ECEN 740: Machine Learning Engineering

Lecture 2: Variational Inference

Tie Liu

Texas A&M University

Outline

- Main topics:
  - Problem setup and main idea
  - Evidence lower bound (ELBO)
  - Mean-field variational family
  - Coordinate ascent mean-field variational inference
- Main reference:
  - "Variational Inference: A Review for Statistician" by David M. Blei, Alp Kucukelbir and Jon D. McAuliffe (`https://arxiv.org/abs/1601.00670`)

Problem setup

- Let $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ be a set of *observed* variables and $\mathsf{z} = (\mathsf{z}_1, \ldots, \mathsf{z}_m)$ be a set of *latent* variables, with joint density $p(z, x)$

- The *inference* problem is to compute the *conditional* density of the latent variables given the observations $p(z|x)$, which can be written as
$$p(z|x) = \frac{p(z, x)}{p(x)}$$

- The denominator contains the *marginal* density of the observations, also known as the *evidence*, which can be calculated by *marginalizing* out the latent variables from the joint density:
$$p(x) = \int p(z, x) dz$$

- For many models, this evidence integral is unavailable in closed form or requires exponential time to compute – this is why exact inference in such models is hard and *approximate* inference is needed

- Note we assume that *all* unobserved quantities of interest are represented as latent random variables. This includes parameters that might govern all the data as well as latent variables that are "local" to individual data points such as the target variables in the learning problems

Markov chain Monte Carlo (MCMC)

For decades, the dominant paradigm for approximate inference has been *Markov chain Monte Carlo (MCMC)*:

- In MCMC, we first construct an ergodic Markov chain on z whose stationary distribution is the posterior $p(z|x)$
- Then, we sample from the chain to collect samples from the stationary distribution
- Finally, we approximate the posterior with an empirical estimate constructed from (a subset of) the collected samples

- MCMC sampling has evolved into an indispensable tool to the modern Bayesian statistician. Landmark developments include the *Metropolis-Hastings* algorithm, the *Gibbs* sampler, and its application to Bayesian statistics

- However, there are problems for which we cannot easily use this approach. These arise particularly when we need an approximate conditional faster than a simple MCMC algorithm can produce, such as when data sets are large or models are very complex

- In these settings, *variational inference* provides a good alternative approach to approximate Bayesian inference

Variational inference

Rather than use sampling, the main idea behind variational inference is to use *optimization*:

- First, we posit a family of approximate densities $\mathcal{Q}$. This is a set of densities over the latent variables
- Then, we try to find the member of that family that minimizes the *Kullback-Leibler (KL) divergence* to the exact posterior:

$$q^*(z) = \arg\min_{q \in \mathcal{Q}} KL\left(q(z) \| p(z|x)\right)$$

- Finally, we approximate the posterior $p(z|x)$ with the *optimized* member of the family $q^*(z)$

Comparing variational inference and MCMC

- MCMC methods tend to be more computationally intensive than variational inference but they also provide *guarantees* of producing (asymptotically) exact samples from the target density

- Variational inference does not enjoy such guarantees – it can only find a density close to the target—but tends to be *faster* than MCMC. Because it rests on *optimization*, variational inference easily takes advantage of methods like stochastic optimization and distributed optimization

- Thus, variational inference is suited to large data sets and scenarios where we want to quickly explore many models; MCMC is suited to smaller data sets and scenarios where we happily pay a heavier computational cost for more precise samples

- The relative accuracy of variational inference and MCMC is still unknown

Bayesian mixture of Gaussians

- In this example, the latent variables are $\mathsf{c} = (\mathsf{c}_1, \ldots, \mathsf{c}_n)$, where each c_i is drawn independently according to a categorical distribution over $[K]$, and $\mu = (\mu_1, \ldots, \mu_K)$, where each μ_k is drawn independently according to $\mathcal{N}(0, \sigma^2)$. The latent variables c and μ are independent of each other

- There are n observations $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$, where given (c, μ), each x_i is drawn independently according to $\mathcal{N}(\mu_{\mathsf{c}_i}, 1)$

- The joint density of the latent and observed variables is

$$p(c, \mu, x) = p(c)p(\mu)\prod_{i=1}^{n} p(x_i | c_i, \mu)$$

- Here, the evidence is

  $$p(x) = \sum_{c} p(c) \int p(\mu) \prod_{i=1}^{n} p(x_i|c_i,\mu) d\mu$$

  where each individual integral is computable, thanks to the *conjugacy* between the Gaussian prior and the Gaussian likelihood. But there are K^n of them, one for each possible value of $c = (c_1, \ldots, c_n)$, so computing the evidence is exponential in n and hence intractable

Evidence lower bound (ELBO)

- Recall that variational inference amounts to solving the optimization problem:

$$q^*(z) = \arg\min_{q \in \mathcal{Q}} KL\left(q(z) \| p(z|x)\right)$$

- Once found, $q^*(z)$ is the best approximation of the conditional, within the family $\mathcal{Q}$

- However, this objective is not computable because it requires computing the evidence $\log p(x)$. To see why, recall that KL divergence is

$$\begin{aligned} KL\left(q(z) \| p(z|x)\right) &= \mathbb{E}[\log q(\mathsf{z})] - \mathbb{E}[\log p(\mathsf{z}|x)] \\ &= \mathbb{E}[\log q(\mathsf{z})] - \mathbb{E}[\log p(\mathsf{z}, x)] + \log p(x) \end{aligned}$$

where all expectations are over $\mathsf{z} \sim q(z)$. This reveals its dependence on $\log p(x)$

