Bayesian Machine Learning - Lecture 8

Guido Sanguinetti

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

March 4, 2015

Today's lecture

1. Variational inference
2. Monte Carlo methods

Free energy revisited

- Recall that by postulating the inference model in terms of approximating probability distributions we obtained a general variational inference principle for Bayesian inference
- The central task is minimising the free energy

$$G(q) = \langle E(x) \rangle_q - H[q(x)]$$

- BP expresses the free energy in terms of families of one node and two node beliefs with constraints
- These do not necessary correspond to any probability distribution $\longrightarrow$ problems with convergence

Variational Bayes

- As opposed to BP, start directly with free energy minimisation within a certain family of approximating distributions q
- Common choices are factorized distributions (Variational Bayes EM, VBEM), or parametric families (e.g. Gaussian)
- In the factorised case, the free energy then becomes

$$\begin{aligned} G[q(x)] =& G[q_1(x_1), \ldots, q_N(x_N)] = \\ =& \langle E(x) \rangle_q + \sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i) \end{aligned} \tag{1}$$

- This leads to update equations

VBEM update equations

- For finite state space, computing an expectation is the same as forming the dot product $\mathbf{p}^T f(x_i)$
- The free energy in (1) then depends on q_i as

$$G[q_i] = \mathbf{q}_i^T \langle E(x) \rangle_{q_{\setminus i}} + \mathbf{q}_i^T \log \mathbf{q}_i + const$$

  where subscript $\setminus i$ indicates all indices except i-th
- Taking derivatives and equating to zero, we get

$$q_i(x_i) \propto \exp \left[-\langle E(x) \rangle_{q_{\setminus i}} \right]$$

- In other words, the distribution over the i-th variable depends only on *the statistics* of the other variables

Variational GMMs

- Suppose we wish to compute the joint posterior over the class variables z AND model parameters Θ in a Gaussian mixture model
- This is analytically intractable; approximate the posterior as $q(z)q(\Theta)$ and minimize KL

$$\begin{aligned} KL[q\|p(z,\Theta|\mathbf{x})] &= \log\frac{1}{Z} + KL[q\|p(z,\Theta,\mathbf{x})] \\ &= \log\frac{1}{Z} + E_{q(z)q(\Theta)}[\log p(\mathbf{x},z,\Theta)] + H[q(z)] + H[q(\Theta)] \end{aligned} \tag{2}$$

where $H = -E_q[\log(q)]$ is the *entropy* of the approximating distributions

Variational GMMs

- The KL divergence in equation (1) can be optimised iteratively by zeroing the *functional derivatives* w.r.t. the approximating distributions
- The resulting update rules are

$$q(z) \propto \exp\left\{-E_{q(\Theta)}\left[\log p(\mathbf{x}, z, \Theta)\right]\right\}$$
$$q(\Theta) \propto \exp\left\{-E_{q(z)}\left[\log p(\mathbf{x}, z, \Theta)\right]\right\}$$

- These are iterated until the KL divergence stops decreasing (or the parameters stop changing)

Sample-based approximations

- Variational methods are limited by being biased (choice of approximating family) and constrained (you still need to be able to compute expectations)
- An alternative is to draw samples from the distribution of interest and compute sample based estimates
- These will be asymptotically exact, with error decreasing with the square root of the number of (effective) samples
- We do not have the distribution of interest, but there are methods to draw samples

Importance sampling

- We always assume to have the distribution of interest p *up to an unknown multiplicative scaling factor* $\frac{1}{Z_p}\tilde{p}$ (the joint is known)
- We introduce a simple distribution q from which we can draw samples
- We then approximate any expectations as

$$\begin{aligned} E_p[f] &= \frac{\int f(x)\tilde{p}(x)dx}{\int \tilde{p}(x)dx} \\ &= \frac{\int f(x)\tilde{p}(x)\frac{q(x)}{q(x)}dx}{\int \tilde{p}(x)\frac{q(x)}{q(x)}dx} \approx \frac{\sum_{j=1}^{M} f(x_j)w_j}{\sum_{j=1}^{M} w_j} \end{aligned} \tag{3}$$

where x_j is a set of M points sampled from q (easy to sample from) and $w_j = \frac{\tilde{p}(x_j)}{q(x_j)}$ are the *importance weights*

Comments on importance sampling

- Adjusting by the importance weights ensures that we are effectively sampling from the correct distribution
- Ideally, one would want the proposal distribution q to be as close as possible to the target p
- Discuss possible drawbacks of importance sampling

Markov Chain Monte Carlo

- An alternative set of algorithms for Bayesian computation relies on constructing an iterative sampling strategy that explores the latent variables according to the posterior distribution
- The trick is to construct a random walk (Markov Chain) and then bias it in a way that (after some time) the variables are effectively sampled from the posterior
- This is the core idea of Markov Chain Monte Carlo methods
- The biasing is usually obtained either through a rejection procedure, or by drawing the samples from distributions that are already very close to what we want

MCMC conditions

- MCMC works by building a random walk that converges to the target distribution
- A necessary condition is the invariance of the target distribution p under $T(x', x)$, the*transition kernel* of the random walk

$$p(x) = \int T(x, x')p(x')dx'$$

- The Markov chain must be *ergodic* (time averages = sample averages); usually one checks for detailed balance (reversibility of transitions, not quite equivalent but sufficient)
- The Markov chain must be irreducible and aperiodic

MCMC: the Metropolis-Hastings algorithm

- The simplest such strategy is the *Metropolis-Hastings algorithm*
- You need a proposal distribution $q(z|z_0)$ from which you can draw samples
- Having drawn a sample z_1, you accept it or reject it with probability

$$r = \min\{1, \frac{p(z_1)}{p(z_0)} \frac{q(z_0|z_1)}{q(z_1|z_0)}\}$$

- You then carry on using $q(z|z_1)$ as the updated proposal. Asymptotically, this converges to the posterior

Validity of MH sampler

- Irreducibility and aperiodicity depend on the proposal distribution (true unless pathological)
- To verify ergodicity we check detailed balance (only needed for when the samples are accepted)

$$\begin{aligned} T(x,y)p(y) =& \min\{1, \frac{p(x)}{p(y)}\frac{q(y|x)}{q(x|y)}\}q(x|y)p(y) = \\ =& \min\{q(x|y)p(y), p(x)q(y|x)\} = T(y,x)p(x) \end{aligned} \tag{4}$$

- MH therefore gives a generic procedure for asymptotically drawing samples from any distribution
- However, the time it may take to reach the asymptotic detailed balance state may be very long

MCMC: Gibbs sampling

- Often the latent variables are structured in groups $z = (z_1, \ldots, z_k)$
- If the *conditional posteriors* $p(z_i|z_{-i}, \mathbf{x})$ can be computed, then one can use Gibbs sampling
- Practically, one samples iteratively from all conditional posteriors one at the time
- Easily shown to be a special case of MH when the acceptance probability is 1
- Generally only slow

Gibbs sampling for a Bayesian GMM

- Exrecise: work out the Gibbs sampler for the Bayesian Gaussian Mixture model with two spherical Gaussian components with fixed variance σ^2
- Place a Beta(1,1) prior over the mixing components π_1, π_2 and a spherical Gaussian prior $\mathcal{N}(0, 10)$ over the component means

Tomorrow's lab class

- Sample 50 2D points each from Gaussians with means (-1,-1), (1,3) and (3,0) all with variance 1
- Implement EM for GMMs with k components of fixed variance 1
- Run EM on the data for $k = 2, \ldots, 8$; compute and plot the BIC score for these 7 models

and thanks to the ERC!