MaxEntropy Pursuit Variational Inference

ISNN'16, Moscow

Evgenii Egorov[1] Kirill Neklydov[2,3] Ruslan Kostoev[1] Evgeny Burnaev[1]

Skolkovo Institute of Science and Technology, Moscow, Russia
{e.egorov, r.kostoev, e.burnaev}@skoltech.ru

National Research University Higher School of Economics, Moscow, Russia

Samsung AI Center in Moscow, Moscow, Russia
k.necludov@gmail.com

June, 2019

Schedule

Overview

Probabilistic Machine Learning: Approach

- A probabilistic model considers the joint distribution over the
  - observed variables x (training data)
  - the hidden variables θ (the parameters of the interest)
- The Bayesian Inference suggests to estimate unknowns through **posterior distribution**:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_{\Theta} p(x|\theta)p(\theta)d\theta},$$

- where

$p(\theta)$ is the prior distribution,

$p(x|\theta)$ is the assumed model.

Probabilistic Machine Learning: Challenge

Benefits

- Prior Knowledge/Structure Incorporation
- Ensembles
- Uncertainty Estimation
- Coherent framework for the Sequential/Distributive Learning

Challenge

Evaluation of the posterior $p(\theta|x)$ is hard as require integration:

$$\int_{\Theta} p(x|\theta)p(\theta)d\theta,$$

Θ high-dimensional space, $p(x|\theta)$ complex model (i.e. Deep Neural Network).

Solution:

Approximate Inference

Approximate Inference: Approaches

MCMC

1. Choose the **proposal** distribution $q_\phi(\theta)$ from the **tractable** family Q_ϕ
2. Draw samples from a Markov chain with the $p(\theta|x)$ invariant distribution
3. Approximate expectations over $p(\theta|x)$ with averaging over the Markov chain samples

Variational Inference

1. Choose the **surrogate** distribution $q_\phi(\theta)$ from the **tractable** family Q_ϕ
2. Define the optimization problem by divergence minimization:

$$\mathcal{B}[p(x|\theta)p(\theta); q_\phi(\theta)]$$

3. $q^*(\theta) = \arg\min_\phi \mathcal{B}[p(x|\theta)p(\theta); q_\phi(\theta)]$

"Tractable"

- Easy to sample from
- Easy to evaluate log-density
- . . .

Approximate Inference: Challenges

MCMC

Pros

- Allow to trade computation time for increased accuracy
- Asymptotically unbiased
- Provide samples

Cons

- Sensitive to proposal selection
- Convergence diagnostic is hard
- Bad scalability (both on data and dimension)
- Provide only samples

Variational Inference

Pros

- Scalability: Fine with stochastic optimization and amortization
- Easy to use incorporate the structure of the problem to efficient optimization
- Flexible Q_λ families parametrized by DNN
- Provide approximations with density

Cons

- Biased (underestimating the posterior variances)
- Optimization is hard

MaxEntropy Pursuit Variational Inference

MPVI: General Idea

Solution Plan

1. Select family of simple "base learners" $q(\theta) \in Q_\lambda$, i.e. Normal Densities
2. Iteratively improve the approximation by additive convex update $q_t(\theta) = (1-\alpha)q_{t-1}(\theta) + \alpha q_t(\theta)$
3. Perform functional gradient descent over KL-divergence to select each component

Challenges

- Avoid degenerate solution (mixture of delta functions)
- Keep inference data scalable and computationally efficient
- Avoid model specific work

MPVI | Component Optimization: Problem

- Given some approximation of the posterior distribution q_t.
- Goal is to improve accuracy of the approximation
- In terms of the KL-divergence by using the additive mixture:

$$q_{t+1} = (1-\alpha)q_t + \alpha h,\ \alpha \in (0;1),\ h \in Q.$$

Using Maximum Entropy Approach we can state the following optimization problem:

$$\max_{h \in \mathcal{Q}} \mathcal{H}[h], s.t.$$

$$\mathcal{F}[q_{t+1}] - \mathcal{F}[q_t] > 0.$$

Using Taylor expansion, we obtain the constraint in the following form:

$$\mathcal{F}[q_{t+1}] - \mathcal{F}[q_t] = \alpha \left\langle h - q_t, \log \frac{L(\theta)}{q_t} \right\rangle - \alpha^2 \int \frac{(h-q_t)^2}{q_t} d\theta + o\left(\alpha \left\| \frac{h-q_t}{q_t} \right\|_2\right).$$