- Because we cannot compute the KL, we optimize an alternative objective that is equivalent to the KL up to an added constant:

$$\begin{aligned} ELBO(q) &= \log p(x) - KL\left(q(z)\|p(z|x)\right) \\ &= \mathbb{E}[\log p(\mathsf{z}, x)] - \mathbb{E}[\log q(\mathsf{z})] \end{aligned}$$

- This function is called the *evidence lower bound (ELBO)*. The ELBO is the negative KL divergence plus $\log p(x)$, which is a constant with respect to $q(z)$. Thus, maximizing the ELBO is equivalent to minimizing the KL divergence

- The name ELBO follows from the fact that

$$ELBO(q) \leq \log p(x)$$

  for any $q(z)$ due to the *non-negativity* of KL divergence

The mean-field variational family

- We now describe a variational family $\mathcal{Q}$, to complete the specification of the optimization problem. The more complex the family is, the closer the optimized member is to the posterior; on the other hand, it is generally more difficult to optimize over a complex family than a simple family

- Here we focus on the *mean-field variational family*, where the latent variables are mutually *independent* and each governed by a *distinct* factor in the variational density:

$$q(z) = \prod_{j=1}^{m} q_j(z_j)$$

  where each latent variable z_j is governed by its own variational factor, the density $q_j(z_j)$. In optimization, these variational factors are chosen to maximize the ELBO

- Finally, though we focus on mean-field inference here, researchers have also studied more complex families. One way to expand the family is to allow *dependencies* between the variables; this is called *structured variational inference*. Another way to expand the family is to consider *mixtures of variational densities*, i.e., additional *latent* variables within the variational family

- Both of these methods potentially improve the fidelity of the approximation, but there is a trade off. Structured and mixture-based variational families come with a more difficult-to-solve variational optimization problem

Coordinate ascent variational inference (CAVI)

- Using the ELBO and the mean-field family, we have cast approximate conditional inference as an optimization problem. Here we describe one of the most commonly used algorithms for solving this optimization problem, *coordinate ascent variational inference (CAVI)*

- CAVI *iteratively* optimizes each factor of the factor of the mean-field variational density, while holding the others *fixed*. It climbs the ELBO to a *local* optimum

- More specifically, consider the jth latent variable z_j and fix the other variational factors $q_\ell(z_\ell)$, $\ell \neq j$. Under the *mean-field* assumption of the variational family, the ELBO can be written as:

$$ELBO(q_j) = \mathbb{E}_j[\mathbb{E}_{-j}[\log p(\mathsf{z}_j, \mathsf{z}_{-j}, x)]] - \sum_{\ell=1}^{m} \mathbb{E}_\ell[\log q_\ell(\mathsf{z}_\ell)]$$

  where $\mathsf{z}_{-j} := (\mathsf{z}_\ell : \ell \neq j)$, and $\mathbb{E}_\ell[\cdot]$ is the expectation taken over $\mathsf{z}_\ell \sim q_\ell(z_\ell)$

- Let $q_j^*(z_j)$ be a density such that

$$q_j^*(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j, \mathsf{z}_{-j}, x)]\}$$

- We have

$$\begin{aligned} ELBO(q_j) &= \mathbb{E}_j[\log q_j^*(\mathsf{z}_\mathsf{j})] - \mathbb{E}_j[\log q_j(\mathsf{z}_j)] - \sum_{\ell \neq j} \mathbb{E}_\ell[\log q_\ell(\mathsf{z}_\ell)] + Const \\ &= -KL(q_j(z_j) \| q_j^*(z_j)) - \sum_{\ell \neq j} \mathbb{E}_\ell[\log q_\ell(\mathsf{z}_\ell)] + Const \end{aligned}$$

and

$$\begin{aligned} ELBO(q_j^*) &= \mathbb{E}_j^*[\log q_j^*(\mathsf{z}_\mathsf{j})] - \mathbb{E}_j^*[\log q_j^*(\mathsf{z}_j)] - \sum_{\ell \neq j} \mathbb{E}_\ell[\log q_\ell(\mathsf{z}_\ell)] + Const \\ &= -\sum_{\ell \neq j} \mathbb{E}_\ell[\log q_\ell(\mathsf{z}_\ell)] + Const \end{aligned}$$

where $\mathbb{E}_j^*[\cdot]$ is the expectation taken over $\mathsf{z}_j \sim q_j^*(z_j)$

- Again by the *non-negativity* of KL divergence, we have

  $$ELBO(q_j) \leq ELBO(q_j^*)$$

  for *any* variational factor q_j

- We thus conclude that the *update rule* for the variational factor q_j (while fixing q_ℓ, $\ell \neq j$) is:

  $$q_j^*(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j, \mathsf{z}_{-j}, x)]\}$$

- For the Bayesian mixture of Gaussians, it can be shown that the update for each c_i is a categorical distribution over $[K]$, and the update for each μ_k is *Gaussian*. All parameters of the updates can be calculated in *closed form*

- The detailed calculations for the Bayesian mixture of Gaussian are referred to *Section 3* of the main reference

Summary

- The inference problem is to compute the *conditional* density of the latent variables given the observations, for which the main challenge is that the *evidence integral* is unavailable in closed form or requires exponential time to compute

- Variational inference is an *approximate* inference technique that is based on maximizing the *ELBO* and is particularly suited for *large-scale* inference problems

- The choice of the variational family (the approximation *model*) plays a crucial role to the viability of this approach

- The importance of modeling and optimization will be continuously highlighted when we move to learning next