



# Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group

Data61 | CSIRO

and

College of Engineering and Computer Science

The Australian National University

Canberra

February – June 2017

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

## Outlines

Overview

Introduction

Linear Algebra

Probability

Linear Regression 1

Linear Regression 2

Linear Classification 1

Linear Classification 2

Kernel Methods

Sparse Kernel Methods

Neural Networks 1

Neural Networks 2

Continuous Latent Variables

Autoencoders

Graphical Models 1

Graphical Models 2

Graphical Models 3

Sampling

Mixture Models and EM 1

Mixture Models and EM 2

Sequential Data 1

Sequential Data 2

Combining Models



# Part VIII

## *Linear Classification 2*

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*



In increasing order of complexity

- Find a discriminant function  $f(\mathbf{x})$  which maps each input directly onto a class label.
- Discriminative Models
  - 1 Solve the inference problem of determining the posterior class probabilities  $p(\mathcal{C}_k | \mathbf{x})$ .
  - 2 Use decision theory to assign each new  $\mathbf{x}$  to one of the classes.
- Generative Models
  - 1 Solve the inference problem of determining the class-conditional probabilities  $p(\mathbf{x} | \mathcal{C}_k)$ .
  - 2 Also, infer the prior class probabilities  $p(\mathcal{C}_k)$ .
  - 3 Use Bayes' theorem to find the posterior  $p(\mathcal{C}_k | \mathbf{x})$ .
  - 4 Alternatively, model the joint distribution  $p(\mathbf{x}, \mathcal{C}_k)$  directly.
  - 5 Use decision theory to assign each new  $\mathbf{x}$  to one of the classes.

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*



- Generative approach: model class-conditional densities  $p(\mathbf{x} | \mathcal{C}_k)$  and priors  $p(\mathcal{C}_k)$  to calculate the posterior probability for class  $\mathcal{C}_1$

$$\begin{aligned} p(\mathcal{C}_1 | \mathbf{x}) &= \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1 + \exp(-a(\mathbf{x}))} = \sigma(a(\mathbf{x})) \end{aligned}$$

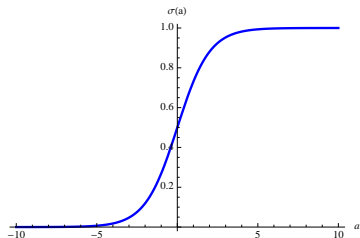
where  $a$  and the **logistic sigmoid** function  $\sigma(a)$  are given by

$$\begin{aligned} a(\mathbf{x}) &= \ln \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)} = \ln \frac{p(\mathbf{x}, \mathcal{C}_1)}{p(\mathbf{x}, \mathcal{C}_2)} \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned}$$

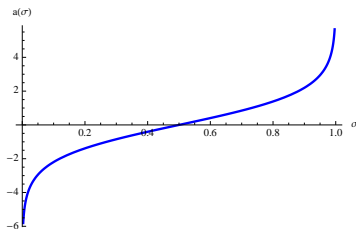
# Logistic Sigmoid



- The **logistic sigmoid** function  $\sigma(a) = \frac{1}{1+\exp(-a)}$
- "squashing function" because it maps the real axis into a finite interval  $(0, 1)$
- $\sigma(-a) = 1 - \sigma(a)$
- Derivative  $\frac{d}{da}\sigma(a) = \sigma(a)\sigma(-a) = \sigma(a)(1 - \sigma(a))$
- Inverse is called **logit** function  $a(\sigma) = \ln\left(\frac{\sigma}{1-\sigma}\right)$



Logistic Sigmoid  $\sigma(a)$



Logit  $a(\sigma)$

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*



- The **normalised exponential** is given by

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) p(C_k)}{\sum_j p(\mathbf{x} | C_j) p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

where

$$a_k = \ln(p(\mathbf{x} | C_k) p(C_k)).$$

- Also called **softmax function** as it is a smoothed version of the max function.
- Example: If  $a_k \gg a_j$  for all  $j \neq k$ , then  $p(C_k | \mathbf{x}) \simeq 1$ , and  $p(C_j | \mathbf{x}) \simeq 0$ .



- Assume class-conditional probabilities are Gaussian, all classes share the **same covariance**. What can we say about the posterior probabilities?

$$\begin{aligned} p(\mathbf{x} | \mathcal{C}_k) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right\} \\ &\quad \times \exp \left\{ \boldsymbol{\mu}_k^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k \right\} \end{aligned}$$

where we separated the quadratic term in  $\mathbf{x}$  and the linear term.



