Machine Learning

Classification: The Laplace Approximation

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

November 6, 2018

http://www.carlhenrik.com

Introduction

Coursework 2

- The second coursework will start next week.
- This week there are no lab sessions

Today

- Classification (Task)
- Logistic Regression (Model)
- Bayesian Logistic Regression (Model)
- Laplace Approximation (Inference)

Conjugacy

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- If we pick the conjugate prior to the likelhood parameter then the posterior is in the same family as the prior
- This means that we do not have to compute the proportionality (evidence)
- We can just multiply likelihood and prior and identify terms

Conjugacy

$$\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_N, \Sigma_N) \propto \mathcal{N}(\mu(\mathbf{w}), \Sigma_1)\mathcal{N}(\mathbf{w}|\mu, \Sigma_2)$$

- Multiply right-hand side
- Identify the terms on the right-hand side

Conjugacy

$$\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_N, \Sigma_N) \propto \mathcal{N}(\mu(\mathbf{w}), \Sigma_1)\mathcal{N}(\mathbf{w}|\mu, \Sigma_2)$$

- Multiply right-hand side
- Identify the terms on the right-hand side
- *what if conjugacy does not make sense?*

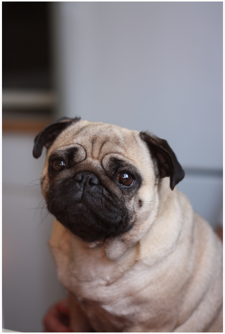
Classification

Classification

- Data $\{\mathcal{D}, \mathcal{C}\}$
  - Variates: $\mathcal{D} = \{x_i\}_{i=1}^{N}$
  - Features: $x \in \mathbb{R}^D$
  - Labels: $\mathcal{C} \in \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$
- Task: given a set of observations and their corresponding class can we associate the correct class to new observations?

Generative model

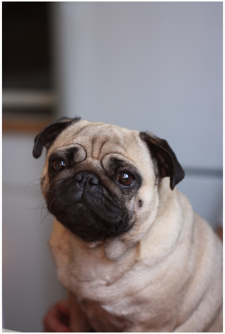
Image This is an image Stella, she was a pug!



Generative model

Image This is an image Stella, she was a pug!

Question does the appearance of the image make her a pug?

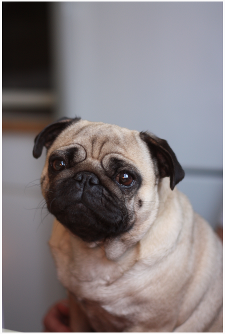


Generative model

Image This is an image Stella, she was a pug!

Question does the appearance of the image make her a pug?

Question does her being a pug make an image of her appear like this?



Generative model

- The image appears this way because she was a pug

Generative model

- The image appears this way because she was a pug
- *Is it possible to have the same image while its not Stella?*

Generative model

- The image appears this way because she was a pug
- *Is it possible to have the same image while its not Stella?*
- *Is it possible to have the same image while its not a pug?*

Generative model

- The image appears this way because she was a pug
- *Is it possible to have the same image while its not Stella?*
- *Is it possible to have the same image while its not a pug?*
- To handle all possible cases we need to think about how the data was created

Generative model

- The image appears this way because she was a pug
- *Is it possible to have the same image while its not Stella?*
- *Is it possible to have the same image while its not a pug?*
- To handle all possible cases we need to think about how the data was created
  - Pug + Camera + *lots of other stuff* $\rightarrow$ Image

$$f(x) = I$$

Generative model

- The image appears this way because she was a pug
- *Is it possible to have the same image while its not Stella?*
- *Is it possible to have the same image while its not a pug?*
- To handle all possible cases we need to think about how the data was created
  - Pug + Camera + *lots of other stuff* $\rightarrow$ Image

$$f(x) = I$$

- *When we formulate models we want to formulate the generation of the data*

Classification

$$p(\mathcal{D}, \mathcal{C}) = p(\mathcal{D}|\mathcal{C})p(\mathcal{C})$$

1. Formulate the likelihood and the prior
2. Formulate the posterior
3. Get updated belief through posterior with new data

