Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Part XVIII

Approximate Inference

Inference and Statistics

- Inference : Drawing conclusions about a population from a random sample drawn from it, or, more generally, about a random process from its observed behavior during a finite period of time.
- Descriptive Statistics describes the main features of a data set in quantitative terms.
- Inductive Statistics : Hypothesis testing.
- Fitting a model to some observed data (finding the parameters of the model) is inference.

Christfried Webers
NICTA
The Australian National University

Need for Approximate Inference

- Central task in the application of probabilistic models is the evaluation of the posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$ of the latent variables $\mathbf{Z}$ given the observed variables $\mathbf{X}$.
- This may be infeasible because
  - Posterior distribution has a too complex form for which expectations are not tractable.
  - Integration (for continuous variables) may not have a closed form solution.
  - Numerical integration (continuous variables) or sums over all possible configuration (discrete variables). Impossible if too many variables.
- Need approximations.

Approximation Schemes

- Stochastic Approximation
  - Produce exact results if given enough computational resources.
  - However, sampling methods can be computational demanding.
  - Example Markov Chain Monte Carlo (next lecture).
- Deterministic Approximation:
  - Based on analytical approximation of the posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$.
  - Example: posterior is assumed to factorise in a particular way.
  - However, these simplifying assumptions can never produce exact results for the original problem.

NICTA
The Australian National University

Variational Optimisation

- Originates from the calculus of variations (Euler, Lagrange, and others)
- Standard calculus uses a function to map one value to another value

$$y = f(x) \qquad\qquad x \xmapsto{f} y$$

- A functional maps a function to a value.
  Example: Entropy $H[p]$

$$H[p] = -\int p(x) \ln p(x) dx$$

- Functional derivative: How does the value change for infinitesimal changes of the input function.
- A large number of problems have the form: Find a function which optimises (maximises/minimises) a functional.
- Approximate solutions can be found by restricting the set of functions over which the functional will be optimised.

Calculus of Variation

- Given a functional of the form

$$F[y] = \int_a^b G(y(x), y'(x), x)dx$$

with $f(a) = f_a$ and $f(b) = f_b$ fixed, find a function $y(\cdot)$ which optimises the functional under the given constraints.

- Consider an infinitesimal change of the function $y(\cdot)$ by $\eta(x)$

$$\begin{aligned} F[y + \epsilon\eta] &= F[y] + \epsilon \int_a^b \left\{ \frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x) \right\} dx + O(\epsilon^2) \\ &= F[y] + \epsilon \int_a^b \left\{ \frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) \right\} \eta(x)dx + O(\epsilon^2) \end{aligned}$$

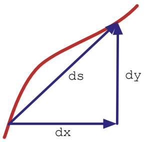
- Euler-Lagrange equations

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0$$

Calculus of Variation - Shortest path

- Shortest path between two points in $\mathbb{R}^N$ or on a manifold.
- In $\mathbb{R}^N$, the path length of a curve $y(x)$ is

$$ds = \sqrt{dx^2 + dy^2} = dx\sqrt{1 + (dy/dx)^2}$$



$$L = \int_a^b \sqrt{1 + y'(x)^2}\, dx$$

and therefore $G(y(x), y'(x), x) = \sqrt{1 + y'(x)^2}$

- Using the Euler-Lagrange equations

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0$$

- we get

$$\frac{d}{dx}\sqrt{1 + y'(x)^2} = 0$$

and therefore $y'(x) =$ constant : a straight line.

Variational Optimisation applied to Inference

- Variational optimisation is exact, there is no approximation involved.
- But it can be used to find approximate solutions to a given problem by restricting the class of functions.
- Examples: consider only quadratic functions, or linear combinations of fixed basis functions with variable coefficients.
- Probabilistic inference: assume the probability functions factorise in a specific way.

NICTA
The Australian National University

Variational Optimisation applied to Inference

- From the lecture on EM

$$
\begin{aligned}
\ln p(\mathbf{X}) &= \mathcal{L}(q) + \mathrm{KL}(q \| p) \\
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \, \mathrm{d}\mathbf{Z} \\
\mathrm{KL}(q \| p) &= -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} \mid \mathbf{X})}{q(\mathbf{Z})} \right\} \, \mathrm{d}\mathbf{Z}.
\end{aligned}
$$

- Here $\mathbf{Z}$ contains
  - latent variables (as before), and
  - parameters $\boldsymbol{\theta}$ which are now assumed to be stochastic (prior distribution over this variables provided).
- Given is the joint distribution $p(\mathbf{X}, \mathbf{Z})$.
- Find the posterior $p(\mathbf{Z} \mid \mathbf{X})$ and the model evidence $p(\mathbf{X})$.

Variational Optimisation applied to Inference

- From the lecture on EM

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

- Maximise the lower bound of $\mathcal{L}(q)$ by optimisation.
- If all probability distributions $p(\mathbf{Z})$ are allowed this will be achieved for $\mathrm{KL}(q\|p) = 0$, or $q(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X})$.
- Assuming the true posterior is intractable, consider a restricted family $q(\mathbf{Z})$ and find a member of this family which minimises the KL divergence.

Factorised Distributions

- Partition $\mathbf{Z}$ into disjoint sets $\mathbf{Z}_i$ where $i = 1, \dots, M$.
- Assume $q(\mathbf{Z})$ factorises as

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

- No restriction on the functional form of the individual $q_i(\mathbf{Z}_i)$.
- This factorisation for variational inference corresponds to Mean Field Theory in physics, replacing the interaction of n particles with a model of one particle and a mean field (created by all the other $n-1$ particles).

Factorised Distributions

- Insert $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$ into

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \, \mathrm{d}\mathbf{Z}$$

- to get (using $(q_j = q_j(\mathbf{Z}_j))$)

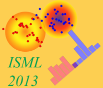
$$\begin{aligned} \mathcal{L}(q) &= \int \prod_{i=1}^{M} q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \, \mathrm{d}\mathbf{Z} \\ &= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j - \text{const} \end{aligned}$$

where

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) - \text{const} = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i = \mathbb{E}_{i \neq j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right]$$

- $\mathbb{E}_{i \neq j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right]$ denotes an expectation with respect to the q distributions over all variables $\mathbf{z}_i$ for $i \neq j$.

Factorised Distributions



- Keep $\{q_{i \neq j}\}$ fixed. Only q_j can vary.

$$\mathcal{L}(q) = \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j - \text{const}$$

- But $\mathcal{L}(q) + \text{const}$ can now be written as a negative Kullback-Leibler divergence

$$\mathcal{L}(q) + \text{const} = \int q_j \ln \left\{ \frac{\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)}{\ln q_j} \right\} = -\mathrm{KL}(q_j \| \widetilde{p}(\mathbf{X}, \mathbf{Z}_j))$$

- Maximising $\mathcal{L}(q)$ is equivalent to minimising $\mathrm{KL}(q_j \| \widetilde{p}(\mathbf{X}, \mathbf{Z}_j))$.

Factorised Distributions

- Optimal solution for

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right] - const \tag{2}$$

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right])}{\int \exp(\mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right])\, \mathrm{d}\mathbf{Z}_j} \tag{3}$$

- The log of the optimal solution for factor q_j is obtained by considering the log of the joint distribution over all hidden and visible variables and taking expectations w.r.t. all other factors $\{q_i\}$ for $i \neq j$.
- Don't use (3), better work with (2) and normalise (when required).

Factorised Distributions - Algorithm

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right] - const$$

1. Initialise all factors.
2. Cycle through the factors and replace each with the revised estimate evaluated using the current estimate.
3. Convergence is guaranteed because the bound is convex w.r.t each of the factors.

Exponential Family

- The exponential family of distributions over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\mathbf{x} \,|\, \boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\}$$

  where $\mathbf{x}$ may be scalar or vector, and may be discrete or continuous.
- Natural parameter $\boldsymbol{\eta}$
- And $\mathbf{u}$ is some function of $\mathbf{x}$.
- The function $g(\boldsymbol{\eta})$ can be interpreted as the coefficient ensuring normalisation

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \, \mathrm{d}\mathbf{x} = 1$$

- Other form with $g(\boldsymbol{\eta}) = \exp\{-G(\boldsymbol{\eta})\}$ we get

$$p(\mathbf{x} \,|\, \boldsymbol{\eta}) = h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) - G(\boldsymbol{\eta})\right\}$$

Example - Normal Distribution

- Normal distribution with mean μ and standard deviation σ

$$
\begin{aligned}
\mathcal{N}(x \mid \mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \\
\boldsymbol{\eta} &= \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)^T \\
h(x) &= \frac{1}{\sqrt{2\pi}} \\
\mathbf{u}(x) &= \left(x, x^2 \right)^T \\
g(\boldsymbol{\eta}) &= \sqrt{-2\eta_2} \exp\left(\frac{\eta_1^2}{4\eta_2} \right) \\
G(\boldsymbol{\eta}) &= -\frac{1}{2} \ln(-2\eta_2) - \frac{\eta_1^2}{4\eta_2}
\end{aligned}
$$

