[Paper review 5]

Ensemble Learning in Bayesian Neural Networks (David Barber and Christopher M. Bishop, 1998)

[Contents]

0. Abstract

Bayesian treatments for NN : three main approaches

- 1) Gaussian Approximation
- 2) MCMC
- 3) Ensemble Learning

Ensemble Learning

- aims to approximate posterior by minimizing KL-divergence between "true posterior" & "parametric approximating distribution"
- original : use of Gaussian approximating distribution with "diagonal" covariance matrix

  this paper : extended to "full-covariance" Gaussian distribution (while remaining computationally tractable)

1. Introduction

posterior distribution : $P(w \mid D) \rightarrow$ corresponding integrals over weight space are "analytically intractable"

1) Gaussian Approximation

- known as Laplace's method
- centered at a mode of $p(w \mid D)$
- covariance of the Gaussian is determined by the local curvature of the posterior

2) MCMC

- more recent method
- generate samples from the posterior
- but, computationally expensive

3) Ensemble Learning

- introduced by Hinton and van Camp (1993)
- finding a simple, analytically tractable approximation to true posterior
- (unlike 1) Laplace's method) approximating distribution is fitted globally (not locally)
- by minimizing KL-divergence
- Hinton and van Camp(1993) : with a diagonal covariance

  (but restriction to diagonal covariance prevents the model from capturing the posterior correlation between the parameters)
- Barber and Bishop (1998) : can be extended to allow a Gaussian approximating distribution with a general covariance matrix, while still leading to a tractable algorithm

2. BNN (Bayesian Neural Networks)

example)

- 2 layer , H hidden units, 1 output unit
- $D=\{\mathrm{x}^\mu, t^\mu\}, \mu=1, \ldots, N$

network : $f(\mathbf{x}, \mathbf{w})=\sum_{i=1}^H v_i \sigma\left(\mathbf{u}_i^{\mathrm{T}} \mathbf{x}\right)$, where $\mathbf{w} \equiv\left\{\mathbf{u}_i, v_i\right\}$

activation function : 'erf' function (=cumulative Gaussian) : $\sigma(a)=\sqrt{\frac{2}{\pi}} \int_0^a \exp \left(-s^2 / 2\right) d s$

standard assumption of Gaussian noise on the target(output) values, with precision β (= variance β^{-1})

1) likelihood : $P(D \mid \mathbf{w}, \beta)=\frac{\exp \left(-\beta E_D\right)}{Z_D}$

- normalizing factor : $Z_D=(2 \pi / \beta)^{N / 2}$
- training error : $E_D(\mathbf{w})=\frac{1}{2} \sum_\mu\left(f\left(\mathbf{x}^\mu, \mathbf{w}\right)-t^\mu\right)^2$

2) prior : $P(\mathbf{w} \mid \mathbf{A})=\frac{\exp \left(-E_W(\mathbf{w})\right)}{Z_P}$

- Gaussian
- normalizing factor : $Z_P=(2 \pi)^{k / 2}|A|^{-1 / 2}$
- matrix of hyperparameters : $E_W(\mathbf{w})=\frac{1}{2} \mathbf{w}^{\mathrm{T}} \mathbf{A} \mathbf{w}, \mathbf{A}$

3) posterior : $P(\mathbf{w} \mid D, \beta, \mathbf{A})=\frac{1}{Z_F} \exp \left(-\beta E_D(\mathbf{w})-E_W(\mathbf{w})\right)$

- normalizing factor : $Z_F=\int \exp \left(-\beta E_D(\mathbf{w})-E_W(\mathbf{w})\right) d \mathbf{w}$
- hard to calculate Z_F (integration above is intractable!)

Predictions for a new input (for given β and A) :

- by integration over the posterior
- predictive mean : $\langle f(\mathrm{x})\rangle = \int f(\mathrm{x}, \mathrm{w}) P(\mathrm{w} \mid D, \beta, \mathrm{A}) d\mathrm{w}$

  (= integration over a high-dimensional space ...... accurate evaluation is really hard!)

3. Laplace's Method

posterior approaches a Gaussian (whose variance goes to zero as $N \rightarrow \infty$)

To calculate Gaussian approximation...

- posterior : $P(\mathrm{w} \mid D, \beta, \mathbf{A}) = \exp(-\phi(\mathbf{w}))$
- expand ϕ around a mode of the distribution ($\mathbf{w}_{*} = \arg\min \phi(\mathbf{w})$)

  $\phi(\mathbf{w}) \approx \phi\left(\mathbf{w}_{*}\right) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}_{*}\right)^{\mathrm{T}} \mathbf{H}\left(\mathbf{w} - \mathbf{w}_{*}\right)$ (by Taylor Expansion)

  (where $\mathbf{H} = \nabla\nabla\phi(\mathbf{w})|_{\mathbf{w}}$ is a Hessian Matrix)
- $P(\mathrm{w} \mid D, \beta, \mathbf{A}) \simeq \frac{|\mathbf{H}|^{1/2}}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{w} - \mathbf{w}_{*}\right)^{\mathrm{T}} \mathbf{H}\left(\mathbf{w} - \mathbf{w}_{*}\right)\right\}$

The expected value of $\langle f(\mathrm{x})\rangle = \int f(\mathrm{x}, \mathrm{w}) P(\mathrm{w} \mid D, \beta, \mathrm{A}) d\mathrm{w}$ can be evaluated by making a further local linearization of $f(x, w)$

4. MCMC

- replace integrals to finite sums
- one of the most successful approaches : "hybrid Monte Carlo"
- $\int P(\mathbf{w} \mid D, \beta, \mathbf{A}) g(\mathbf{w}) d\mathbf{w} \approx \frac{1}{m} \sum_{i=1}^{m} g\left(\mathbf{w}_{i}\right)$

5. Ensemble Learning

- introduce a distribution $Q(w)$

$$
\begin{aligned}
\ln P(D \mid \beta, \mathbf{A}) &= \ln \int P(D \mid \mathbf{w}, \beta) P(\mathbf{w} \mid \mathbf{A}) d\mathbf{w} \\
&= \ln \int \frac{P(D \mid \mathbf{w}, \beta) P(\mathbf{w} \mid \mathbf{A})}{Q(\mathbf{w})} Q(\mathbf{w}) d\mathbf{w} \\
&\geq \int \ln\left\{\frac{P(D \mid \mathbf{w}, \beta) P(\mathbf{w} \mid \mathbf{A})}{Q(\mathbf{w})}\right\} Q(\mathbf{w}) d\mathbf{w} \\
&= \mathcal{F}[Q]
\end{aligned}
$$

ELBO = Variational Free Energy = $\mathcal{F}[Q]$

difference between $\ln P(D \mid \beta, \mathbf{A})$ and $\mathcal{F}[Q]$: $\mathrm{KL}(Q \| P) = \int Q(\mathbf{w}) \ln\left\{\frac{Q(\mathbf{w})}{P(\mathbf{w} \mid D, \beta, \mathbf{A})}\right\} d\mathbf{w}$

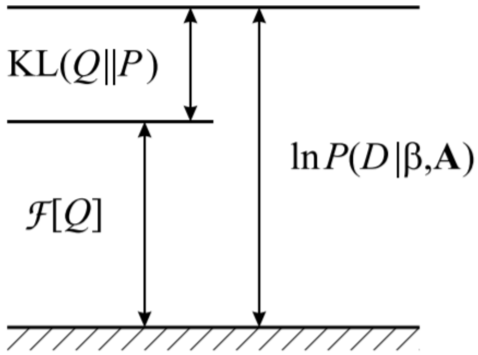

Figure 1: The quantity $\mathcal{F}[Q]$ provides a rigorous lower bound on the log marginal likelihood $\ln P(D|\beta, \mathbf{A})$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between the approximating distribution $Q(\mathbf{w})$ and the true posterior $P(\mathbf{w}|D, \beta, \mathbf{A})$.

Goal : to choose a form for $Q(w)$ so that $\mathcal{F}[Q]$ can be evaluated efficiently

- Maximizing $\mathcal{F}[Q]$
- Minimizing $\mathrm{KL}(Q\|P) = \int Q(\mathbf{w}) \ln \left\{ \frac{Q(\mathbf{w})}{P(\mathbf{w}|D,\beta,\mathbf{A})} \right\} d\mathbf{w}$
- The richer family of Q dist'n considered, the better the resulting bound

Key to a successful application of variational methods therefore lies in "the choice of the Q distribution"

- should be close to true posterior
- analytically tractable integration

Relationship between Variational Framework & EM Algorithm

- Standard EM algorithm :
  - E step : posterior distribution of hidden variables is used to evaluate the expectation of the complete-data log likelihood
  - M step : expected complete-data log likelihood is maximized w.r.t the parameters
- In Variational Framework :
  - E step : alternate maximization of $\mathcal{F}$ with respect to a free-form Q distribution
  - M step : alternate maximization of $\mathcal{F}$ with respect to hyper-parameters

5-1. Gaussian Variational Distribution

- Hinton and van Camp (1993) : diagonal covariance matrix
- Mackay (1995) : general class of Gaussian approximating distributions can be considered by "allowing the linear transformation of input variables"

  (even with this generalization, incapable of capturing strong correlations between parameters)
- This paper : such restrictions are unnecessary!

Consider a Q given by a Gaussian, with mean $\overline{\mathbf{w}}$ and covariance $\mathbf{C}$

Variational Free Energy (= ELBO) :

$$\mathcal{F}[Q] = -\int Q(\mathbf{w}) \ln Q(\mathbf{w}) d\mathbf{w} - \int Q(\mathbf{w}) \left\{ E_W + E_D \right\} d\mathbf{w} - \ln Z_P - \ln Z_D$$

- first term : entropy of Gaussian distribution

  $-\int Q(\mathbf{w}) \ln Q(\mathbf{w}) d\mathbf{w} = \frac{1}{2} \ln |\mathbf{C}| + \frac{k}{2}(1 + \ln 2\pi)$
- second term : prior term

  $\int Q(\mathbf{w}) E_W(\mathbf{w}) d\mathbf{w} = \mathrm{Tr}(\mathbf{CA}) + \frac{1}{2} \overline{\mathbf{w}}^{\mathrm{T}} \mathbf{A} \overline{\mathbf{w}}$
- third term : data dependent term

  $\int Q(\mathbf{w}) E_D(\mathbf{w}) d\mathbf{w} = \frac{1}{2} \sum_{\mu=1}^{N} l\left(\mathbf{x}^{\mu}, t^{\mu}\right)$

  (where $l(\mathbf{x}, t) = \int Q(\mathbf{w}) f(\mathbf{x}, \mathbf{w})^2 d\mathbf{w} - 2t \int Q(\mathbf{w}) f(\mathbf{x}, \mathbf{w}) d\mathbf{w} + t^2$)