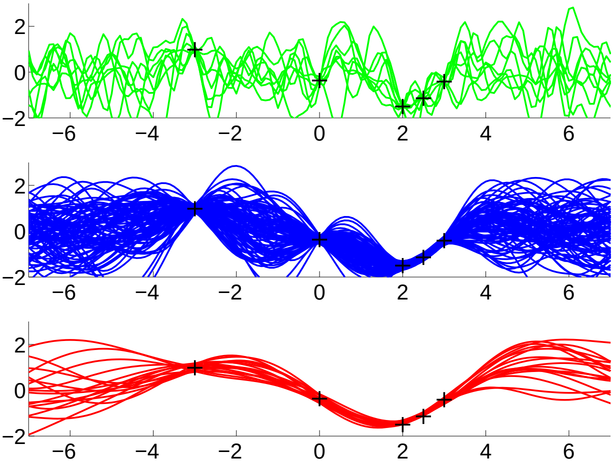
As $T \to \infty$ the approximation (under very mild conditions) converges to z.
This algorithm is called *Simple Monte Carlo*.

Understanding the marginal likelihood (3). Posterior

Posterior samples for each of the models obtained by rejection sampling.

- For each model we draw 1 million samples from the prior.
- We only keep the samples that have non-zero likelihood.

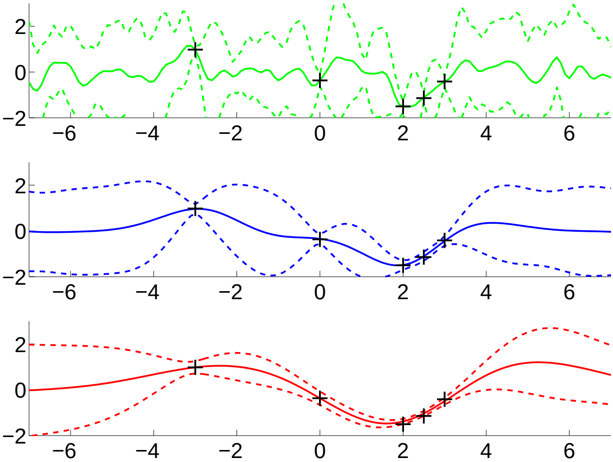
S_a	$p(\mathbf{y}\vert\mathcal{M}_i)$
8	8×10^{-4}
88	$\mathbf{9 \times 10^{-3}}$
17	2×10^{-3}



Predictive distribution

Predictive distribution for each of the models obtained.

- For each model we take all the posterior functions from rejection sampling.
- We compute the average and standard deviation of $f_s(x)$.



Conclusions

Probability theory provides a framework for

- making inferences from data in a model
- making probabilistic predictions

It also provides a *principled* and *automatic* way of doing

- model comparison

In the following lectures, we'll demonstrate how to use this framework to solve challenging machine learning problems.