Classification

- Lets assume we want to update our belief of class $\mathcal{C}_1$ from the information in x

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x)}$$

Classification

- Lets assume we want to update our belief of class $\mathcal{C}_1$ from the information in x

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x)}$$

- What is the evidence in this case?

$$p(x) = \int p(x|\mathcal{C})p(\mathcal{C})\mathrm{d}\mathcal{C} = \sum_{i=1}^{k} p(x|\mathcal{C}_i)p(\mathcal{C}_i)$$

Classification

- Lets assume we want to update our belief of class $\mathcal{C}_1$ from the information in x

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x)}$$

- What is the evidence in this case?

$$p(x) = \int p(x|\mathcal{C})p(\mathcal{C})\mathrm{d}\mathcal{C} = \sum_{i=1}^{k} p(x|\mathcal{C}_i)p(\mathcal{C}_i)$$

- Posterior

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{\sum_{i=1}^{k} p(x|\mathcal{C}_i)p(\mathcal{C}_i)}$$

Binary Classification

- Lets assume we have only two classes i.e. $\mathcal{C} \in \{\mathcal{C}_1, \mathcal{C}_2\}$

$$p(x) = p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)$$

Binary Classification

- Lets assume we have only two classes i.e. $\mathcal{C} \in \{\mathcal{C}_1, \mathcal{C}_2\}$

$$p(x) = p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)$$

- Posterior:

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)} =$$

Binary Classification

- Lets assume we have only two classes i.e. $\mathcal{C} \in \{\mathcal{C}_1, \mathcal{C}_2\}$

$$p(x) = p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)$$

- Posterior:

$$\begin{aligned} p(\mathcal{C}_1|x) &= \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)} = \\ &= \frac{\left(\frac{1}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}\right)}{\left(\frac{1}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}\right)} \cdot \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)} = \end{aligned}$$

Binary Classification

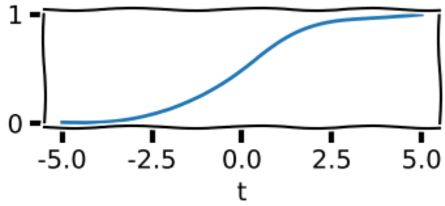
- Lets assume we have only two classes i.e. $\mathcal{C} \in \{\mathcal{C}_1, \mathcal{C}_2\}$

$$p(x) = p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)$$

- Posterior:

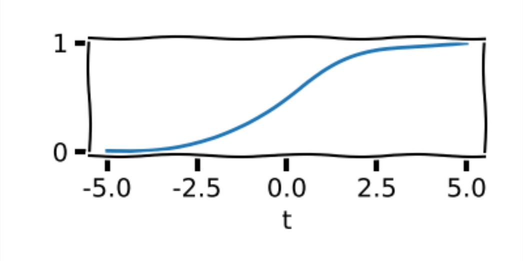
$$\begin{aligned}
p(\mathcal{C}_1|x) &= \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)} = \\
&= \frac{\left(\frac{1}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}\right)}{\left(\frac{1}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}\right)} \cdot \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)} = \\
&= \frac{1}{1 + \frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}}
\end{aligned}$$

Binary Classification



$$y = \frac{1}{1 + e^{-t}}$$

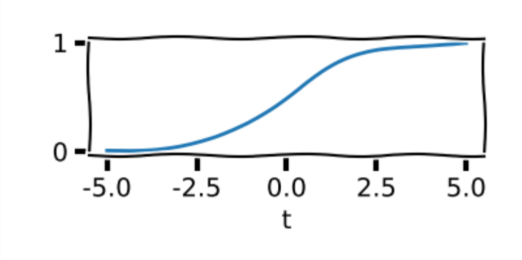
Binary Classification



$$y = \frac{1}{1 + e^{-t}}$$

$$p(\mathcal{C}_1|x) = \frac{1}{1 + \frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}}$$

