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.

$$f(w) = \text{likelihood} \times \text{prior}$$

The Laplace Approximation: Univariate case 3

- 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:

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

$\frac{1}{\sqrt{2\pi A^{-1}}}$ variance mean

- 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

$$\log f(\mathbf{w}) \approx \log f(\mathbf{w}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T \mathbf{A} (\mathbf{w} - \mathbf{w}_0)$$

$$\mathbf{A} = -\nabla\nabla \log f(\mathbf{w}); \mathbf{w} = \mathbf{w}_0$$

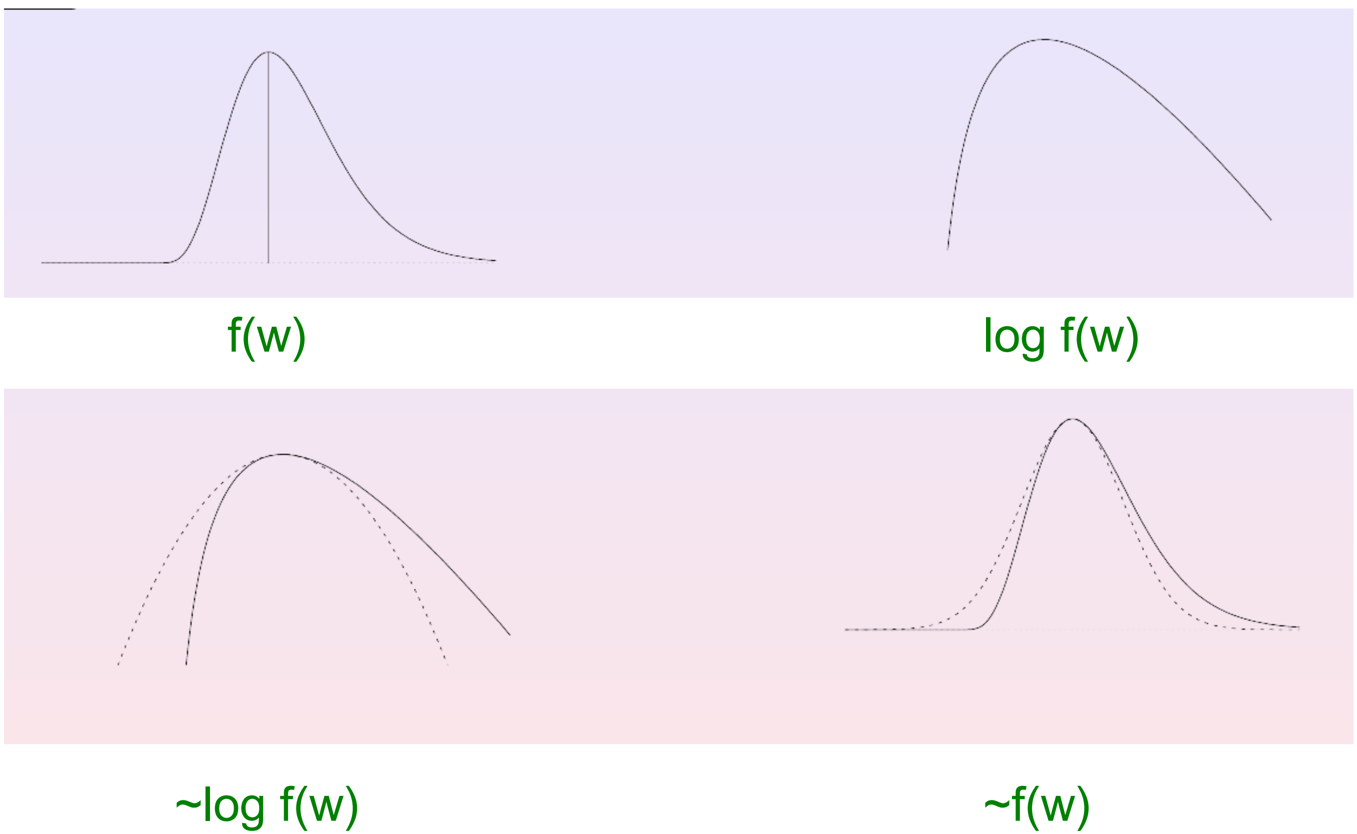
$$f(\mathbf{w}) \approx f(\mathbf{w}_0) \exp\left\{\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T \mathbf{A} (\mathbf{w} - \mathbf{w}_0)\right\}$$