- For two classes

$$p(\mathcal{C}_1 | \mathbf{x}) = \sigma(a(\mathbf{x}))$$

- and  $a(\mathbf{x})$  is

$$\begin{aligned} a(\mathbf{x}) &= \ln \frac{p(\mathbf{x} | \mathcal{C}_1) p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_2) p(\mathcal{C}_2)} \\ &= \ln \frac{\exp \left\{ \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 \right\}}{\exp \left\{ \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \right\}} + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \end{aligned}$$

- Therefore

$$p(\mathcal{C}_1 | \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$$

where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \end{aligned}$$

Probabilistic Generative  
Models

Continuous Input

Discrete Features

Probabilistic  
Discriminative Models

Logistic Regression

Iterative Reweighted  
Least Squares

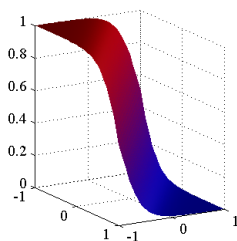
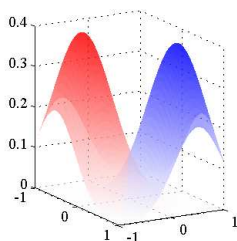
Laplace Approximation

Bayesian Logistic  
Regression





Class-conditional densities for two classes (left). Posterior probability  $p(C_1 | \mathbf{x})$  (right). Note the logistic sigmoid of a linear function of  $\mathbf{x}$ .



Probabilistic Generative  
Models

Continuous Input

Discrete Features

Probabilistic  
Discriminative Models

Logistic Regression

Iterative Reweighted  
Least Squares

Laplace Approximation

Bayesian Logistic  
Regression



- Use the normalised exponential

$$p(\mathcal{C}_k | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x} | \mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

where

$$a_k = \ln(p(\mathbf{x} | \mathcal{C}_k)p(\mathcal{C}_k)).$$

- to get a linear function of  $\mathbf{x}$

$$a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}.$$

where

$$\begin{aligned}\mathbf{w}_k &= \Sigma^{-1} \boldsymbol{\mu}_k \\ w_{k0} &= -\frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + p(\mathcal{C}_k).\end{aligned}$$

Probabilistic Generative  
Models

Continuous Input

Discrete Features

Probabilistic  
Discriminative Models

Logistic Regression

Iterative Reweighted  
Least Squares

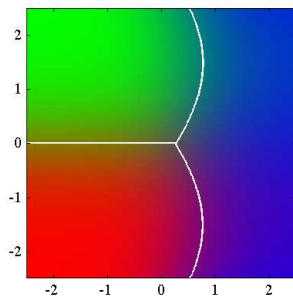
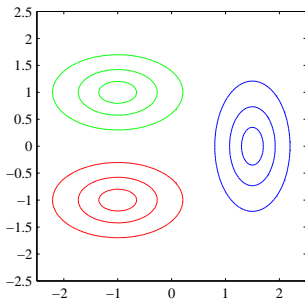
Laplace Approximation

Bayesian Logistic  
Regression

# General Case - $K$ Classes, Different Covariance



- If each class-conditional probability has a **different** covariance, the quadratic terms  $-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}$  do not longer cancel each other out.
- We get a **quadratic** discriminant.





- Given the functional form of the class-conditional densities  $p(\mathbf{x} | \mathcal{C}_k)$ , can we determine the parameters  $\mu$  and  $\Sigma$  ?

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*

# Maximum Likelihood Solution



- Given the functional form of the class-conditional densities  $p(\mathbf{x} | \mathcal{C}_k)$ , can we determine the parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  ?
- Not without data ;-)
- Given also a data set  $(\mathbf{x}_n, t_n)$  for  $n = 1, \dots, N$ . (Using the coding scheme where  $t_n = 1$  corresponds to class  $\mathcal{C}_1$  and  $t_n = 0$  denotes class  $\mathcal{C}_2$ .)
- Assume the class-conditional densities to be Gaussian with the same covariance, but different mean.
- Denote the prior probability  $p(\mathcal{C}_1) = \pi$ , and therefore  $p(\mathcal{C}_2) = 1 - \pi$ .
- Then

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n | \mathcal{C}_1) = \pi \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n | \mathcal{C}_2) = (1 - \pi) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$



# Maximum Likelihood Solution

- Thus the likelihood for the whole data set  $\mathbf{X}$  and  $\mathbf{t}$  is given by

$$p(\mathbf{t}, \mathbf{X} | \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^N [\pi \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} \times [(1 - \pi) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

- Maximise the log likelihood
- The term depending on  $\pi$  is

$$\sum_{n=1}^N (t_n \ln \pi + (1 - t_n) \ln(1 - \pi))$$

- which is maximal for

$$\pi = \frac{1}{N} \sum_{n=1}^N t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

where  $N_1$  is the number of data points in class  $\mathcal{C}_1$ .