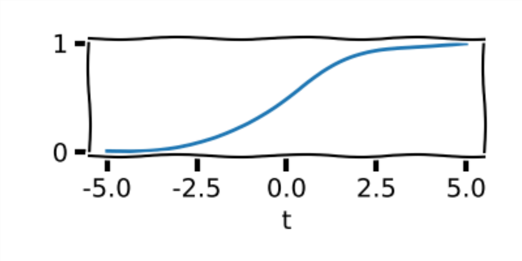
Binary Classification



$$y = \frac{1}{1 + e^{-t}}$$

$$p(\mathcal{C}_1|x) = \frac{1}{1 + \frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}} = \frac{1}{1 + \exp\left(\log\left(\frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}\right)\right)}$$

Binary Classification

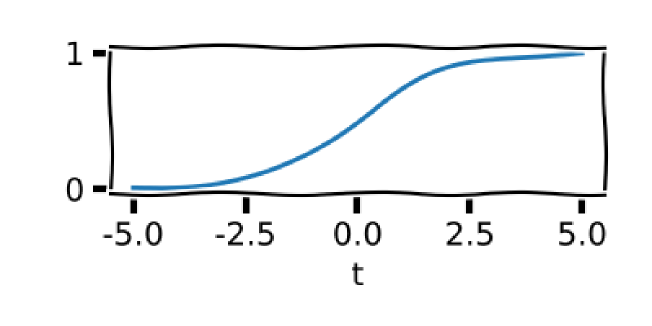


$$y = \frac{1}{1 + e^{-t}}$$

$$p(\mathcal{C}_1|x) = \frac{1}{1 + \frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}} = \frac{1}{1 + \exp\left(\log\left(\frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}\right)\right)}$$

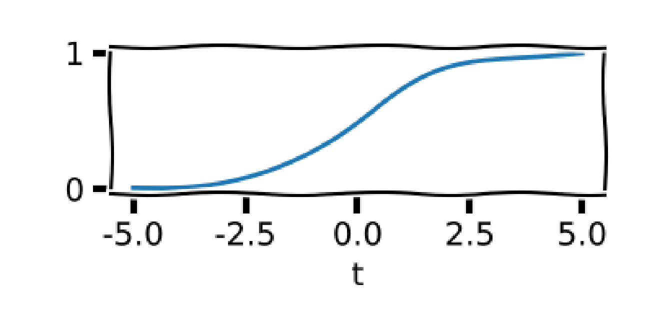
$$= \frac{1}{1 + \exp\left(-\log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right)\right)}$$

Binary Classification



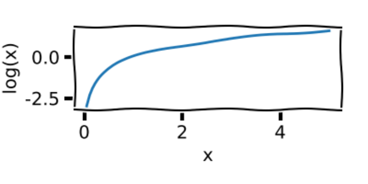
$$t = \log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right)$$

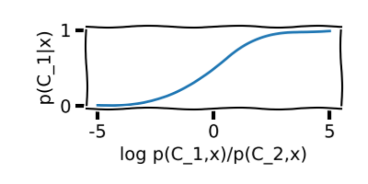
Binary Classification



$$t = \log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right) = \log\left(\frac{p(x,\mathcal{C}_1)}{p(x,\mathcal{C}_2)}\right)$$

