T-61.5040
Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer:	**Petteri Pajunen**
Course Assistant:	**Ville Viitaniemi**

Lecture 9: 21.3.2007

Posterior Approximation: Laplace and Variational Methods

Parametric Approximation

- some ways to compute the posterior:
  - normalizable closed-form (rarely possible)
  - MCMC (works in theory, convergence is generally very slow)
  - parametric approximation

- closed-form is obviously a good choice when possible
- MCMC can become infeasible in 'large' problems
- MCMC works if computational costs and memory costs are ignored
- since costs always matter, parametric approximation can be a practical compromise between accuracy and cost of implementing and computing the solution

- this compromise is generally not optimal
- approximate posterior is a wrong posterior, so the choice is between different wrong solutions
- since wrong solutions have consequences, their cost should also be considered
- a probability model does not contain information about such costs

- therefore any single approximation even for the same model cannot be generally optimal
- for example, the same model might describe the absorption of aspirin and the absorption of an antibiotic. Consequences of errors are certainly different.
- in practice, approximations have to be made even if non-probability costs are not explicity considered
- it is better to think about different approximation methods as heuristics with varying properties

- pros/cons of parametric approximation
  - simple representation of the posterior
  - possibly low computational cost
  - overfitting (or underfitting)
  - accurate approximation often means high computational cost

Point Estimation

- simplest posterior approximation method is *point estimation*
- meaning: choose a single value θ_0, and use it to represent the posterior
- formally the point estimate is the Dirac function $\delta(\theta - \theta_0)$

- examples:
  - posterior mean: $\theta_0 = \mathrm{E}(\theta|y)$
  - posterior median (cont θ): $\mathbb{P}(\theta \leq \theta_0) = 0.5$
  - posterior mode: $\theta_0 = \mathrm{argmax}_{\theta_0}\, p(\theta_0|y)$
  - minimum mean-square: $\theta_0 = \mathrm{argmin}_{\theta_0}\, \mathrm{E}((\theta - \theta_0)^2|y)$

- for some approximations, many of these are equivalent
- e.g. a Normal distribution has the same mean, mode, median and MSE (computed wrt to the Normal)
- the posterior mode has some properties that warrant further discussion
- if $p(\theta) \propto c$, then the posterior mode is the maximum likelihood estimate

- in general, the posterior mode is called the MAP-estimate (maximum a posteriori)
- there are certain classes of models, for which the posterior mode can be found iteratively using the EM-algorithm
- more on this in the next lecture

- "Bayes Central Limit Theorem": if

$$y_1, y_2, \ldots, y_n \text{ is i.i.d. from } p(y|\theta_0) \implies$$
$$p(\theta|y_1, \ldots, y_n) \to N(\theta|\theta_0, (nI(\theta_0))^{-1})$$

  as n goes to infinity (see Gelman for assumptions)
- $I(\theta_0)$ is the Fisher information (used earlier to define the Jeffreys' prior)
- the covariance $(nI)^{-1}$ goes to zero when $n \to \infty$
- motivates the use of posterior mode, although this holds only asymptotically

Laplace Approximation

- point estimates ignore *posterior uncertainty*
- variance can be added to a point estimate to obtain *local* uncertainty
- Bayes CLT suggests approximating the posterior by a Normal distribution
- the mean is the posterior mode, and the variance is fitted to the posterior at the mode

- suppose we have found the posterior mode θ_0
- Taylor-expand $\log p(\theta|y)$ at the mode:

$$\log p(\theta|y) = \log p(\theta_0|y) + (\log p(\theta_0|y))'(\theta - \theta_0) + \frac{1}{2}(\log p(\theta_0|y))''(\theta - \theta_0)^2 + \dots$$

