Variational Inference - Lecture 2

Guido Sanguinetti

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
gsanguin@inf.ed.ac.uk

April 3, 2019

Today's lecture

1. Variational inference - mathematical foundations
2. Mean field - Variational Bayes
3. Parametric variational inference

The Bayesian Inference problem

- Bayesian inference provides an appealing mathematical formulation to perform learning/ prediction in uncertain scenarios
- The world (system) is divided in two sets of random variables: latent (or hidden) θ and visible (or observed) $\mathbf{x}$
- Assumptions are encoded in a *prior distribution* $p(\theta)$ **and** a likelihood function connecting latent to visibles $p(\mathbf{x}|\theta)$
- Then we update our beliefs on the latent world according to Bayes rule

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} \tag{1}$$

where $p(\mathbf{x})$ is the *marginal likelihood* (probability of the visibles regardless of the latents).

The Bayesian Inference problem

- Bayesian inference provides an appealing mathematical formulation to perform learning/ prediction in uncertain scenarios
- The world (system) is divided in two sets of random variables: latent (or hidden) θ and visible (or observed) $\mathbf{x}$
- Assumptions are encoded in a *prior distribution* $p(\theta)$ **and** a likelihood function connecting latent to visibles $p(\mathbf{x}|\theta)$
- Then we update our beliefs on the latent world according to Bayes rule

  $$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} \tag{1}$$

  where $p(\mathbf{x})$ is the *marginal likelihood* (probability of the visibles regardless of the latents).
- **IMPOSSIBLE**: we'd need to evaluate the likelihood for *all possible* configurations of the latents!!!!

The Variational Principle

- Various strategies exist for approximating posterior distributions.
- One popular class constructs Markov chains that asymptotically sample from the posterior (MCMC).
- Variational methods recast inference as optimisation in function space, using methods of calculus of variation.
- Specifically, one minimises the *Kullback-Leibler divergence* (or cross-entropy)

$$KL[q(\theta)\|p(\theta|\mathbf{x})] = \int d\theta q(\theta) \log \frac{q(\theta)}{p(\theta|\mathbf{x})} \tag{2}$$

  where $q(\theta)$ is an approximating distribution
- Since the marginal likelihood does not depend on the data, its knowledge is not required to find the optimum
- Free form optimisation problem is just as hard; need approximations

Why the KL: the ELBO

- We've seen in the last lecture that EM is based on optimising a lower bound on the log-marginal likelihood (*evidence*)
- Explicitly

$$\log p(\mathbf{x}) = \log \int d\theta p(\mathbf{x},\theta) = \log \int d\theta \frac{p(\mathbf{x},\theta)}{q(\theta)} q(\theta) \geq \int d\theta q(\theta) \log \frac{p(\mathbf{x},\theta)}{q(\theta)} = p(\mathbf{x}) - KL[q(\theta) \| p(\theta|\mathbf{x})] \tag{3}$$

- Minimising the KL divergence makes the *evidence lower bound (ELBO)* tight

Why KL: the perturbative expansion

- Consider a complicated probability distribution $P = \exp(H)$
- We would like to replace it with an easier probability distribution $Q = \exp(H_0)$
- We define an *intermediate* distribution $Q_\lambda = \exp(H_0)\exp[-\lambda(H_0 - H)]$ that is P for $\lambda = 1$ and Q for $\lambda = 0$
- Taylor expand

$$P = \exp(H) = \exp(H_0)\exp[-\lambda(H_0 - H)] =$$
$$= \exp(H_0)\left[1 - \lambda(H_0 - H) + O(\lambda^2)\right] = Q\left[1 - \lambda Q \log\frac{Q}{P} + O(\lambda^2)\right]$$

- So minimizing KL (on average) minimises the first order correction

How to minimise KL

- KL is a *functional* of the approximating distribution q
- Functionals can be thought of as *functions of functions*
- To minimise a functional, one sets its *functional derivative* to zero
- **Excursus**: let's work out on the board!

What about parameters?

- Functional optimisation of KL enables approximate posterior inference
- What about model parameters?

What about parameters?

- Functional optimisation of KL enables approximate posterior inference
- What about model parameters?
- ELBO can be used as a surrogate of the marginal likelihood and optimised w.r.t. to model parameters (either in the prior or likelihood), directly (gradient descent) or analytically when possible
- Sometimes called VBEM

Talk outline

1. Variational inference - mathematical foundations
2. Mean field - Variational Bayes
3. Parametric variational inference

Factorizing complicated distributions

- Most complex models involve several latent variables $\theta_1, \ldots, \theta_N$
- Even if they are *a priori* independent, the data usually couples the latent variables making inference complicated
- *Mean-field* variational inference breaks these dependencies by replacing them with *averaged effects*

Coordinate Ascent Variational Inference (CAVI)

- Assume the approximating distribution is factorized

$$q\left(\theta_1, \dots, \theta_N\right) = q_1(\theta_1), \cdots, q_N(\theta_N)$$

- Computing functional derivatives of (3) and setting to zero we get

$$q_j \propto \exp\langle \log p(\mathbf{x}, \theta) \rangle_{\tilde{j}}$$

  where $\langle\rangle_{\tilde{j}}$ means expectation w.r.t. all the latent variables except θ_j

- Provided you can compute these expectations, iterating these fixed point equations leads to a (local) optimum

Exercise

Consider a Gaussian mixture model with Dirichlet priors over the mixing components and normal-inverse Wishart priors over the component means/ variances. Work out the CAVI algorithm. See excellent worked out example here
`https://rpubs.com/cakapourani/variational-bayes-gmm`

Talk outline

1. Variational inference - mathematical foundations
2. Mean field - Variational Bayes
3. Parametric variational inference

Families of distributions

- Mean-field posits a factorised form for the approximating distribution but does not restrict the functional form of the factors
- Alternatively, one could choose a parametric family for the approximating distribution (e.g. a Gaussian)
- Then KL becomes a normal *function* of the parameters of the distribution and one may compute its gradient and optimise
- **CAVEAT**: you will still need to be able to compute expectations to get this gradient analytically

Exercise

Compute the Gaussian variational approximation for a standard normal latent variable observed through an exponential link with Poisson noise, i.e.

$$p(\mathbf{x}|\theta) = Poisson(\exp(\theta)), \theta \sim \mathcal{N}(0, 1)$$