[Paper review 20]

Uncertainty in Deep Learning - Chapter 2

(Yarin Gal, 2016)

[Contents]

2. The Language of Uncertainty

2.1 Bayesian Modeling

in Bayesian (Parametric) Model

- would like to find ω of $\mathrm{y} = \mathrm{f}^{\omega}(\mathrm{x})$
- that are "likely to have generated" our output

Likelihood

- classification : $p(y = d \mid \mathbf{x}, \boldsymbol{\omega}) = \frac{\exp(f_d^{\omega}(\mathbf{x}))}{\sum_{d'} \exp(f_{d'}^{\omega}(\mathbf{x}))}$ (softmax)
- regression : $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\omega}) = \mathcal{N}\left(\mathbf{y}; \mathbf{f}^{\omega}(\mathbf{x}), \tau^{-1} I\right)$ (Gaussian likelihood)

  (model precision τ : corrupt the model output with observation noise with variance τ^{-1})

Structure

- posterior distribution : $p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})$

  $p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega})p(\boldsymbol{\omega})}{p(\mathbf{Y}|\mathbf{X})}$
- predictive distribution : $p\left(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{Y}\right)$

  $p\left(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{Y}\right) = \int p\left(\mathbf{y}^* \mid \mathbf{x}^*, \boldsymbol{\omega}\right) p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y}) \mathrm{d}\boldsymbol{\omega}$ (known as "inference")
- model evidence : $p(\mathbf{Y} \mid \mathbf{X})$

  $p(\mathbf{Y} \mid \mathbf{X}) = \int p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega}) p(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}$

If conjugate prior, but if not , difficlut

(to make more interesting model, marginalization can not be done analytically ... APPROXIMATION is needed!)

2.1.1 Variational Inference

minimizing KL-divergence = maximizing ELBO

- $\mathrm{KL}\left(q_\theta(\omega) \| p(\omega \mid \mathrm{X}, \mathrm{Y})\right) = \int q_\theta(\omega) \log \frac{q_\theta(\omega)}{p(\omega \mid \mathrm{X}, \mathrm{Y})} \mathrm{d}\omega$

It allows us to approximate our "predictive distribution" as

-

ELBO : $\mathcal{L}_{\mathrm{VI}}(\theta) := \int q_\theta(\omega) \log p(\mathrm{Y} \mid \mathrm{X}, \omega) \mathrm{d}\omega - \mathrm{KL}\left(q_\theta(\omega) \| p(\omega)\right) \leq \log p(\mathrm{Y} \mid \mathrm{X}) = \log \text{ evidence}$

- (1) first term : encourage $q_\theta(\boldsymbol{\omega})$ to explain the data well
- (2) second term : encourages $q_\theta(\boldsymbol{\omega})$ to be close to the prior

Variational inference replaces "marginalization" $\rightarrow$ "OPTIMIZATION"

(replace calculation of "integral" $\rightarrow$ "derivatives" ... much easier and makes many approximations tractable)

We OPTIMIZE over "distributions" instead of point estimates

But.....

- does not scale to large data
- does not adapt to complex models

2.2 Bayesian Neural Networks

prior distribution : $P(W_i) = N(0, I)$

2.2.1 Brief history (PREVIOUS)

1) placing a prior distribution over the space of weight (Denker et al, 1987)

2) network generalization error (Tishby et al, 1989)

3) Only statistical interpretation of a NN euclidean loss is as maximum likelihood w.r.t a Gaussian likelihood over the network outputs

4) Laplace approximation (Denker and LeCun, 1991)

(optimized the NN weights to find a mode & fitted a Gaussian to that mode)

5) use model evidence for model compariosn (MacKay, 1992)

(model evidence correlates to the generalization error, thus can be used to select model size)

6) showed that model misspecification can lead to Bayes Failure

7) MDL (Hinton and Van Camp, 1993)

(first variational inference approximation to BNN)

8) HMC (Neal, 1995)

9) Different prior distribution in BNN (Neal, 1995)

(+ showed that in the limit of the number of units, the model would converge to various stable processes (depending on the prior used))

10) replaced Hinton and Van Camp (1993) 's diagonal matrices with full covariance matrices (Baraber and Bishop, 1998)

(+ gamma prior over the network hyper-parameters)

(+ VI with free-form variational distributions over the hyper-parameters)

2.2.2 Modern Approximate Inference (NEW)

can be divided into

- 1) variational inference
- 2) sampling based techniques
- 3) ensemble methods

1) variational inference

(1) Hinton and Van Camp (1993)