- posterior is maximized at θ_0 so

$$(\log p(\theta_0|y))' = 0$$

- we also know that $(\log p(\theta_0|y))'' < 0$ since θ_0 is the mode
- dropping higher terms and taking exponents gives

$$p(\theta|y) \approx p(\theta_0|y)\exp(-\frac{1}{2}(-(\log p(\theta_0|y))'')(\theta - \theta_0)^2)$$

- this is a Normal distribution of θ

- this works even if $p(\theta|y)$ is not normalized, since the exponent tells us the mean and the variance
- we obtain $p(\theta|y) \approx N(\theta|\theta_0, \sigma^2)$
- the mean is obviously θ_0
- a Normal density has the exponent $-\frac{1}{2\sigma^2}(\theta - \theta_0)^2$

- therefore the variance is

$$\sigma^2 = \frac{-1}{(\log p(\theta_0|y))''}$$

- the variance does not depend on the normalization of $p(\theta|y)$ due to the logarithm and the differentiation
- if needed, one can compute the normalization factor

$$\frac{1}{\sqrt{2\pi\sigma^2}} = \sqrt{\frac{-(\log p(\theta_0|y))''}{2\pi}}$$

- one can also Laplace-approximate integrals

$$\mathrm{E}(h(\theta)|y) = \int h(\theta)p(\theta|y)d\theta$$

- use the same idea, but approximate the integrand $z(\theta) = h(\theta)p(\theta|y)$ around its mode θ_0
- Taylor-expand $\log(z(\theta))$ and take exponent:

$$z(\theta) \approx z(\theta_0)\exp(\frac{1}{2}[\log(z(\theta_0))]''(\theta-\theta_0)^2)$$

- this is an unnormalized Normal distribution with mean θ_0 and variance

$$\sigma^2 = \frac{1}{-(\log(z(\theta_0)))''}$$

- the integral is approximated as follows:

$$\begin{aligned}\mathrm{E}(h(\theta)|y) &\approx \int z(\theta_0)\exp(-\frac{1}{2\sigma^2}(\theta-\theta_0)^2)d\theta \\ &= z(\theta_0)\sqrt{2\pi\sigma^2}\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(\theta-\theta_0)^2)d\theta \\ &= z(\theta_0)\sqrt{2\pi\sigma^2}\end{aligned}$$

- the last integral vanished since its over $N(\theta|\theta_0,\sigma^2)$, which is a normalized probability distribution
- the Laplace-approximation is

$$\mathrm{E}(h(\theta)|y) \approx z(\theta_0)\sqrt{\frac{2\pi}{-(\log(z(\theta_0)))''}}$$

- note that θ_0 is the mode of $z(\theta)$ and not $p(\theta|y)$

- Laplace approximation allows local model averaging around the model: i.e. it describes the posterior uncertainty locally
- however, the approximation is centered on posterior mode so most point estimation problems remain
- when there are multiple modes and the highest one is a bad solution, Laplace approximation fails just as point estimates do
- demo: `laplace.R`

Kullback-Leibler Divergence

- posterior approximation can be thought like this:
  1. the true posterior is $p(\theta|y)$
  2. the approximate posterior is $q(\theta)$, constrained in some way (e.g. a parametric distribution)
  3. the problem is to find $q(\theta)$ that is 'as close as possible' to $p(\theta|y)$
- how to measure the closeness of distributions?

- denote a measure between distributions as $D(p, q)$
- there is no unique way of choosing the measure: the reasons are the same that discredit point estimation
- for example if $D(p, q) \ll D(p, q')$, it is possible that q causes a very costly error while q' doesn't
- probabilites cannot measure such costs so there is no generally optimal measure

- *Kullback-Leibler divergence* is an information-theoretic measure between distributions
- KL divergence is defined as

$$D(q\|p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$$

- some properties: $D(p\|q)$ is
  - nonnegative: $D(p\|q) \geq 0$
  - equal to zero iff $q = p$
  - not symmetric: $D(p\|q) \neq D(q\|p)$