- The approximation has mean of $\mathbf{w}_0$ and covariance matrix $\mathbf{A}^{-1}$

$$p(\mathbf{w}|D) \approx N(\mathbf{w}; \mathbf{w}_0, \mathbf{A}^{-1})$$

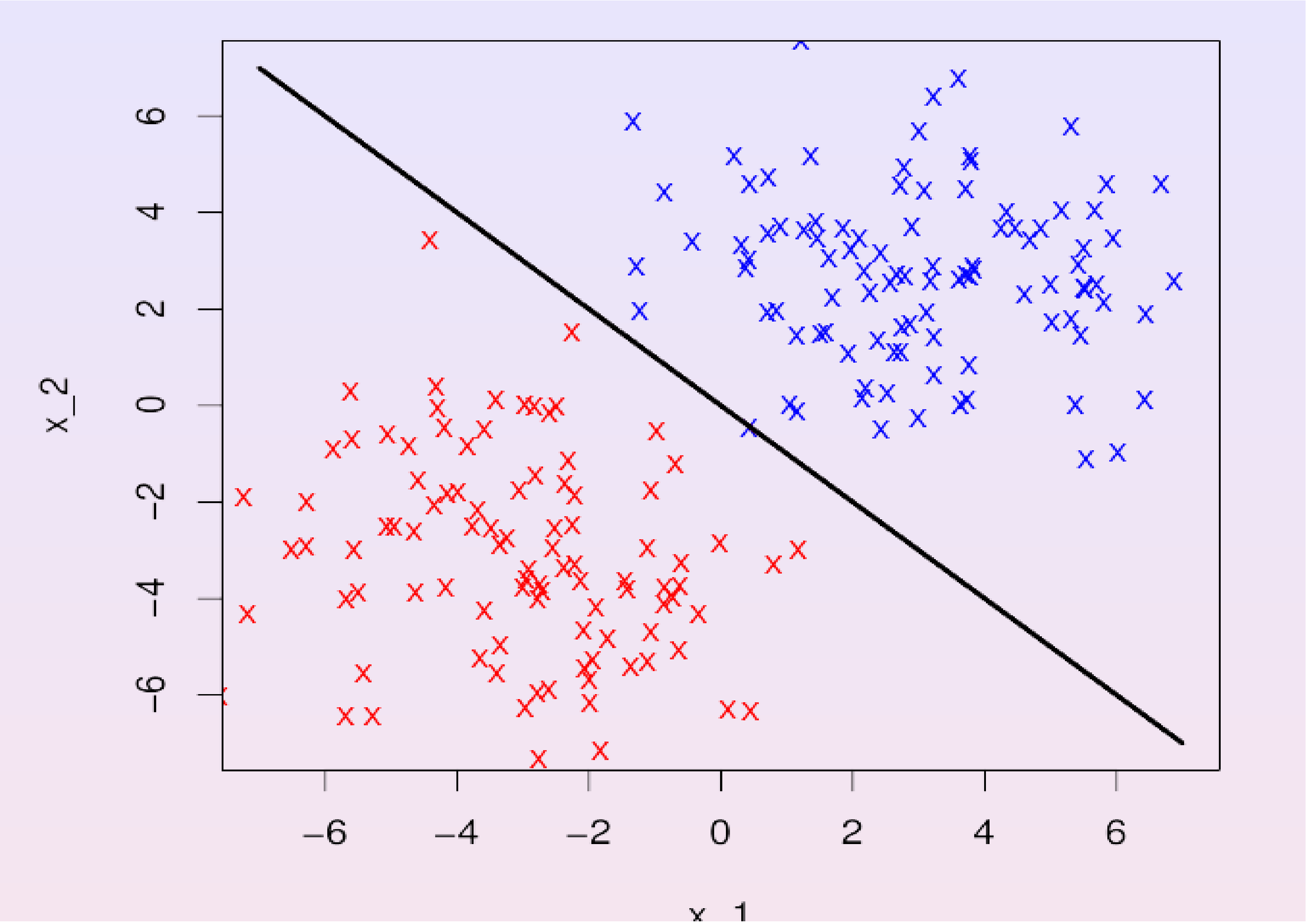
The Laplace Approximation: example

- 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.

