Information Theory, Inference, and Learning Algorithms Chapter 28 Summary

CSCI-GA 3033 Bayesian Machine Learning
New York University
Yilun Kuang

Sep 29, 2022

28. 1 Occam's Razor

Model Comparison and Occam's Razor

Occam's Razor is a model comparison principle that favors small complexities. Specifically, consider the Bayesian model selection problem

$$\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_2|D)} = \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)} \frac{p(D|\mathcal{H}_1)}{p(D|\mathcal{H}_2)},$$

where $p(\mathcal{H}_i)$ is the prior over the hypothesis $\mathcal{H}_i$ and the probability of the data D given model $\mathcal{H}_i$ $p(D|\mathcal{H}_i)$ is called *model evidence*. It turns out that *evidence* encodes the Occam's Razor principle automatically even if we assume equal priors (see examples in chapter 28). The following figure shows that a simple model $\mathcal{H}_1$ has supports on a limited intervals while a more complicated model $\mathcal{H}_2$ with more free parameters has heavy tails.

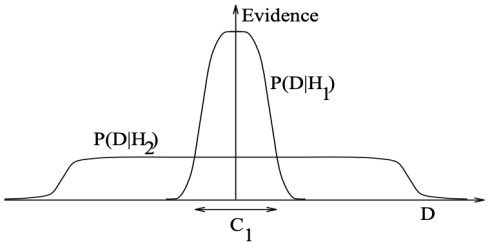

Figure 1: Model Evidence

Remark: If we select models based on a hold-out validation set with relatively small distribution shift compared to the training set, it's possible that a less complicated model will be selected. Such model might not be able to handle out-of-distribution datasets. It'll also be interesting to see how modern over-parametrized neural networks or other more powerful hypothesis classes fit into the Bayesian model selection framework.

The Mechanism of the Bayesian Razor: The Evidence and the Occam Factor

For frequentist, the maximum likelihood model choice prefer complex, over-parametrized models as it fits the data better but might generalizes poorly (take it with a grain of salt). Thus we need the Occam's Razor, which is embodied in the Bayesian inference as we'll show below.
There are two levels of inference in the Bayesian framework. The first level is the estimation of posterior measures on the parameters.

$$p(\mathbf{w}|D, \mathcal{H}_i) = \frac{p(D|\mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\mathcal{H}_i)}{p(D|\mathcal{H}_i)} \iff \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

It's common to use gradient-based methods to find the maximum posterior $\mathbf{w}_{\text{MP}}$. We can then summarize the posterior distribution by the value of $\mathbf{w}_{\text{MP}}$, and error bars or confidence intervals on these best-fit parameters.

Error bars can be obtained from the curvature of the posterior by evaluating the Hessian at $\mathbf{w}_{\text{MP}}$, $\mathbf{A} = -\nabla\nabla \ln p(\mathbf{w}|D, \mathcal{H}_i)|_{\mathbf{w}_{\text{MP}}}$. We can then Taylor expand the log posterior probability with $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\text{MP}}$.

$$p(\mathbf{w}|D, \mathcal{H}_i) \approx p(\mathbf{w}_{\text{MP}}|D, \mathcal{H}_i) \exp(-1/2\Delta\mathbf{w}^\top \mathbf{A}\Delta\mathbf{w})$$

i.e. the posterior can be locally approximated as a Gaussian with covariance matrix (equivalent to error bars) $\mathbf{A}^{-1}$.

The second level involves the inference process during model comparison.

$$p(\mathcal{H}_i|D) \propto p(D|\mathcal{H}_i)p(\mathcal{H}_i)$$

where

$$p(D|\mathcal{H}_i) = \int p(D|\mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\mathcal{H}_i)d\mathbf{w}$$

This integral can be approximated using the Laplace method by the peak of the integrand $p(D|\mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\mathcal{H}_i)$ times its width $\sigma_{\mathbf{w}|D}$:

$$\underbrace{p(D|\mathcal{H}_i)}_{\text{Evidence}} \approx \underbrace{p(D|\mathbf{w}_{\text{MP}}, \mathcal{H}_i)}_{\text{Best Fit Likelihood}} \times \underbrace{p(\mathbf{w}_{\text{MP}}|\mathcal{H}_i)\sigma_{\mathbf{w}|D}}_{\text{Occam Factor}}$$

Interpretation of the Occam Factor

Suppose $p(\mathbf{w}_{\text{MP}}|\mathcal{H}_i) = \frac{1}{\sigma_w}$, then the Occam Factor is given by

$$\text{Occam Factor} = \frac{\sigma_{w|D}}{\sigma_w}$$

which measures the factor by which the hypothesis space collapses when the data arrives. The Bayesian model comparison is an extension of the MLE model comparison as it's the likelihood itself multiplied by the Occam factor.

28. 2 Examples

Skipped

28. 3 Minimum Description Length (MDL)

The MDL framework can be viewed from a Bayesian perspective. One should prefer models that can communicate data in the smallest number of bits.

Remark: In fact, we can also connect this to the efficient / sparse coding ideas in neuroscience literatures (also Barlow-Twins). A criteria for learning representations is to have compact and efficient representations, i.e. the principle of parsimony and the maximal coding rate distortion idea. This is perhaps not as related to the model comparison itself but a reasonable desiderata in general.