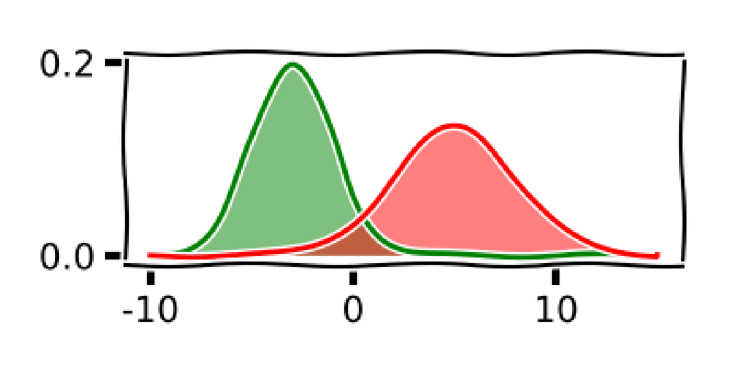
Binary Classification





- $p(x, \mathcal{C}_1) > p(x, \mathcal{C}_2)$
  - $p(\mathcal{C}_1|x) > 0.5$
- $p(x, \mathcal{C}_1) < p(x, \mathcal{C}_2)$
  - $p(\mathcal{C}_1|x) < 0,5$
- $p(x, \mathcal{C}_1) = p(x, \mathcal{C}_2)$
  - $p(\mathcal{C}_1|x) = 0.5$
- $p(x, \mathcal{C}_1) = 0$
  - $p(\mathcal{C}_1|x) = 0$
- $p(x, \mathcal{C}_2) \to 0$
  - $p(\mathcal{C}_1|x) \to 1$
- $p(x, \mathcal{C}_1) = p(x, \mathcal{C}_2) = 0$
  - Undefined

Likelihoods



- We haven't specified the model yet, lets make a Gaussian likelihood

$$p(x|\mathcal{C}_i) = \mathcal{N}(x|\mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

Posterior

$$t = \log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right)$$

Posterior

$$t = \log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right) = \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}p(\mathcal{C}_1)}{\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}p(\mathcal{C}_2)}\right)$$

Posterior

$$t = \log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right) = \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}p(\mathcal{C}_1)}{\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}p(\mathcal{C}_2)}\right)$$

$$= \log\left(\frac{A}{B}\right) = \log(A) - \log(B)$$

Posterior cont.

$$\log(A) = \log\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} p(\mathcal{C}_1)\right)$$

Posterior cont.

$$\log(A) = \log\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} p(\mathcal{C}_1)\right) =$$

$$= -\frac{1}{2}\log\left(2\pi\sigma_1^2\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \log\left(p(\mathcal{C}_1)\right)$$

Posterior cont.

$$\log(A) = \log\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} p(\mathcal{C}_1)\right) =$$

$$= -\frac{1}{2}\log\left(2\pi\sigma_1^2\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \log\left(p(\mathcal{C}_1)\right) =$$

$$= -\frac{1}{2}\log\left(2\pi\sigma_1^2\right) - \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} + \log\left(p(\mathcal{C}_1)\right)$$

Posterior cont.

$$\log(A) - \log(B)$$

Posterior cont.

$$\log(A) - \log(B) = -\frac{1}{2}\log\left(\frac{2\pi\sigma_1^2}{2\pi\sigma_2^2}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$$

Posterior cont.

$$\log(A) - \log(B) = -\frac{1}{2}\log\left(\frac{2\pi\sigma_1^2}{2\pi\sigma_2^2}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) - \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} + \frac{x^2 - 2x\mu_2 + \mu_2^2}{2\sigma_2^2}$$

Posterior cont.

$$\begin{aligned}\log(A) - \log(B) &= -\frac{1}{2}\log\left(\frac{2\pi\sigma_1^2}{2\pi\sigma_2^2}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)\\ &- \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} + \frac{x^2 - 2x\mu_2 + \mu_2^2}{2\sigma_2^2} =\\ &= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)\end{aligned}$$

Posterior cont.

$$\begin{aligned}
\log(A) - \log(B) &= -\frac{1}{2}\log\left(\frac{2\pi\sigma_1^2}{2\pi\sigma_2^2}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&- \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} + \frac{x^2 - 2x\mu_2 + \mu_2^2}{2\sigma_2^2} = \\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&- x^2\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right) + x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) - \left(\frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right)
\end{aligned}$$

Posterior cont.

$$\begin{aligned}\log(A) - \log(B) &= -\frac{1}{2}\log\left(\frac{2\pi\sigma_1^2}{2\pi\sigma_2^2}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\ &- \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} + \frac{x^2 - 2x\mu_2 + \mu_2^2}{2\sigma_2^2} = \\ &= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\ &- x^2\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right) + x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) - \left(\frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) = \\ &= t\end{aligned}$$

Posterior cont.

