Reading Notes for Mackey ch28 Model Comparison and Occam's Razor

Xiang Pan

September 21, 2021

1 Occam's Razor

Occam's razor: Accept the simplest explanation that fits the data. Reason:

- aesthetic ('A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data')
- past empirical success of Occam's razor

It is essential to use proper priors – otherwise the evidences and the Occam factors are not meaningful.

$$\frac{P(\mathcal{H}_1 \mid D)}{P(\mathcal{H}_2 \mid D)}=\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}\frac{P(D \mid \mathcal{H}_1)}{P(D \mid \mathcal{H}_2)}$$

Occam's Razor gives a favor of simplicity to the model.
Model fitting

$$\begin{aligned} P(\mathbf{w} \mid D, \mathcal{H}_i) &= \frac{P(D \mid \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} \mid \mathcal{H}_i)}{P(D \mid \mathcal{H}_i)} \\ \text{Posterior } &= \frac{\text{Likelihood } \times \text{ Prior}}{\text{Evidence}} \end{aligned} \tag{1}$$

Model Comparison
Models H_i are ranked by evaluating the evidence,

$$P(D \mid \mathcal{H}_i) \simeq \underbrace{P(D \mid \mathbf{w}_{\mathrm{MP}}, \mathcal{H}_i)} \times \underbrace{P(\mathbf{w}_{\mathrm{MP}} \mid \mathcal{H}_i) \sigma_{w \mid D}} . \tag{2}$$

$$\begin{aligned} \text{Evidence } &\simeq \text{ Best fit likelihood } \times \text{ Occam factor} \\ \text{Occam factor } &= \frac{\sigma_{w \mid D}}{\sigma_w} \end{aligned}$$

Occam factor is equal to the ratio of the posterior accessible volume of H_i's parameter space to the prior accessible volume.

Occam factor for several parameters If the posterior is well approximated by a Gaussian, then the Occam factor is obtained from the determinant of the corresponding covariance matrix.

$$P(D \mid \mathcal{H}_i) \simeq P(D \mid \mathbf{w}_{\mathrm{MP}}, H_i) \times P(\mathbf{w}_{\mathrm{MP}} \mid \mathcal{H}_i) \det^{-\frac{1}{2}}(\mathbf{A}/2\pi) \tag{3}$$

$$\mathbf{A} = -\nabla\nabla \ln P(\mathbf{w} \mid D, \mathcal{H}_i) \tag{4}$$

On-line learning and cross-validation.

$$\begin{aligned} \log P(D \mid \mathcal{H}) = &\log P\left(\mathbf{t}^{(1)} \mid \mathcal{H}\right) + \log P\left(\mathbf{t}^{(2)} \mid \mathbf{t}^{(1)}, \mathcal{H}\right) \\ &+ \log P\left(\mathbf{t}^{(3)} \mid \mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \mathcal{H}\right) + \cdots + \log P\left(\mathbf{t}^{(N)} \mid \mathbf{t}^{(1)} \ldots \mathbf{t}^{(N-1)}, \mathcal{H}\right) \end{aligned} \tag{5}$$

Cross-validation examines the average value of just the last term. The evidence, on the other hand, sums up how well the model predicted all the data, starting from scratch.