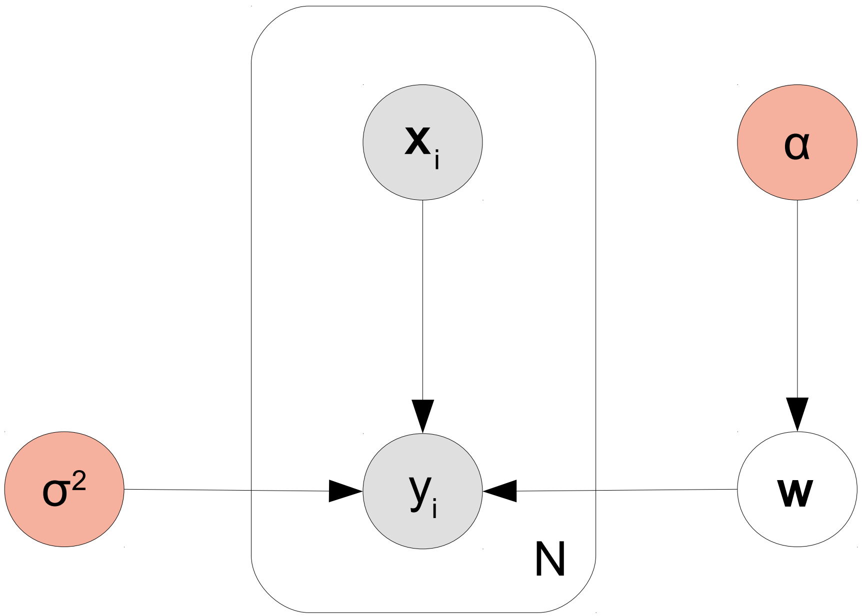


Probit regression model

- Similar to logistic regression
- Useful for binary classification



Probit regression: graphical model



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

$$\mathbf{w} \sim N(\mathbf{0}, \mathbf{I}\alpha)$$
$$\epsilon_i \sim N(0, \sigma^2)$$

$$p(y_i | \mathbf{x}_i; \mathbf{w}) = \underbrace{\Phi}_{\text{Probit function}}(y_i \mathbf{w}^t \mathbf{x}_i; 0, \sigma^2)$$

- **w** are our parameters
- y_i,x_i are our observations – data **D**

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

$$p(\mathbf{y} | \mathbf{x}; \mathbf{w}) p(\mathbf{w}) = \prod_{i=1}^{N} p(y_i | \mathbf{x}_i; \mathbf{w}) p(\mathbf{w}) = \prod_{i=1}^{N} \Phi(y_i \mathbf{w}^t \mathbf{x}_i; 0, \sigma^2) N(\mathbf{w}; \mathbf{0}, \mathbf{I}\alpha)$$

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**

$$p(\mathbf{w}|\mathbf{y},\mathbf{x};\alpha,\sigma^2) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{w};\sigma^2)p(\mathbf{w};\alpha)}{p(\mathbf{y}|\mathbf{x})}$$

The Laplace Approximation: Probit Regression 1

- For simplicity, we consider that $\sigma^2 = 1$ and that $\alpha = 1$.
- The posterior distribution is:

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

- Recall 1

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

- Recall 2

$$f(\mathbf{w}) = p(\mathbf{y}|\mathbf{x}; \mathbf{w})p(\mathbf{w})$$

$$\Downarrow$$

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

The Laplace Approximation: Probit Regression 2

- Using some numerical optimisation algorithm
  - find $\mathbf{w}_0$ – a local maximum of