Considering the first order terms, we get the following optimization problem:

$$\max_{h \in Q} \mathcal{H}[h] + \lambda \left\langle h, \log \frac{L(\theta)}{q_t} \right\rangle.$$

MPVI | Component Optimization: Solution

$$\max_{h \in Q} \mathcal{H}[h] + \lambda \left\langle h, \log \frac{L(\theta)}{q_t} \right\rangle .$$

Problem Proprieties

- Strictly concave over h
- Could be solved by stochastic gradient optimization, i.e. scalable over dataset size
- Exact solution is

$$h^* = \frac{1}{Z(\lambda)} \left[\frac{L(\theta)}{q_t} \right]^\lambda = \arg\min_{h \in Q} D_{KL} \left(h \middle|\middle| \frac{1}{Z(\lambda)} \left[\frac{L(\theta)}{q_t} \right]^\lambda \right) .$$

λ selection heuristic

For U (uniform) and $p : \mathcal{H}[p] > \mathcal{H}[U]$, $T_\lambda : p \to \frac{p^\lambda(\theta)}{\int p^\lambda(\theta) d\theta}$, $\lambda > 0$ holds:

$$D_{KL}(U||p) > D_{KL}(U||T_\lambda p), \text{ for } \lambda > 1,$$
$$D_{KL}(U||p) < D_{KL}(U||T_\lambda p), \text{ for } \lambda < 1.$$

MPVI | Connection with Variational Inference

Variational Inference optimization problem:

$$\arg\max_{h \in Q} \int h \log \frac{L(\theta)}{h} d\theta.$$

For $\lambda = 1$ **MPVI** optimization problem:

$$\arg\max_{h \in Q} \mathcal{H}[h] + \left\langle h, \log \frac{L(\theta)}{q_t} \right\rangle = \arg\max_{h \in Q} \underbrace{\int h \log \frac{L(\theta)}{h} d\theta}_{\text{term (1)}} - \underbrace{\int h \log q_t d\theta}_{\text{term (2)}}.$$

We can note than:

- Term (1) corresponds to the standard **Variational Inference** objective
- Term (2) plays the role of **similarity penalty** with the current solution q_t

MPVI | Weight Optimization

After we obtain the new mixture component h for the current variational approximation q_t, we should select the mixture weight α to obtain a new variational approximation as a convex combination:

$$q_{t+1}(\theta) = (1-\alpha)q_t(\theta) + \alpha h(\theta).$$

Hence, let us state the optimization problem over $\alpha \in (0;1)$:

$$\min_{\alpha \in (0;1)} D_{KL}((1-\alpha)q_t(\theta) + \alpha h(\theta) || p(\theta|X)).$$

Theoretical solution

- Convex problem
- $\alpha^* = -\dfrac{\int \frac{1}{p(\theta|X)} q_t(h-q_t)d\theta}{\int \frac{1}{p(\theta|X)}(h-q_t)^2 d\theta} = -\dfrac{\int \frac{1}{L(\theta)} q_t(h-q_t)d\theta}{\int \frac{1}{L(\theta)}(h-q_t)^2 d\theta}.$

Implementation

In practise we use stochastic gradient descent over α

MPVI Incremental Learning

Problem: Neural Networks suffer from Catastrophic Forgetting
Solution: $p(\theta|x, x^{\text{new}}) \approx (1-\alpha)q(\theta|x) + \alpha q(\theta|x^{\text{new}})$

Experiment

- Dataset: MNIST, 10 classes classification
- Incremental setting: pair classes arrive: 0 vs 1, 2 vs 3, .etc
- Neural Network: LeNet-5
- Prior: Factorized Normal
- Metric: Accuracy