- VI perspective
  - intractable posterior $\rightarrow$ aprroximate with $q_\theta(w)$
  - minimize KL divergence

$$\begin{aligned}\mathrm{KL}\left(q_\theta(\boldsymbol{\omega}) \| p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})\right) & \propto-\int q_\theta(\boldsymbol{\omega}) \log p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega}) \mathrm{d} \boldsymbol{\omega}+\mathrm{KL}\left(q_\theta(\boldsymbol{\omega}) \| p(\boldsymbol{\omega})\right) \\ & =-\sum_{i=1}^N \int q_\theta(\boldsymbol{\omega}) \log p\left(\mathbf{y}_i \mid \mathbf{f}^\omega\left(\mathbf{x}_i\right)\right) \mathrm{d} \boldsymbol{\omega}+\mathrm{KL}\left(q_\theta(\boldsymbol{\omega}) \| p(\boldsymbol{\omega})\right)\end{aligned}$$

- fully factorized approximation

  (define $q_\theta(w)$ to factorize over the weights)

  $q_\theta(\boldsymbol{\omega})=\prod_{i=1}^L q_\theta\left(\mathbf{W}_i\right)=\prod_{i=1}^L \prod_{j=1}^{K_i} \prod_{k=1}^{K_{i+1}} q_{m_{ijk}, \sigma_{ijk}}\left(w_{ijk}\right)=\prod_{i, j, k} \mathcal{N}\left(w_{ijk} ; m_{ijk}, \sigma_{ijk}^2\right)$
- but, expected log likelihood $\int q_\theta(\boldsymbol{\omega}) \log p\left(\mathbf{y}_i \mid \mathbf{f}^\omega\left(\mathbf{x}_i\right)\right) \mathrm{d} \boldsymbol{\omega}$ is intractable

  thus, only used "single hidden layer"
- work bad in practice

  (losing important information about weight correlations)

(2) Barber and Bishop (1998)

- modeling correlation between the weights
- required covariance matrices $\rightarrow$ computational complexity $\rightarrow$ impractical

(3) Graves (2011)

- Data sub-sampling techniques (MC estimates / Mini-batch)

  $\rightarrow$ allows to scale to large amounts of data & complex models (for the first time, PRACTICAL!)
- but still performed bad in practice

  (due to the lack of correlations over the weight)

(4) Blundell et al (2015)

- re-parametrise the "expected log likelihood" MC estimates
- put a mixture of Gaussian prior & optimize the mixture components

  $\rightarrow$ improved model performance
- but computationally expensive

  (Gaussian approximating distributions increases the number of parameters)

(5) After...

- Probabilistic Back Propagation
- α- divergence minimization

2) sampling based techniques

(1) Neal (1995)

- HMC (Hamiltonian Dynamics)
- difficult in practice

  (setting the leapfrog step size & do not scale to large data)

(2) Langevin method

- simplification of Hamiltoninan dynamics

  (only a single leapfrog step)
- simplifies the inference $\rightarrow$ scale to large data

(3) Welling and Teh (2011)

- SGLD (Stochastic Gradient Langevin Dynamics)
- generate a set of samples $\{\widehat{\omega}_t\}$from posterior, by adding stochastic gradient steps to the previous samples

$$\Delta\omega = \frac{\epsilon}{2}\left(\nabla \log p(\omega) + \frac{N}{M}\sum_{i\in S}\nabla \log p\left(\mathbf{y}_i \mid \mathbf{x}_i, \omega\right)\right) + \eta$$
$$\eta \sim \mathcal{N}(0, \epsilon)$$

- unlike fully-factorized VI, can capture "weight correlation"
- difficulty : collapse to a single mode (do not explore well)

  (since ϵ decreases so rapidly ! probability of jumping out is too small)
- also, in practice, generate correlated samples $\rightarrow$ need to sample a lot

3) ensemble methods

- produces point estimate! (not a distribution)
- evaluating the sample variance of the outputs from all deterministic models
- more computationally efficient
- BUT, uncertainty estimate lacks in many ways

  (ex) RBF network : test data far from train data, will output zero(0)
  then sample variance of this technique will be zero at the given test point)
- can be alleviated by using "probabilistic models"

  (GP predictions far from the training data will have large model uncertainty)

2.2.3 Challenges

Important properties

- 1) scale well to large data & complex model
- 2) do not change the existing model architectures
- 3) easy for non-experts to use

Next chapter, will deal with approximate inference technique that meet those three!