$$
\begin{aligned}
\log(A) - \log(B) &= -\frac{1}{2}\log\left(\frac{2\pi\sigma_1^2}{2\pi\sigma_2^2}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&\quad - \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} + \frac{x^2 - 2x\mu_2 + \mu_2^2}{2\sigma_2^2} = \\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&\quad - x^2\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right) + x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) - \left(\frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) = \\
&= t \\
p(\mathcal{C}_1|x) &= \frac{1}{1 + e^{-t}}
\end{aligned}
$$

Posterior analysis

- The posterior over $\mathcal{C}_1$ (and its the same for $\mathcal{C}_2$) depends on x as

$$-x^2\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2}\right)+x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)$$

Posterior analysis

- The posterior over $\mathcal{C}_1$ (and its the same for $\mathcal{C}_2$) depends on x as

$$-x^2\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2}\right)+x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)$$

- if $\sigma_1=\sigma_2$

Posterior analysis

- The posterior over $\mathcal{C}_1$ (and its the same for $\mathcal{C}_2$) depends on x as
$$-x^2\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2}\right)+x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)$$
- if $\sigma_1=\sigma_2$
    - the posterior is linear in x

Posterior analysis

- The posterior over $\mathcal{C}_1$ (and its the same for $\mathcal{C}_2$) depends on x as

$$-x^2 \left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2} \right) + x \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right)$$

- if $\sigma_1 = \sigma_2$
  - the posterior is linear in x
- if $\mu_2 = \mu_1$ and $\sigma_1 = \sigma_2$

Posterior analysis

- The posterior over $\mathcal{C}_1$ (and its the same for $\mathcal{C}_2$) depends on x as

$$-x^2\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2}\right)+x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)$$

- if $\sigma_1=\sigma_2$
  - the posterior is linear in x
- if $\mu_2=\mu_1$ and $\sigma_1=\sigma_2$
  - the posterior does not depend on x

$$p(\mathcal{C}_1|x)=\frac{1}{1+e^{-\log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}}}=\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_1)+p(\mathcal{C}_2)}=\frac{p(\mathcal{C}_1)}{p(\mathcal{C})}=p(\mathcal{C}_1)$$

Posterior analysis

- The posterior over $\mathcal{C}_1$ (and its the same for $\mathcal{C}_2$) depends on x as

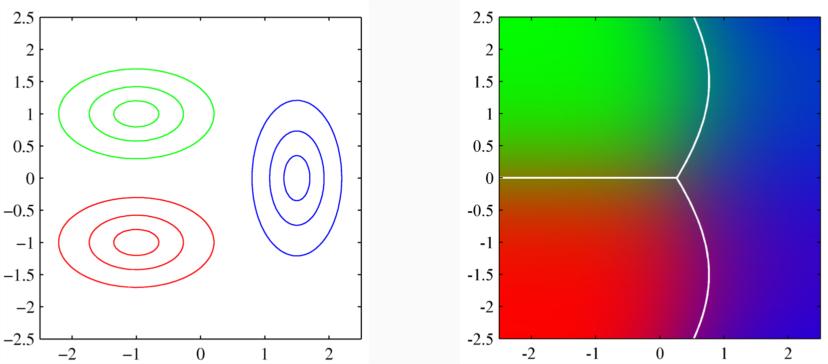
$$-x^2\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2}\right)+x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)$$

- if $\sigma_1=\sigma_2$
  - the posterior is linear in x
- if $\mu_2=\mu_1$ and $\sigma_1=\sigma_2$
  - the posterior does not depend on x

$$p(\mathcal{C}_1|x)=\frac{1}{1+e^{-\log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}}}=\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_1)+p(\mathcal{C}_2)}=\frac{p(\mathcal{C}_1)}{p(\mathcal{C})}=p(\mathcal{C}_1)$$

- *if the observations does not provide me with any information to update my belief my posterior belief is equal to my prior belief*

Posterior analysis[1]



- Red & Green share the same covariance ⇒ linear separation
- Blue & (Red & Green) different covariance ⇒ curved separation

