Reading Notes for Bayesian Model Selection

Xiang Pan

September 22, 2021

1 Occam's Razor

Occam's Razor focus on complex models.

Non-Bayesian Netease Music Model complexity is often regulated by adjusting the number of free parameters in the model and sometimes complexity is further constrained by the use of regularizers (such as weight decay).

Depending on the scaling properties of the prior over parameters, both the Occam's Razor view and the large models view can seem appropriate.

Bayesian

- One view is to infer the probability of the model for each of several different model sizes and use these probabilities when making predictions.
- we simply choose a "large enough" model and sidestep the problem of model size selection.

View1 The evidence is the probability that if you randomly selected parameter values from your model class, you would generate data set Y .

View2 We don't seriously believe that the "true" generative process can be implemented exactly with a small model. Moreover, optimizing (or integrating) over continuous hyperparameters may be easier than optimizing over the discrete space of model sizes.

We ought not to limit the number of basis functions in function approximation a priori since we don't really believe that the data was actually generated from a small number of fixed basis functions.

scaling exponent γ: Large values of γ correspond to priors with most probability mass on simple functions, whereas small values of γ correspond to priors that allow more complex functions.

2 Bayesian model averaging is not model combination

- Model combination works by enriching the space of hypotheses, not by approximating a Bayesian model average.
- BMA is 'soft model selection'. The soft weights in BMA only reflect a statistical inability to distinguish the hypothesis based on limited data.

BMA as soft model selection

Bayesian model averaging is best thought of as a method for 'soft model selection.' It answers the question: "Given that all of the data so far was generated by exactly one of the hypotheses, what is the probability of observing the new pair (c, x)?"

BMA for stacked models

"Given that all of the data so far was generated by some linear combination of the hypotheses, what is the probability of observing the new pair (c, x)?"

This is BMA applied to a new hypotheses space of "stacked" models.

$$p((c,x) \mid D) \propto \sum_h p((c,x), D \mid h)p(h) \tag{1}$$

, which emphasizes the assumption that exactly one hypothesis is responsible for all of the data.

3 Bayesian Averaging of Classifiers and the Overfitting Problem[1]

Bayesian model averaging posterior probability is the product of the model's prior probability, which reflects our domain knowledge (or assumptions) before collecting data, likelihood, which is the probability of the data given the model, [1]

Given the "correct" model space and prior distribution, Bayesian model averaging is the optimal method for making predictions;

[1] You can get the updated and full version of note at https://github.com/Xiang-Pan/NYU_Bayesian_Machine_Learning/blob/master/reading_notes/reading6/build/note6.pdf

3.1 Bayesian Model Averaging in Classification

$$\Pr(h \mid \vec{x}, \vec{c}) = \frac{\Pr(h)}{\Pr(\vec{x}, \vec{c})} \prod_{i=1}^{n} \Pr\left(x_i, c_i \mid h\right) \tag{2}$$

$$\text{posterior probability of h given datasets} = \frac{\text{prior probability of h}}{\text{data prior}} \times \text{likelihood} \tag{3}$$

Each example's class is corrupted with probability ϵ,

$$\Pr(h \mid \vec{x}, \vec{c}) \propto \Pr(h)(1-\epsilon)^s \epsilon^{n-s}, \tag{4}$$

where s is the number of examples correctly classified by h.

An unseen example x is assigned to the class that maximizes:

$$\Pr(c \mid x, \vec{x}, \vec{c}, H) = \sum_{h \in H} \Pr(c \mid x, h) \Pr(h \mid \vec{x}, \vec{c}) \tag{5}$$

3.2 Bayesian Model Averaging of Bagged C4.5 Rule Sets

$$\sum f(x)p(x) = \sum f(x) \left[\frac{p(x)}{q(x)}\right] q(x) \tag{6}$$

p(x) being the model posterior probabilities, importance sampling distribution q(x).

Bagging($p(x) = q(x)$) can be viewed as an approximation of Bayesian model averaging by importance sampling.

More closely approximate: weighting models by their posteriors leads to a better approximation of Bayesian model averaging than weighting them uniformly

3.3 The Overfitting Problem in Bayesian Model Averaging

Bayesian model averaging in effect acts more like model selection than model averaging, and is equivalent to amplifying the search process that produces overfitting in the underlying learner in the first place.

Overfitting is the result of attempting too many hypotheses, and consequently finding a poor hypothesis that appears good. In this light, Bayesian model averaging's potential to aggravate the overfitting problem relative to learning a single model becomes clear.

Bagging is at one extreme, with uniform weights, and never increases Overfitting. Methods like boosting and stacking produce weights that are more variable, and can sometimes lead to overfitting.Bayesian model averaging is at the opposite end of the spectrum, producing highly asymmetric weights, and being correspondingly more prone to overfitting.

4 Bayesian PCA[2]

4.1 Probabilistic PCA

$$\begin{aligned} \mathbf{x} &= \sum_{j=1}^{k} \mathbf{h}_j w_j + \mathbf{m} + \mathbf{e} \\ &= \mathbf{H}\mathbf{w} + \mathbf{m} + \mathbf{e} \\ p(\mathbf{e}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \end{aligned} \tag{7}$$

k < d(dimension of x).

The goal of PCA is to estimate the basis vectors H and the noise variance v from a data set D.

References

[1] Pedro Domingos. Bayesian Averaging of Classifiers and the Overfitting Problem. page 8.

[2] Thomas P Minka. Automatic choice of dimensionality for PCA. page 16.