- there is a coding interpretation of KL divergence
- a discrete random variable x can be coded with average number of bits equal to its *entropy*
- each x_i has codelength $-\log p(x_i)$
- entropy is $H(x) = -\sum_i p(x_i) \log p(x_i)$
- assume we guess the distribution of x as $q(x)$

- the average codelength must be computed over the *real distribution*, so we obtain

$$H_q(x) = \sum_i -p(x_i) \log q(x_i)$$

- the KL divergence $D(p\|q)$ is now

$$D(p\|q) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} = -H(x) + H_q(x)$$

- then $H(x) + D(p\|q) = H_q(x)$, which means that the KL divergence is the number of extra bits caused by using the wrong distribution $q(x)$

Variational Approximation

- variational approximation means finding $q(\theta)$ which minimizes $D(q\|p)$, and using it as the approximate posterior
- the coding interpretation involved $D(p\|q)$, not $D(q\|p)$
- variation approximation uses $D(q\|p)$ because integration is performed over $q(\theta)$, not over the intractable $p(\theta|D)$
- since KL divergence is not symmetric, in general, $D(q\|p) \neq D(p\|q)$

- example: discrete $\theta \in \{\theta_0, \theta_1, \dots\}$
- point estimate: $q_0, q_1, \dots$ with $q_i \in \{0, 1\}$, $\sum_i q_i = 1$
- posterior: $p_0, p_1, \dots, \sum_i p_i = 1$, $p_i \geq 0$
- KL divergence is

$$\begin{aligned} D(q\|p) &= \sum_i q_i \log \frac{q_i}{p_i} \\ &= \log \frac{1}{p_j} \text{ (only one nonzero } q_j) \end{aligned}$$

- min D is obtained by maximizing p_j, so we get the posterior mode as the result

- the variational approximation in general tries to match the posterior area with the largest probability mass
- naturally the way the probability mass is measured depends on the approximating distribution $q(\theta)$
- in the posterior mode example, the posterior mass is measured at a single point
- if $q(\theta)$ is for example a Normal distribution, then the results are different

- example: multimodal posterior
- posterior mode is at a point where $p(\theta|y)$ is maximized
- but if the mode is on top a very narrow peak, it may be that the posterior mass contained in that peak is small
- variational approximation generally finds the peak with most posterior mass (exercise)
- demo: `varnorm.R`

- in variational approximation $q(\theta)$ must be constrained: otherwise $q(\theta) = p(\theta|y)$ and we get nowhere
- if $q(\theta)$ is a parametric distribution, minimizing KL-divergence is a parametric optimization problem
- in *free form approximation* it is only assumed that $q(\theta)$ is factorizable: $q(\theta) = q(\theta_1)q(\theta_2)\dots q(\theta_d)$
- this makes minimizing KL divergence a functional optimization problem

Freeform Approximation Example

- independent data $y_1, \dots, y_n$ from $N(\mu, \lambda^{-1})$
- λ is the *precision* (inverse of variance)
- the Jeffreys' prior for the mean is $p(\mu) \propto 1$
- for the variance σ^2, Jeffreys' prior is $p(\sigma^2) \propto \sigma^{-2}$
- since $\lambda = \sigma^{-2}$, we have to transform the prior:

$$p(\lambda) = p(\sigma^2)\left|\frac{d\sigma^2}{d\lambda}\right| \propto \lambda\left|\frac{d\sigma^2}{-\lambda^2 d\sigma^2}\right| = \lambda^{-1}$$

- posterior is

$$p(\mu, \lambda | y) \propto p(y | \mu, \lambda) \lambda^{-1}$$

- denote $\overline{y} = \frac{1}{n} \sum_i y_i$
- solve the variational approximation $q(\mu, \lambda) = q(\mu) q(\lambda)$