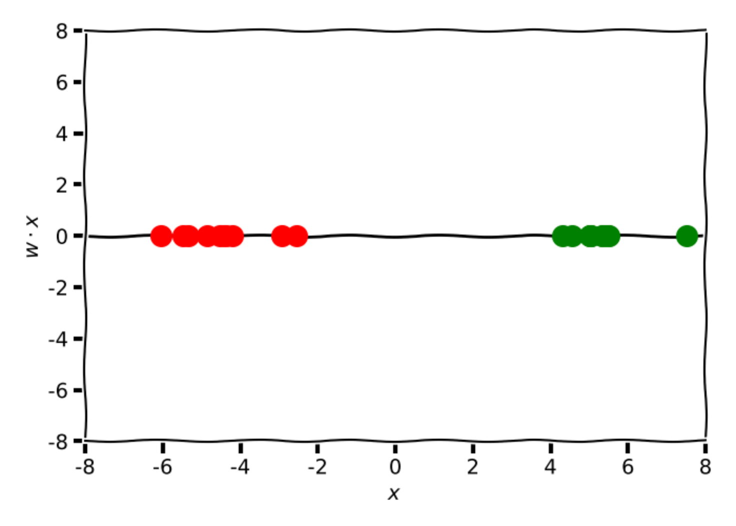
[1] Bishop, C. M. (2006). Figure 4.11

Logistic Regression

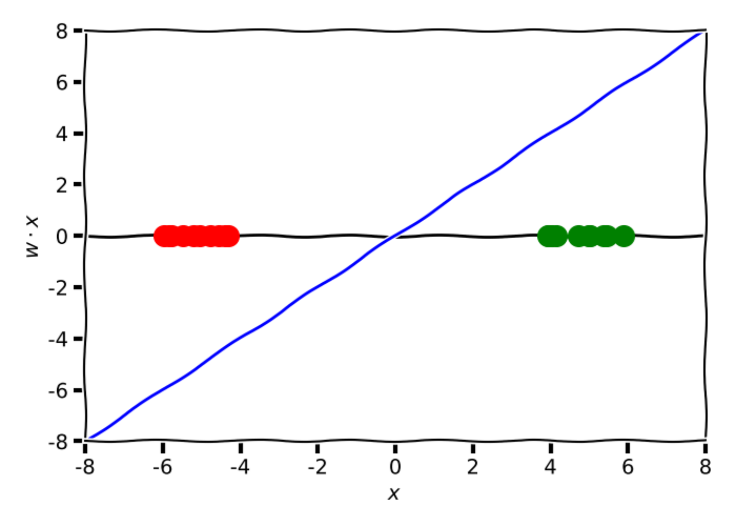
Logistic Regression

$$p(\mathcal{C}_1|x) = \frac{1}{1 + exp(-f(x))}$$

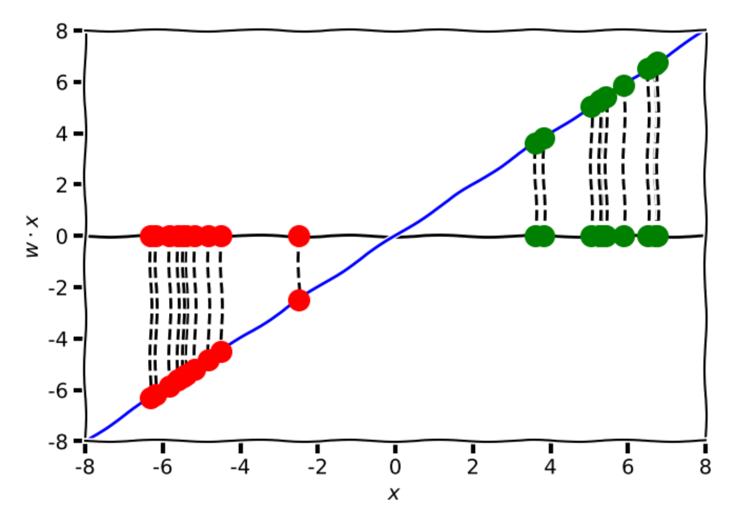
- We can derive the posterior through principle by Bayes Rule
  - This forces us to make our assumptions clear
- We can directly parametrise the posterior as we know its form
  - this is called logistic regression


