HYPERPARAMETERS: OPTIMIZE, OR INTEGRATE OUT?

Semkin Kirill

BBM
2024

Problem statement

Fully bayessian framework:

$$P(D, \mathrm{w}, \alpha, \beta|\mathcal{H}) = P(D|\mathrm{w}, \beta, \mathcal{H})P(\mathrm{w}|\alpha, \mathcal{H})P(\alpha, \beta|\mathcal{H}) \quad (1)$$

Applied to a linear model:

$$P(D|\mathrm{w}, \beta, \mathcal{H}) = \frac{1}{Z_D(\beta)} \exp(-\frac{1}{2}\beta E_D(\mathrm{w}))$$

$$P(\mathrm{w}|\alpha, \mathcal{H}) = \frac{1}{Z_W(\alpha)} \exp(-\frac{1}{2}\alpha \mathrm{w}^{\mathsf{T}}\mathrm{w}))$$

Problems to consider:

1. Infer the parameters: $P(\mathrm{w}|D, \mathcal{H})$ or marginal $P(w_i|D, \mathcal{H})$
2. Infer evidence $P(D|\mathcal{H})$
3. Prediction for new data $P(D_2|D, \mathcal{H})$

Ideal approach

Find closed forms of $P(\mathrm{w}|D,\mathcal{H})$, $P(D|\mathcal{H})$, $P(D_2|D,\mathcal{H})$ by integrating out unnecessary"variables from the model's probability (1).

$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) =$$

$$\sum_{i=1}^{K} p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}},\ M_i)p(M_i|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) =$$

$$\sum_{i=1}^{K} p_i(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})p(M_i|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})$$

$$p(M_i|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \frac{p(\mathbf{y}_{\text{train}},\ M_i|\mathbf{X}_{\text{train}})}{P(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})} \propto p(\mathbf{y}_{\text{train}}, M_i|\mathbf{X}_{\text{train}}) =$$

$$\int p(\mathbf{y}_{\text{train}},\ \mathbf{w},\ M_i|\mathbf{X}_{\text{train}})d\mathbf{w} = p(M_i)p_i(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})$$

Byassian inference for new data

Evidence Framework

- Having posterior for each α:

$$P(\mathrm{w}|D,\alpha,\mathcal{H}) = \frac{P(D|\mathrm{w},\alpha,\mathcal{H})P(\mathrm{w}|\alpha,\mathcal{H})}{P(D|\alpha,\mathcal{H})}$$

  Approximate it as gaussian (2nd-order approximation). Then evidence $P(D|\alpha,\mathcal{H})$ is evaluated from determinant of this distribution.
- Maximize evidence and estimate α_{MP} with its deviation $\sigma_{\alpha|D}$
- Put α_{MP} into posterior and obtain approximate solution for Problem 1. The result is gaussian on w with mean $\mathrm{w}_{\mathrm{MAP}|\alpha_{\mathrm{MP}}}$
- Estimate evidence (Problem 2):

$$P(D|\mathcal{H}) = P(D|\alpha_{\mathrm{MP}},\mathcal{H})P(\alpha_{\mathrm{MP}}|\mathcal{H})/\sqrt{2\pi}\sigma_{\alpha|D}$$

- Estimate probability of new data using MAP-posterior:

$$P(D_2|D,\mathcal{H}) = \int P(D_2|\mathrm{w},\mathcal{H})P(\mathrm{w}|D,\alpha_{\mathrm{MP}},\mathcal{H})d\mathrm{w}$$

Implicit estimation of α_{MP}

We can determine α_{MP} using intuitive implicit equation

$$\frac{1}{\alpha_{\text{MP}}} = \frac{\sum_{i=1}^{k} w_i^2}{\gamma} \tag{2}$$

Where γ is the number of *well-determined parameters*. It is charactersitcs of the input data

$$\gamma = k - \alpha \text{tr}(\Sigma) = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i + \alpha}$$

Therefore we estimate α only from well-determined parameters.

Why does $P(\alpha|D,\mathcal{H})$ vanish?

Our approach relies on yet another approximation:

$$P(D_2|D,\mathcal{H}) = \int P(D_2|D,\alpha,\mathcal{H})P(\alpha|D,\mathcal{H})d\alpha \approx P(D_2|D,\alpha_{\mathsf{MP}},\mathcal{H})$$

This is true if $P(\alpha|D,\mathcal{H})$ incentive to changes in α on scale $\sigma_{\alpha|D}$. Such fact is true, for instance, in thermodynamics assembles with high freedom degrees. Having two different distributions

$$P(r) = \frac{\beta E_r}{Z}$$

$$P(r) = \begin{cases} \frac{1}{\Omega}, & E_r \in U_\varepsilon(\bar{E}_r) \\ 0, & \text{else} \end{cases}$$

They are indistinguishable in terms of probability mass and marginals

MAP Method

Find "true"prior first

$$P(\mathrm{w}|\mathcal{H}) = \int P(\mathrm{w}|\alpha, \mathcal{H}) P(\alpha|\mathcal{H}) d\alpha$$

Then write "true"posterior and maximize it

$$P(\mathrm{w}|D, \mathcal{H}) \propto P(D|\mathrm{w}, \mathcal{H}) P(\mathrm{w}|\mathcal{H})$$

To solve stated problems:

- Approximate $P(\mathrm{w}|D, \mathcal{H})$ with a gaussian near w_{MP} (Problem 1)
- Estimate evidence $P(\mathcal{H}|D)$ using determinant of the approximation (Problem 2)
- Estimate probability of new data with $P(D|\mathrm{w}_{\mathrm{MP}}, \mathcal{H})$ (Problem 3)

Similarity between Evidence and MAP

Take improper prior $p(\log \alpha) = const$, then

$$p(\mathrm{w}|\mathcal{H}) = \int_0^{+\infty} \frac{\sum_{i=1}^{k} w_i^2 e^{-\alpha}}{Z_W(\alpha)} d\log\alpha \propto \frac{1}{(\sum_{i=1}^{k} w_i^2)^{k/2}}$$

We can define $\alpha_{\mathrm{eff}}(w)$ as such α that the maximum of $p(\mathrm{w}|\mathcal{H})$ and $p(\mathrm{w}|\mathcal{H}, \alpha)$ are the same. Then

$$\frac{1}{\alpha_{\mathrm{eff}}(w)} = \frac{\sum_{i=1}^{k} w_i^2}{k} \tag{3}$$

Choosing the framework

MAP method has several advances over Evidence:

1. It uses only one approximation in true posterior whilst Evidence approximates the whole integral over α
2. MAP method is easily transferred across different data as it requires only one computation of true prior. Evidence needs recalculation with every new dataset.

Nonetheless if the data D contains ill-determined parameters, the MAP estimator will become greatly biased while the Evidence estimator will not change. It can be seen from the (3) and (2). High level reason is that the MAP estimation is not viable if the posterior distribution differs from $\delta(\mathrm{w} - \mathrm{w}_{\mathrm{MP}})$. The situation worsens in high dimensional hyperparameter spaces.