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

- Perform Taylor expansion of

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

$$\log f(\mathbf{w}) = \sum_{i=1}^{N} \log \Phi(y_i \mathbf{w}^t \mathbf{x}_i; 0, 1) - \frac{1}{2}\mathbf{w}^T \mathbf{w} - \frac{1}{2} \log 2\pi$$

The Laplace Approximation: Probit Regression 3

- Let $\mathbf{w}_0$ be a maximum of f($\mathbf{w}$)
- Computing the negative Hessian at w_0 of log f(w)

$$\mathbf{A} = -\nabla\nabla \log f(\mathbf{w}) = \sum_{i=1}^{N} [v_i(y_i \mathbf{w}_0^T \mathbf{x}_i + v_i)\mathbf{x}_i \mathbf{x}_i^T] + \mathbf{I}$$

$$v_i = \frac{N(y_i \mathbf{w}_0^T \mathbf{x}_i; 0, 1)}{\Phi(y_i \mathbf{w}_0^T \mathbf{x}_i)}$$

- Approximation of $p(\mathbf{w}|\mathbf{y}, \mathbf{x})$ is

$$p(\mathbf{w}; \mathbf{w}_0, \mathbf{A}^{-1}) = N(\mathbf{w}; \mathbf{w}_0, \mathbf{A}^{-1})$$

Predictive distribution

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

$$p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{x}, \alpha, \sigma^2) = \int p(y_{new}|\mathbf{x}_{new}, \mathbf{w})p(\mathbf{w}|\mathbf{y}, \mathbf{x}, \alpha, \sigma^2)d\mathbf{w}$$

The Laplace Approximation: Probit Regression 4

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

$$p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{x}, \alpha, \sigma^2) = \int p(y_{new}|\mathbf{x}_{new}, \mathbf{w}) N(\mathbf{w}|\mathbf{w}_0, \mathbf{A}^{-1}) d\mathbf{w}$$

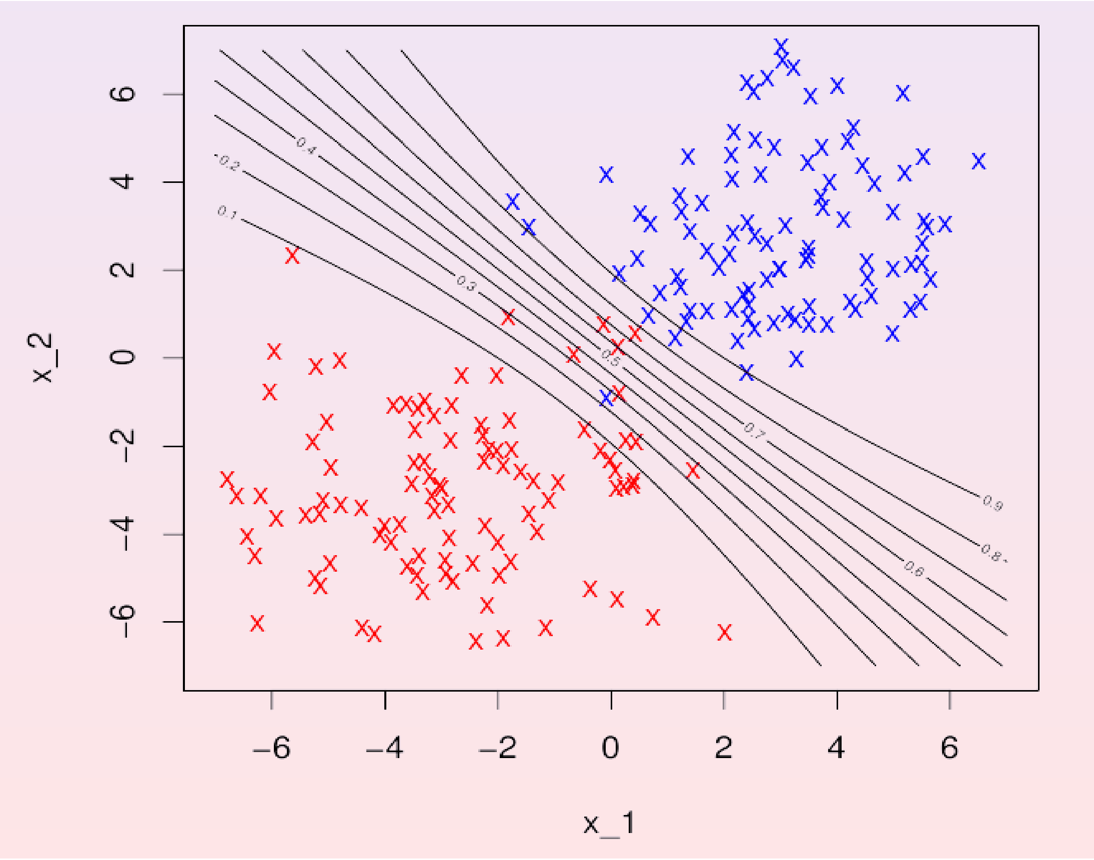
$$= \int \Phi(y_{new}\mathbf{w}^T\mathbf{x}_{new}) N(\mathbf{w}|\mathbf{w}_0, \mathbf{A}^{-1}) d\mathbf{w}$$