8
6
4
2
0
-2
-4
-6
-8
$w \cdot x$
-8
-6
-4
-2
0
2
4
6
8
x

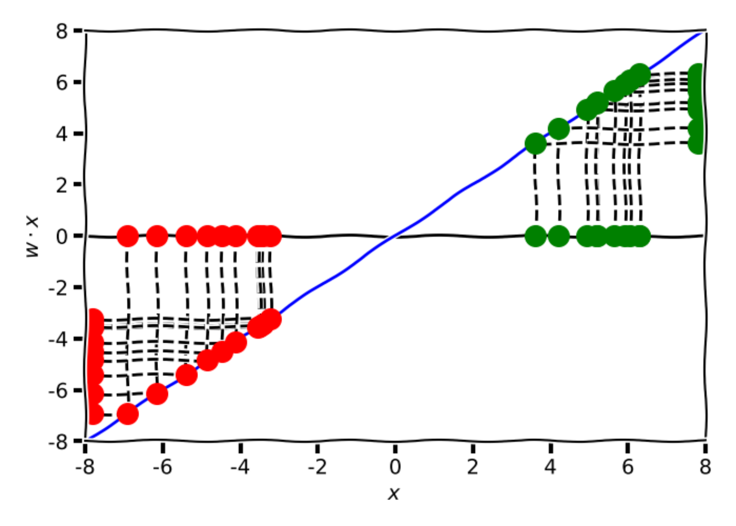
Logistic Regression



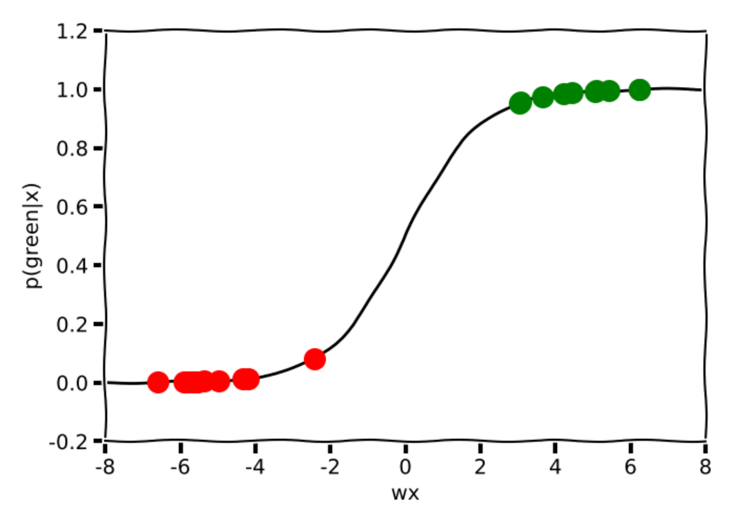
Logistic Regression



Logistic Regression



Logistic Regression



Why

- If we have $\mathbf{x} \in \mathbb{R}^{100}$

Why

- If we have $\mathbf{x} \in \mathbb{R}^{100}$
  - Each Gaussian has

$$\mu_i \in \mathbb{R}^{100}$$
$$\Sigma_i \in \mathbb{R}^{100 \times 100}$$

Why

- If we have $\mathbf{x} \in \mathbb{R}^{100}$
    - Each Gaussian has

$$\mu_i \in \mathbb{R}^{100}$$
$$\Sigma_i \in \mathbb{R}^{100 \times 100}$$

- Binary classification implies 20200 parameters

Why

- If we have $\mathbf{x} \in \mathbb{R}^{100}$
  - Each Gaussian has

$$\mu_i \in \mathbb{R}^{100}$$
$$\Sigma_i \in \mathbb{R}^{100\times 100}$$

- Binary classification implies 20200 parameters
- However good our priors are this is likely to require a lot of data to learn

