NPFL108 – Bayesian inference

Approximate Inference

Laplace approximation

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek

Version: 21/03/2014
Outline

• Laplace approximation
• Probit regression model
The Laplace Approximation

- The **simplest deterministic** method for approximate inference

- Restricted to models in which the variables of interest are **continuous**

- The factors for the continuous random variables will generally be some continuous parametric functions
The Laplace Approximation: Univariate case 1

- The Laplace approximation will find a Gaussian approximation to the conditional distribution of a set of continuous variables.
- We are interested in approximating posteriors.
- Consider a single scalar variable $w$:

$$p(w|D) = \frac{1}{Z} p(w, D) = \frac{1}{Z} p(D|w)p(w) = \frac{1}{Z} f(w)$$

- $D$ are observed variables, therefore fixed and can be omitted.
- $Z$ is a normalisation constant:

$$Z = \int p(w, D)dw = \int f(w)dw$$

- We want to find $w_0$ and $A$ such that

$$p(w|D) \approx N(w; w_0, A^{-1})$$
The Laplace Approximation: Univariate case 2

- First, **find a mode** (i.e. **local** maximum $w_0$) of $p(w|D)$

\[
\frac{df(w)}{dw} = 0
\]

$=> w_0$

- Any algorithm can be used
  - including numerical solution

- **We do not work with** $p(w|D)$ because **we do not know** $Z$!
  - We do not need it to find maximum!

- Instead we work with $f(w)$ which is typically easily available.

\[
f(w) = \text{likelihood} \times \text{prior}
\]
Second, compute a truncated Taylor expansion of $\log f(w)$ centre at the mode

$$\log f(w) \approx \log f(w_0) + \frac{1}{2} A(w - w_0)^2$$

where

$$A = -\frac{d^2}{dw^2} \log f(w); w = w_0$$

Taking the exponential:

$$f(w) \approx f(w_0) \exp \left\{ \frac{1}{2} A(w - w_0)^2 \right\}$$

One can see that this looks like a normal distribution

$$p(w|D) \approx N(w; w_0, A^{-1}) = \frac{1}{\sqrt{2\pi A^{-1}}} \exp \left\{ \frac{(w - w_0)^2}{2A^{-1}} \right\}$$
The Laplace Approximation: Multi-variate Case

- The same principle can be applied to approximate an M-dimensional distribution

\[
\log f(w) \approx \log f(w_0) + \frac{1}{2}(w - w_0)^T A (w - w_0)
\]

\[
A = -\nabla \nabla \log f(w); \ w = w_0
\]

\[
f(w) \approx f(w_0) \exp \left\{ \frac{1}{2}(w - w_0)^T A (w - w_0) \right\}
\]

- The approximation has mean of \( w_0 \) and covariance matrix \( A^{-1} \)

\[
p(w|D) \approx N(w; w_0, A^{-1})
\]
The Laplace Approximation: example

- The Gaussian approximation will only be defined if $A$ is positive semidefinite, i.e., $w_0$ must be a local maximum not a minimum or a saddle point.
Probit regression model

- Similar to logistic regression
- Useful for binary classification
Probit regression: graphical model

\[ y_i = \begin{cases} 
1 & \text{if } \mathbf{w}^T \mathbf{x}_i + \epsilon_i \geq 0 \\
-1 & \text{if } \mathbf{w}^T \mathbf{x}_i + \epsilon_i < 0
\end{cases} \]

\[ \mathbf{w} \sim N(0, \mathbf{I}_\alpha) \]

\[ \epsilon_i \sim N(0, \sigma^2) \]

\[ p(y_i | x_i; \mathbf{w}) = \Phi(y_i \mathbf{w}^T \mathbf{x}_i; 0, \sigma^2) \]

\[ p(y, \mathbf{w} | \mathbf{x}) = p(y | \mathbf{x}; \mathbf{w})p(\mathbf{w}) \]

\[ p(y | \mathbf{x}; \mathbf{w})p(\mathbf{w}) = \prod_{i=1}^{N} p(y_i | x_i; \mathbf{w})p(\mathbf{w}) = \prod_{i=1}^{N} \Phi(y_i \mathbf{w}^T \mathbf{x}_i; 0, \sigma^2)N(\mathbf{w}; 0, \mathbf{I}_\alpha) \]