Exponential Family - Properties

- Differentation of $-\ln g(\boldsymbol{\eta})$ provides the moments

$$\frac{d}{d\boldsymbol{\eta}} - \ln g(\boldsymbol{\eta}) = \mathbb{E}\left[\mathbf{u}(\mathbf{x})\right]$$

$$\frac{d^2}{d\boldsymbol{\eta}^2} - \ln g(\boldsymbol{\eta}) = \operatorname{cov}[\mathbf{u}(\mathbf{x})]$$

$$\cdots$$

Prove using $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\}\, \mathrm{d}\mathbf{x} = 1$:

$$\begin{aligned}
\frac{d}{d\boldsymbol{\eta}} - \ln g(\boldsymbol{\eta}) &= \frac{d}{d\boldsymbol{\eta}} - \ln \left(\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\}\, \mathrm{d}\mathbf{x} \right)^{-1} \\
&= -\left(\int h(\mathbf{x}) e^{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})}\, \mathrm{d}\mathbf{x} \right) \frac{d}{d\boldsymbol{\eta}} \left(\int h(\mathbf{x}) e^{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})}\, \mathrm{d}\mathbf{x} \right)^{-1} \\
&= \left(\int h(\mathbf{x}) e^{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})}\, \mathrm{d}\mathbf{x} \right)^{-1} \frac{d}{d\boldsymbol{\eta}} \int h(\mathbf{x}) e^{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})}\, \mathrm{d}\mathbf{x} \\
&= \left(\int h(\mathbf{x}) e^{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})}\, \mathrm{d}\mathbf{x} \right)^{-1} \int \mathbf{u}(\mathbf{x})\, h(\mathbf{x}) e^{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})}\, \mathrm{d}\mathbf{x}
\end{aligned}$$

Exponential Family - Properties

- Differentation of $-\ln g(\boldsymbol{\eta})$ provides the moments

$$\frac{d}{d\boldsymbol{\eta}} - \ln g(\boldsymbol{\eta}) = \mathbb{E}\left[\mathbf{u}(\mathbf{x})\right]$$
$$\frac{d^2}{d\boldsymbol{\eta}^2} - \ln g(\boldsymbol{\eta}) = \operatorname{cov}[\mathbf{u}(\mathbf{x})]$$
$$\dots$$

or using $G(\boldsymbol{\eta})$ defined by $g(\boldsymbol{\eta}) = \exp\{-G(\boldsymbol{\eta})\}$,

$$\frac{d}{d\boldsymbol{\eta}} G(\boldsymbol{\eta}) = \mathbb{E}\left[\mathbf{u}(\mathbf{x})\right]$$
$$\frac{d^2}{d\boldsymbol{\eta}^2} G(\boldsymbol{\eta}) = \operatorname{cov}[\mathbf{u}(\mathbf{x})]$$
$$\dots$$

where

$$p(\mathbf{x} \,|\, \boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} = h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) - G(\boldsymbol{\eta})\right\}$$

Expectation Propagation

- Given a joint distribution over observed data $\mathcal{D}$ and stochastic variables $\boldsymbol{\theta}$ in the form of product of factors

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$$

- Goal: Approximate the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ by a distribution of the form

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \widetilde{f}_i(\boldsymbol{\theta}).$$

- Also the model evidence $p(\mathcal{D})$.

Expectation Propagation with Exponential Family

1. Initialise all approximating factors $\widetilde{f}_i(\boldsymbol{\theta})$.
2. Initialise the posterior approximation by setting

$$q(\boldsymbol{\theta}) \propto \prod_i \widetilde{f}_i(\boldsymbol{\theta})$$

3. Choose a factor $\widetilde{f}_j(\boldsymbol{\theta})$ to refine.
4. Remove $\widetilde{f}_j(\boldsymbol{\theta})$ from the posterior by division

$$q^{\setminus j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\widetilde{f}_j(\boldsymbol{\theta})}$$

5. Evaluate the new posterior by setting the sufficient statistics (moments) of q^{new} equal to those of $q^{\setminus j}\widetilde{f}_j(\boldsymbol{\theta})$ including evaluation of the normalisation constant

$$Z_j = \int q^{\setminus j}(\boldsymbol{\theta})\widetilde{f}_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

6. Evaluate and store the new factor, and goto 3.

$$\widetilde{f}_j(\boldsymbol{\theta}) = Z_j \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\setminus j}(\boldsymbol{\theta})}$$

Expectation Propagation

ISML 2013

- Finally, approximate the model evidence

$$p(\mathcal{D}) \approx \int \prod_i \widetilde{f}_i(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

- Expectation propagation is not guaranteed to converge.