- Similarly, we can maximise the log likelihood (and thereby the likelihood  $p(\mathbf{t}, \mathbf{X} | \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ ) depending on the mean  $\boldsymbol{\mu}_1$  or  $\boldsymbol{\mu}_2$ , and get

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^N t_n \mathbf{x}_n$$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^N (1 - t_n) \mathbf{x}_n$$

- For each class, this are the means of all input vectors assigned to this class.



- Finally, the log likelihood  $\ln p(\mathbf{t}, \mathbf{X} | \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$  can be maximised for the covariance  $\boldsymbol{\Sigma}$  resulting in

$$\boldsymbol{\Sigma} = \frac{N_1}{N} \mathbf{S}_1 + \frac{N_2}{N} \mathbf{S}_2$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Probabilistic Generative  
Models

Continuous Input

Discrete Features

Probabilistic  
Discriminative Models

Logistic Regression

Iterative Reweighted  
Least Squares

Laplace Approximation

Bayesian Logistic  
Regression





- Assume the input space consists of discrete features, in the simplest case  $x_i \in \{0, 1\}$ .
- For a  $D$ -dimensional input space, a general distribution would be represented by a table with  $2^D$  entries.
- Together with the normalisation constraint, this are  $2^D - 1$  independent variables.
- Grows exponentially with the number of features.
- The **Naive Bayes** assumption is that all features conditioned on the class  $\mathcal{C}_k$  are independent of each other.

$$p(\mathbf{x} | \mathcal{C}_k) = \prod_{i=1}^D \mu_{k_i}^{x_i} (1 - \mu_{k_i})^{1-x_i}$$



- With the naive Bayes

$$p(\mathbf{x} | \mathcal{C}_k) = \prod_{i=1}^D \mu_{k_i}^{x_i} (1 - \mu_{k_i})^{1-x_i}$$

- we can then again find the factors  $a_k$  in the normalised exponential

$$p(\mathcal{C}_k | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x} | \mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

- as a linear function of the  $x_i$

$$a_k(\mathbf{x}) = \sum_{i=1}^D \{x_i \ln \mu_{k_i} + (1 - x_i) \ln(1 - \mu_{k_i})\} + \ln p(\mathcal{C}_k).$$



In increasing order of complexity

- Find a discriminant function  $f(\mathbf{x})$  which maps each input directly onto a class label.
- **Discriminative Models**
  - 1 Solve the inference problem of determining the posterior class probabilities  $p(\mathcal{C}_k | \mathbf{x})$ .
  - 2 Use decision theory to assign each new  $\mathbf{x}$  to one of the classes.
- **Generative Models**
  - 1 Solve the inference problem of determining the class-conditional probabilities  $p(\mathbf{x} | \mathcal{C}_k)$ .
  - 2 Also, infer the prior class probabilities  $p(\mathcal{C}_k)$ .
  - 3 Use Bayes' theorem to find the posterior  $p(\mathcal{C}_k | \mathbf{x})$ .
  - 4 Alternatively, model the joint distribution  $p(\mathbf{x}, \mathcal{C}_k)$  directly.
  - 5 Use decision theory to assign each new  $\mathbf{x}$  to one of the classes.

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*



- Maximise a likelihood function defined through the conditional distribution  $p(\mathcal{C}_k | \mathbf{x})$  directly.
- **Discriminative** training
- Typically fewer parameters to be determined.
- As we learn the posterior  $p(\mathcal{C}_k | \mathbf{x})$  directly, prediction may be better than with a generative model where the class-conditional density assumptions  $p(\mathbf{x} | \mathcal{C}_k)$  poorly approximate the true distributions.
- But: discriminative models can not create synthetic data, as  $p(\mathbf{x})$  is not modelled.

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

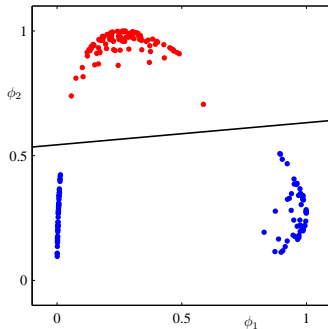
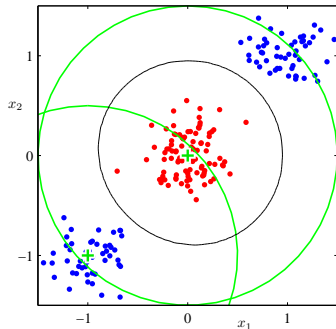
*Laplace Approximation*

*Bayesian Logistic  
Regression*



# Original Input versus Feature Space

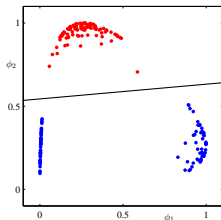
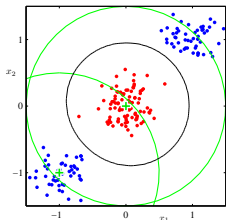
- Used direct input  $\mathbf{x}$  until now.
- All classification algorithms work also if we first apply a fixed nonlinear transformation of the inputs using a vector of basis functions  $\phi(\mathbf{x})$ .
- Example: Use two Gaussian basis functions centered at the green crosses in the input space.



# Original Input versus Feature Space



- Linear decision boundaries in the feature space correspond to nonlinear decision boundaries in the input space.
- Classes which are NOT linearly separable in the input space can become linearly separable in the feature space.
- BUT: If classes overlap in input space, they will also overlap in feature space.
- Nonlinear features  $\phi(\mathbf{x})$  can not remove the overlap; but they may increase it !



# Original Input versus Feature Space



- Fixed basis functions do not adapt to the data and therefore have important limitations (see discussion in Linear Regression).
- Understanding of more advanced algorithms becomes easier if we introduce the feature space now and use it instead of the original input space.
- Some applications use fixed features successfully by avoiding the limitations.
- We will therefore use  $\phi$  instead of  $\mathbf{x}$  from now on.

# Logistic Regression is Classification



- Two classes where the posterior of class  $\mathcal{C}_1$  is a logistic sigmoid  $\sigma(\cdot)$  acting on a linear function of the feature vector  $\phi$

$$p(\mathcal{C}_1 | \phi) = y(\phi) = \sigma(\mathbf{w}^T \phi)$$

- $p(\mathcal{C}_2 | \phi) = 1 - p(\mathcal{C}_1 | \phi)$
- Model dimension is equal to dimension of the feature space  $M$ .
- Compare this to fitting two Gaussians

$$\underbrace{2M}_{\text{means}} + \underbrace{M(M+1)/2}_{\text{shared covariance}} = M(M+5)/2$$

- For larger  $M$ , the logistic regression model has a clear advantage.



# Logistic Regression is Classification



- Determine the parameter via maximum likelihood for data  $(\phi_n, t_n)$ ,  $n = 1, \dots, N$ , where  $\phi_n = \phi(\mathbf{x}_n)$ . The class membership is coded as  $t_n \in \{0, 1\}$ .
- Likelihood function

$$p(\mathbf{t} | \mathbf{w}) = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n}$$

where  $y_n = p(C_1 | \phi_n)$ .

- Error function : negative log likelihood resulting in the **cross-entropy** error function

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}) = -\sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

# Logistic Regression is Classification



- Error function (cross-entropy error )

$$E(\mathbf{w}) = - \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

- $y_n = p(\mathcal{C}_1 | \phi_n) = \sigma(\mathbf{w}^T \phi_n)$
- Gradient of the error function (using  $\frac{d\sigma}{da} = \sigma(1 - \sigma)$  )

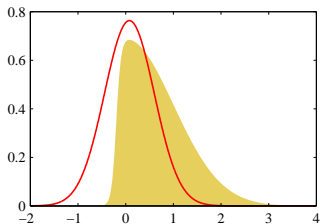
$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n$$

- gradient does not contain any sigmoid function
- for each data point error is product of deviation  $y_n - t_n$  and basis function  $\phi_n$ .
- BUT : maximum likelihood solution can exhibit over-fitting even for many data points; should use regularised error or MAP then.

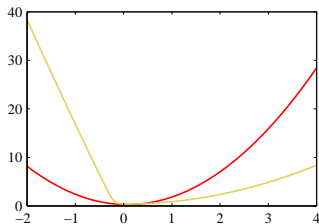
# Laplace Approximation



- Given a continuous distribution  $p(x)$  which is not Gaussian, can we approximate it by a Gaussian  $q(x)$  ?
- Need to find a mode of  $p(x)$ . Try to find a Gaussian with the same mode.



Non-Gaussian (yellow) and  
Gaussian approximation (red).



Negative log of the  
Non-Gaussian (yellow) and  
Gaussian approx. (red).



- Assume  $p(x)$  can be written as

$$p(z) = \frac{1}{Z} f(z)$$

with normalisation  $Z = \int f(z) dz$ .

- Furthermore, assume  $Z$  is unknown !
- A mode of  $p(z)$  is at a point  $z_0$  where  $p'(z_0) = 0$ .
- Taylor expansion of  $\ln f(z)$  at  $z_0$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

where

$$A = - \frac{d^2}{dz^2} \ln f(z) \Big|_{z=z_0}$$

Probabilistic Generative  
Models

Continuous Input

Discrete Features

Probabilistic  
Discriminative Models

Logistic Regression

Iterative Reweighted  
Least Squares

Laplace Approximation

Bayesian Logistic  
Regression



- Exponentiating

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}A(z - z_0)^2$$

- we get

$$f(z) \simeq f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}.$$

- And after normalisation we get the **Laplace approximation**

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}.$$

- Only defined for precision  $A > 0$  as only then  $p(z)$  has a maximum.



# Laplace Approximation - Vector Space

- Approximate  $p(\mathbf{z})$  for  $\mathbf{z} \in \mathbb{R}^M$

$$p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z}).$$

- we get the Taylor expansion

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)$$

- where the Hessian  $\mathbf{A}$  is defined as

$$\mathbf{A} = -\nabla \nabla \ln f(\mathbf{z}) \big|_{\mathbf{z}=\mathbf{z}_0}.$$

- The Laplace approximation of  $p(\mathbf{z})$  is then

$$\begin{aligned} q(\mathbf{z}) &= \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp \left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0) \right\} \\ &= \mathcal{N}(\mathbf{z} | \mathbf{z}_0, \mathbf{A}^{-1}) \end{aligned}$$

Probabilistic Generative  
Models

Continuous Input

Discrete Features

Probabilistic  
Discriminative Models

Logistic Regression

Iterative Reweighted  
Least Squares

Laplace Approximation

Bayesian Logistic  
Regression



- Exact Bayesian inference for the logistic regression is intractable.
- Why? Need to normalise a product of prior probabilities and likelihoods which itself are a product of logistic sigmoid functions, one for each data point.
- Evaluation of the predictive distribution also intractable.
- Therefore we will use the Laplace approximation.

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*



- Assume a Gaussian prior because we want a Gaussian posterior.

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$$

for fixed **hyperparameter**  $\mathbf{m}_0$  and  $\mathbf{S}_0$ .

- **Hyperparameters** are parameters of a prior distribution. In contrast to the model parameters  $\mathbf{w}$ , they are not learned.
- For a set of training data  $(\mathbf{x}_n, t_n)$ , where  $n = 1, \dots, N$ , the posterior is given by

$$p(\mathbf{w} | \mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t} | \mathbf{w})$$

where  $\mathbf{t} = (t_1, \dots, t_N)^T$ .

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*





- Using our previous result for the cross-entropy function

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}) = -\sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

we can now calculate the log of the posterior

$$p(\mathbf{w} | \mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t} | \mathbf{w})$$

using the notation  $y_n = \sigma(\mathbf{w}^T \phi_n)$  as

$$\begin{aligned} \ln p(\mathbf{w} | \mathbf{t}) = & -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\ & + \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \end{aligned}$$



- To obtain a Gaussian approximation to

$$\ln p(\mathbf{w} | \mathbf{t}) = -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) + \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

- 1 Find  $\mathbf{w}_{MAP}$  which maximises  $\ln p(\mathbf{w} | \mathbf{t})$ . This defines the mean of the Gaussian approximation. (Note: This is a nonlinear function in  $\mathbf{w}$  because  $y_n = \sigma(\mathbf{w}^T \phi_n)$ .)
- 2 Calculate the second derivative of the negative log likelihood to get the inverse covariance of the Laplace approximation

$$\mathbf{S}_N = -\nabla \nabla \ln p(\mathbf{w} | \mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^N y_n(1 - y_n) \phi_n \phi_n^T.$$

Probabilistic Generative Models

Continuous Input

Discrete Features

Probabilistic Discriminative Models

Logistic Regression

Iterative Reweighted Least Squares

Laplace Approximation

Bayesian Logistic Regression



- The approximated Gaussian (via Laplace approximation) of the posterior distribution is now

$$q(\mathbf{w} | \phi) = \mathcal{N}(\mathbf{w} | \mathbf{w}_{MAP}, \mathbf{S}_N)$$

where

$$\mathbf{S}_N = -\nabla \nabla \ln p(\mathbf{w} | \mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^N y_n(1 - y_n) \phi_n \phi_n^T.$$

*Probabilistic Generative  
Models*

*Continuous Input*

*Discrete Features*

*Probabilistic  
Discriminative Models*

*Logistic Regression*

*Iterative Reweighted  
Least Squares*

*Laplace Approximation*

*Bayesian Logistic  
Regression*