- **\( \mathbf{w} \)** are our parameters
- **\( y_i, x_i \)** are our observations – data \( \mathbf{D} \)

\( \sigma^2 \)

\( N \)

\( \Phi \)

Probit function
Probit regression model

- For the sake of completeness, **probit function**

\[
\Phi(a; \mu, \sigma^2) = \int_{-\infty}^{a} N(a; \mu, \sigma^2) \, da
\]

- We want to make inference of \( w \) given some observed labels \( y \) and \( x \)

\[
p(w|y, x; \alpha, \sigma^2) = \frac{p(y|x, w; \sigma^2)p(w; \alpha)}{p(y|x)}
\]
For simplicity, we consider that $\sigma^2 = 1$ and that $\alpha = 1$.

The posterior distribution is:

$$p(w|y, x) \propto p(y|x; w)p(w)$$

Recall 1

$$p(y|x; w)p(w) = \prod_{i=1}^{N} p(y_i|x_i; w)p(w) = \prod_{i=1}^{N} \Phi(y_iw^t x_i; 0, 1)N(w; 0, I)$$

Recall 2

$$f(w) = p(y|x; w)p(w)$$

$$f(w) = \prod_{i=1}^{N} \Phi(y_iw^t x_i; 0, 1)N(w; 0, I)$$
The Laplace Approximation: Probit Regression 2

- Using some numerical optimisation algorithm
  - find $w_0$ – a local maximum of

$$f(w) = \prod_{i=1}^{N} \Phi(y_i w^t x_i; 0, 1) N(w; 0, I)$$

- Perform Taylor expansion of

$$\log f(w) = \log p(y|x; w) + \log p(w)$$

$$\log f(w) = \sum_{i=1}^{N} \log \Phi(y_i w^t x_i; 0, 1) - \frac{1}{2} w^T w - \frac{1}{2} \log 2\pi$$
Let $w_0$ be a maximum of $f(w)$

Computing the negative Hessian at $w_0$ of $\log f(w)$

$$A = -\nabla \nabla \log f(w) = \sum_{i=1}^{N} [v_i (y_i w_0^T x_i + v_i) x_i x_i^T] + I$$

$$v_i = \frac{N(y_i w_0^T x_i; 0, 1)}{\Phi(y_i w_0^T x_i)}$$

Approximation of $p(w|y, x)$ is

$$p(w; w_0, A^{-1}) = N(w; w_0, A^{-1})$$
Predictive distribution

- We also want to compute a predictive distribution for new unlabelled instances

\[
p(y_{new} | x_{new}, y, x, \alpha, \sigma^2) = \int p(y_{new} | x_{new}, w)p(w | y, x, \alpha, \sigma^2) dw
\]
The Laplace Approximation: Probit Regression 4

- Thanks to probit model and the Laplace Approximation
  - It is possible to compute an approximate predictive distribution

\[
p(y_{new}|x_{new}, y, x, \alpha, \sigma^2) = \int p(y_{new}|x_{new}, w)N(w|w_0, A^{-1})dw
\]

\[
= \int \Phi(y_{new}w^T x_{new})N(w|w_0, A^{-1})dw
\]

Hurray! We know how to compute the integral.

\[
= \Phi \left( \frac{y_{new}w^T x_{new}}{\sqrt{x_{new}^T x_{new} + A^{-1} x_{new} + 1}} \right)
\]
• Uncertainty is high near the decision boundary and progressively decreases as we move away from it.
• Uncertainty is significantly larger in regions where there is no data.
The Laplace Approximation: Considerations

- The mode of log f can be found using a numerical optimization method.
- The Hessian can be approximated by differences.
- Many distributions can be multi-modal, what leads to many different Laplace approximations, depending on the mode.
- In many cases, the posterior distribution of z will converge to a Gaussian as the number of observations (evidence) increases.
- Only applicable on real variables.
- Only focuses around the mode and can fail to capture global properties.
- No need to know Z.
Thank you!

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek