[Paper review 4]

Practical Variational Inference for Neural Networks (Alex Graves, 2011)

[Contents]

0. Abstract

Variational methods : tractable approximation to Bayesian Inference

previous works : have only been applicable to few simple network architectures

This paper introduces "stochastic variational method" (= MDL loss function) that can be applied to most NN!

1. Introduction

at first, V.I. has not been widely used (due to difficulty of deriving analytical solutions to the integrals)

Key point :

- forget about analytical solutions! can be efficiently approximated with NUMERICAL INTEGRATION
- "Stochastic method" for V.I with a diagonal Gaussian posterior

takes a view of MDL (Minimum Description Length)

- 1) clear separation between "prediction accuracy" and "model accuracy"
- 2) recasting inference as "optimization" makes it easier to implement in "gradient-descent based NN"

2. Neural Networks

network loss (defined as the "negative log probability") :

$L^N(\mathbf{w}, \mathcal{D}) = -\ln \Pr(\mathcal{D} \mid \mathbf{w}) = -\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \ln \Pr(\mathbf{y} \mid \mathbf{x}, \mathbf{w})$

3. Variational Inference

prior of weights: $P(\mathbf{w} \mid \boldsymbol{\alpha})$

posterior of weights : $\Pr(\mathrm{w} \mid \mathcal{D}, \alpha) \rightarrow$ can not be calculated analytically in most cases

solve this problem by approximating $\Pr(\mathrm{w} \mid \mathcal{D}, \alpha)$ with a more tractable distribution $Q(\mathrm{w} \mid \beta)$

by minimizing "VARIATIONAL FREE ENERGY" : $\mathcal{F} = -\left\langle \ln\left[\frac{\Pr(\mathcal{D}|\mathbf{w})P(\mathbf{w}|\alpha)}{Q(\mathbf{w}|\beta)}\right]\right\rangle_{\mathbf{w}\sim Q(\boldsymbol{\beta})}$

($\langle g \rangle_{x\sim p}$ denotes the expectation of g over p)

4. MDL (Minimum Description Length)

Variational Free Energy $\mathcal{F}$ can be viewed with MDL principle!

$\mathcal{F} = \left\langle L^N(\mathbf{w}, \mathcal{D})\right\rangle_{\mathbf{w}\sim Q(\boldsymbol{\beta})} + D_{KL}(Q(\boldsymbol{\beta}) \| P(\boldsymbol{\alpha}))$

- 1) error loss : $L^E(\boldsymbol{\beta}, \mathcal{D}) = \left\langle L^N(\mathbf{w}, \mathcal{D})\right\rangle_{\mathbf{w}\sim Q(\boldsymbol{\beta})}$
- 2) complexity loss : $L^C(\boldsymbol{\alpha}, \boldsymbol{\beta}) = D_{KL}(Q(\boldsymbol{\beta}) \| P(\boldsymbol{\alpha}))$

with MDL view : $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{D}) = L^E(\boldsymbol{\beta}, \mathcal{D}) + L^C(\boldsymbol{\alpha}, \boldsymbol{\beta})$

- 1) cost of transmitting the model with w unspecified
- 2) cost of transmitting the prior

Network is then trained on $\mathcal{D}$ by minimizing $L(\alpha, \beta, \mathcal{D})$

5. Choice of Distributions

Should derive the form of $L^E(\boldsymbol{\beta}, \mathcal{D})$ and $L^C(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for various choices of $Q(\beta)$ and $P(\alpha)$

will limit to diagonal posteriors of the form

- $Q(\boldsymbol{\beta}) = \prod_{i=1}^{W} q_i(\beta_i)$
- $L^C(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{W} D_{KL}(q_i(\beta_i) \| P(\boldsymbol{\alpha}))$

5-1. Delta Posterior

- Delta posterior : simplest non-trivial distribution for $Q(\beta)$

  (assign probability 1 to a particular set of weights w , and 0 to all other weights)

- Prior : Laplace distribution with $\mu = 0 \rightarrow L1$ regularization
  - $\alpha = \{\mu, b\}$

- $P(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{W} \frac{1}{2b} \exp\left(-\frac{|w_i - \mu|}{b}\right)$
- $L^C(\boldsymbol{\alpha}, \mathbf{w}) = W \ln 2b + \frac{1}{b} \sum_{i=1}^{W} |w_i - \mu| + C$
- $\frac{\partial L^C(\boldsymbol{\alpha}, \mathbf{w})}{\partial w_i} = \frac{\operatorname{sgn}(w_i - \mu)}{b}$

- Prior : Gaussian distribution with $\mu = 0 \rightarrow L2$ regularization
  - $\alpha = \left\{\mu, \sigma^2\right\}$
  - $P(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w_i - \mu)^2}{2\sigma^2}\right)$
  - $L^C(\boldsymbol{\alpha}, \mathbf{w}) = W \ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2\sigma^2} \sum_{i=1}^{W} (w_i - \mu)^2 + C$
  - $\frac{\partial L^C(\boldsymbol{\alpha}, \mathbf{w})}{\partial w_i} = \frac{w_i - \mu}{\sigma^2}$

5-2. Gaussian Posterior

- diagonal Gaussian posterior
- each weight requires a separate mean $ variance ($\beta = \{\mu, \sigma^2\}$, both of size w)
- cannot compute derivative exactly, so apply MC integration :
  $L^E(\boldsymbol{\beta}, \mathcal{D}) \approx \frac{1}{S} \sum_{k=1}^{S} L^N\left(\mathbf{w}^k, \mathcal{D}\right)$
- derive the following identities for the derivatives:

  $\nabla_{\boldsymbol{\mu}} \langle V(\boldsymbol{a}) \rangle_{\boldsymbol{a} \sim \mathcal{N}} = \langle \nabla_{\boldsymbol{a}} V(\boldsymbol{a}) \rangle_{\boldsymbol{a} \sim \mathcal{N}}, \quad \nabla_{\boldsymbol{\Sigma}} \langle V(\boldsymbol{a}) \rangle_{\boldsymbol{a} \sim \mathcal{N}} = \frac{1}{2} \langle \nabla_{\boldsymbol{a}} \nabla_{\boldsymbol{a}} V(\boldsymbol{a}) \rangle_{\boldsymbol{a} \sim \mathcal{N}}$