Variational Optimisation by Marginal Matching

Neil D. Lawrence
neill@cs.man.ac.uk

7th December 2007

Variational Approximation

- Variational deconstruction of the log likelihood, $L = \log p(\mathbf{y})$

$$L = \sum_{n=1}^{N} \langle \log p(y_n|f_n) \rangle + \langle \log p(\mathbf{f}) \rangle - \langle \log q(\mathbf{f}) \rangle + \left\langle \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right\rangle.$$

- Rewrite by introducing 'site functions', $\{t(f_n)\}_{n=1}^{N}$.

$$L = \sum_{n=1}^{N} \left\langle \log \frac{p(y_n|f_n)}{t(f_n)} \right\rangle + \left\langle \log \prod_{n=1}^{N} t(f_n)\, p(\mathbf{f}) \right\rangle - \langle \log q(\mathbf{f}) \rangle + \left\langle \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right\rangle.$$

- If 'zeroth moment' of $q(f_n)$ and $\hat{p}(f_n)$ are matched we can write

$$L = \log \int \prod_{n=1}^{N} t(f_n)\, p(\mathbf{f})\, d\mathbf{f} - \sum_{n=1}^{N} \mathrm{KL}(q(f_n)\,||\,\hat{p}(f_n)) + \mathrm{KL}(q(\mathbf{f})\,||\,p(\mathbf{f}|\mathbf{y})).$$

Bound on EP Energy

- Dropping the last term gives us the variational lower bound on the likelihood,

$$\mathscr{L} = \log \int \prod_{n=1}^{N} t(f_n)\, p(\mathbf{f})\, \mathrm{d}\mathbf{f} - \sum_{n=1}^{N} \mathrm{KL}\left(q(f_n) \,||\, \hat{p}(f_n)\right).$$

- The negative EP energy is given by

$$\mathscr{L}_{\mathrm{EP}} = \log \int \prod_{n=1}^{N} t(f_n)\, p(\mathbf{f})\, \mathrm{d}\mathbf{f}$$

  so we also have a *lower bound on the negative EP energy.*
- The bound can be maximised with respect to the site functions, $t(f_n)$.

Quality of EP Approximation

- Difference between truth and EP approximation is given by

$$\Delta_{\mathrm{EP}} = L - L_{\mathrm{EP}} = \mathrm{KL}\left(q\left(\mathbf{f}\right) \| p\left(\mathbf{f}|\mathbf{y}\right)\right) - \sum_{n=1}^{N} \mathrm{KL}\left(q\left(f_n\right) \| \hat{p}\left(f_n\right)\right)$$

- Minima of these KL divergences *coincide*!
- Zero forcing is a problem though! *e.g.* Gaussian case and $p\left(y_n|f_n\right)$ is a step function.
  1. Perhaps EP is doing a much better job and *matching* these KL divergences.
  2. Can use EP energy in conjunction with variational methods. [1].

References

[1] M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.