Hurray! We know how to compute the integral.

$$= \Phi\left(\frac{y_{new}\mathbf{w}^T\mathbf{x}_{new}}{\sqrt{\mathbf{x}_{new}^T + \mathbf{A}^{-1}\mathbf{x}_{new} + 1}}\right)$$

The Laplace Approximation: Probit Regression 5

- 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.



Задача 1

По условию задачи:

$$p(\mathbf{X}, \mathbf{y}, \mathbf{w}|\mathbf{A}) = \prod_i N(\mathbf{x}_i|\mathbf{0}, \sigma^2\mathbf{I}_n)N(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \prod_j p(y_j|\mathbf{x}_j, \mathbf{w}), \tag{1.1}$$

где $p(y_j = 1|\mathbf{x}_j, \mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_j)}$

Для простоты запишем (1.1) в следующем общем виде:

$$p(\mathbf{X}, \mathbf{y}, \mathbf{w}|\mathbf{A}) = p(\mathbf{X})p(\mathbf{w}|\mathbf{A})p(\mathbf{y}|\mathbf{X}, \mathbf{w}). \tag{1.2}$$

а) По формуле Байеса:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{A}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A})}{\int_{\mathbf{w}'\in\mathbb{R}^n} p(\mathbf{y}|\mathbf{X}, \mathbf{w}')p(\mathbf{w}'|\mathbf{A})d\mathbf{w}'} = \frac{\mathcal{Q}(\mathbf{w})}{\int_{\mathbf{w}'\in\mathbb{R}^n} \mathcal{Q}(\mathbf{w}')}, \tag{1.3}$$

где введено обозначение $\mathcal{Q}(\mathbf{w}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A})$.

Выполним аппроксимацию Лапласа:

$$\log\mathcal{Q}(\mathbf{w}) \approx \log\mathcal{Q}(\mathbf{w}_{\text{MAP}}) + \cancel{\nabla\log\mathcal{Q}(\mathbf{w}_{\text{MAP}})} + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T}\nabla\nabla^\mathsf{T}\log\mathcal{Q}(\mathbf{w}_{\text{MAP}})(\mathbf{w} - \mathbf{w}_{\text{MAP}}) =$$

$$= \log\mathcal{Q}(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T}\mathbf{H}^{-1}(\mathbf{w} - \mathbf{w}_{\text{MAP}}), \tag{1.4}$$

где введено обозначение $\mathbf{H}^{-1} = -\nabla\nabla^\mathsf{T}\log\mathcal{Q}(\mathbf{w}_{\text{MAP}})$.