- assume that $q(\lambda)$ is known
- find $q(\mu)$ that minimizes $D(q \| p)$, keeping $q(\lambda)$ fixed
- to simplify notation, all constant terms wrt the current minimization are dropped (e.g $\log q(\lambda)$ in the next slide)
- the notation E_λ means an expectation over $q(\lambda)$
- E_μ is defined correspondingly

$$\begin{aligned} D(q \| p) &= \mathrm{E}_\mu \, \mathrm{E}_\lambda \log \frac{q(\mu)}{p(y|\mu, \lambda)} \\ &= \mathrm{E}_\mu \, \mathrm{E}_\lambda \left[\log q(\mu) + \frac{\lambda}{2} \sum (y_i - \mu)^2 \right] \\ &= \mathrm{E}_\mu \left(\log q(\mu) + \frac{1}{2} \left[\mathrm{E}_\lambda \, \lambda \right] \sum_i (y_i - \mu)^2 \right) \\ &= \mathrm{E}_\mu \left(\log q(\mu) + \frac{1}{2} A \sum_i (y_i - \mu)^2 \right) \end{aligned}$$

- the quantity A is known, since $\mathrm{E}_\lambda \, \lambda = \int \lambda q(\lambda) d\lambda$

- we can write

$$\begin{aligned} E_\mu \left(\log q(\mu) + \frac{1}{2} A \sum_i (y_i - \mu)^2 \right) &= \mathrm{E}_\mu \log \frac{q(\mu)}{p'(\mu)} \\ &= D(q(\mu) \| p'(\mu)) \end{aligned}$$

- straightforward calculation gives

$$p'(\mu) = N(\mu | \overline{y}, \lambda_1^{-1}), \;\; \lambda_1 = An$$

- setting $q(\mu) = p'(\mu)$ minimizes $D(q \| p')$ and thus minimizes the original KL-divergence
- first step has been completed and $q(\mu)$ is known

- next, find $q(\lambda)$ by minimizing the KL-divergence

$$D(q\|p) = \mathrm{E}_\lambda \, \mathrm{E}_\mu \log \frac{q(\lambda)}{p(y|\mu,\lambda)p(\lambda)}$$
$$= \mathrm{E}_\lambda \left(\log \frac{q(\lambda)}{p(\lambda)} + \mathrm{E}_\mu \left[-\log p(y|\mu,\lambda) \right] \right)$$

- the expectation over $q(\mu)$ gives a function of λ:

$$\mathrm{E}_\mu \left[-\log p(y|\mu,\lambda) \right] = -\frac{n}{2} \log \lambda + \frac{1}{2} \lambda (n\lambda_1^{-1} + S)$$
$$= F(\lambda)$$
$$S = \sum_i (y_i - \overline{y})^2$$

- then the KL-divergence is

$$D(q\|p) = \mathrm{E}_\lambda \left(F(\lambda) + \log \frac{q(\lambda)}{p(\lambda)} \right)$$
$$= \mathrm{E}_\lambda \log \frac{q(\lambda)}{p'(\lambda)} = D(q\|p')$$

- this results in a distribution

$$\log p'(\lambda) = -F(\lambda) + \log p(\lambda)$$
$$= \frac{n}{2} \log \lambda - \frac{1}{2} \lambda (n\lambda_1^{-1} + S) + \log p(\lambda) \implies$$
$$p'(\lambda) = \lambda^{n/2-1} \exp\left(-\frac{\lambda}{2}(n\lambda_1^{-1} + S)\right)$$

- the Gamma distribution $\Gamma(a, b)$ has density

$$\Gamma(\theta|a,b) \propto \theta^{a-1} \exp(-b\theta)$$

- therefore $q(\lambda) = \Gamma(\frac{n}{2}, \frac{1}{2}(n\lambda_1^{-1} + S))$

- each step started from an assumption that the other distribution is known
- since in the beginning neither distribution is known, these steps should be iterated several times
- the parametric form of each distribution is obtained in the first iteration but the parameter values can change
- demo: `freeform.R`