Для нашей задачи найдем $\mathbf{H}^{-1}$:

$$\mathbf{H}^{-1} = -\nabla\nabla^\mathsf{T}(-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}^{-1}\mathbf{w} - \mathbf{1}^\mathsf{T}\log(1 + \exp(-\mathbf{X}^\mathsf{T}\mathbf{w}))) =$$

$$= \mathbf{A}^{-1} + \nabla\nabla^\mathsf{T}\mathbf{1}^\mathsf{T}\log(1 + \exp(-\mathbf{X}^\mathsf{T}\mathbf{w})) =$$

$$= \mathbf{A}^{-1} + \sum_{i=1}^{m}\nabla\nabla^\mathsf{T}\log(1 + \exp(-\mathbf{x}_i^\mathsf{T}\mathbf{w})) =$$

$$= \mathbf{A}^{-1} + \sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\frac{\exp(-\mathbf{x}_i^\mathsf{T}\mathbf{w})}{1 + \exp(-\mathbf{x}_i^\mathsf{T}\mathbf{w})} - \mathbf{x}_i\mathbf{x}_i^\mathsf{T}\frac{\exp(-2\mathbf{x}_i^\mathsf{T}\mathbf{w})}{(1 + \exp(-\mathbf{x}_i^\mathsf{T}\mathbf{w}))^2}. \tag{1.5}$$

Тогда получаем:

$$\mathcal{Q}(\mathbf{w}) \approx \mathcal{Q}(\mathbf{w}_{\text{MAP}})\exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T}\mathbf{H}^{-1}(\mathbf{w} - \mathbf{w}_{\text{MAP}})\right). \tag{1.6}$$

Подставляя (1.6) в (1.3) получим:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{A}) \approx \frac{\mathcal{Q}(\mathbf{w}_{\text{MAP}})\exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T}\mathbf{H}^{-1}(\mathbf{w} - \mathbf{w}_{\text{MAP}})\right)}{\int_{\mathbf{w}'\in\mathbb{R}^n}\mathcal{Q}(\mathbf{w}_{\text{MAP}})\exp\left(-\frac{1}{2}(\mathbf{w}' - \mathbf{w}_{\text{MAP}})^\mathsf{T}\mathbf{H}^{-1}(\mathbf{w}' - \mathbf{w}_{\text{MAP}})\right)d\mathbf{w}'} =$$

$$= \frac{\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})\right)}{\int_{\mathbf{w}'\in\mathbb{R}^n}\exp\left(-\frac{1}{2}(\mathbf{w}'-\mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}'-\mathbf{w}_{\mathrm{MAP}})\right)d\mathbf{w}'} =$$

$$= N(\mathbf{w}_{\mathrm{MAP}}, \mathbf{H}). \tag{1.7}$$

Оценим $\mathbf{w}_{\mathrm{MAP}}$:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}\in\mathbb{R}^n} p(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}) = \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{-\log p(\mathbf{y}|\mathbf{X},\mathbf{w}) - \log p(\mathbf{w}|\mathbf{A})\}, \tag{1.8}$$

где

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_i \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}; \quad -\log p(\mathbf{w}|\mathbf{A}) = \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}; \quad \hat{\mathbf{p}} = \frac{1}{1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})}, \tag{1.9}$$

Подставляя (1.9) в (1.8) получаем:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{-\mathbf{y}^{\mathsf{T}}\log\hat{\mathbf{p}} - (1-\mathbf{y})^{\mathsf{T}}\log(1-\hat{\mathbf{p}}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}\} =$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{-\mathbf{y}^{\mathsf{T}}\log\frac{1}{1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})} - (\mathbf{1}-\mathbf{y})^{\mathsf{T}}\log\frac{\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})}{1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})} + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}\}\} =$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{(\mathbf{1}-\mathbf{y})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{w} + \mathbf{1}^{\mathsf{T}}\log(1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}\}, \tag{1.10}$$

где введя обозначения $\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}) = (\mathbf{1}-\mathbf{y})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{w} + \mathbf{1}^{\mathsf{T}}\log(1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}$ получим следующую оптимизационую задачу для нахождения $\mathbf{w}_{\mathrm{MAP}}$:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n}\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}). \tag{1.11}$$

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