Why

- If we have $\mathbf{x} \in \mathbb{R}^{100}$
  - Each Gaussian has

$$\mu_i \in \mathbb{R}^{100}$$
$$\Sigma_i \in \mathbb{R}^{100 \times 100}$$

- Binary classification implies 20200 parameters
- However good our priors are this is likely to require a lot of data to learn
- Logistic regression implies 101

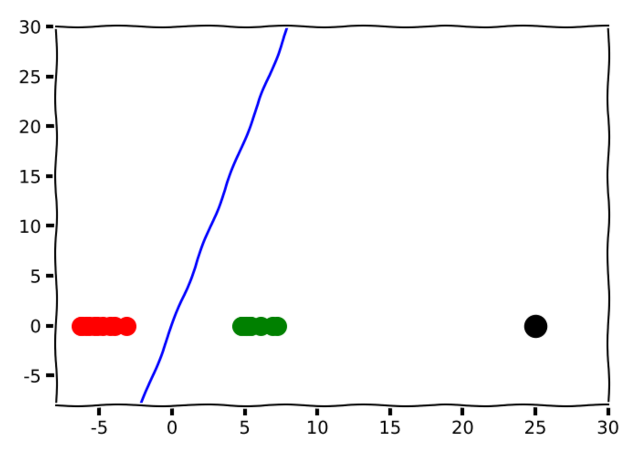
$$\mathbf{w} \in \mathbb{R}^{100}$$

Logistic Regression

$$p(\mathcal{C}|\mathcal{D})$$

- Reaching this through principle makes it a posterior
- If we parametrise it directly we have to see it as a likelihood
  - we do not model $\mathcal{D}$
  - we do not quantify our uncertainty in $\mathcal{D}$
    - denominator in Bayes Rule $p(\mathcal{D})$

Why Not



$$t = \log\left(\frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}\right)$$

Logistic Regression

$$p(\mathcal{C}_1|x) = \frac{1}{1 + exp(-f(x))} = \sigma(x)$$

- We seek a function that is positive for $\mathcal{C}_1$ and negative for $\mathcal{C}_2$
- Linear classifier

$$f(x) = w^{\mathrm{T}}x$$

- Maximum Likelihood

$$\hat{w} = \mathrm{argmax}_w p(\mathcal{C}|\mathcal{D}) = \mathrm{argmin}_w - \log(p(\mathcal{C}|\mathcal{D}))$$

Logistic Regression

$$E(w) = -\log p(\mathcal{C}|\mathcal{D}) = -\log\left(\prod_i^N \sigma(x_i)^{c_i} \cdot (1-\sigma(x_i))^{1-c_i}\right)$$

- we want to minimise the above
- take derivatives with respect to w

$$\frac{\delta E(w)}{\delta w} = \sum_i^N (\sigma(x_i) - c_i)x_i$$

Logistic Regression

$$\frac{\delta E(w)}{\delta w} = \sum_{i}^{N} (\sigma(x_i) - c_i) x_i$$

- $(\sigma(x_i) - c_i)$ is the classification error
    - if 0 no gradient
    - if $\neq 0$ then the gradient goes in the direction of the data-point x
- Using the gradient we can now update w and find a solution

Bayesian Logistic Regression

- We have made no assumptions about the function

- We have made no assumptions about the function
- We have seen how to do linear regression in a principled way

Bayesian Logistic Regression

- We have made no assumptions about the function
- We have seen how to do linear regression in a principled way
    - specify prior over $\mathbf{w}$

Bayesian Logistic Regression

- We have made no assumptions about the function
- We have seen how to do linear regression in a principled way
  - specify prior over $\mathbf{w}$
  - derive posterior

Bayesian Logistic Regression

We want to use the same motivation as we did for normal regression

- Prior

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

- Likelihood

$$p(c_i|\mathbf{w}, \mathbf{x}_i) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$$
$$p(\mathbf{c}|\mathbf{w}, \mathbf{x}) = \prod_i^N \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^{c_i} \cdot (1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i))^{1-c_i}$$

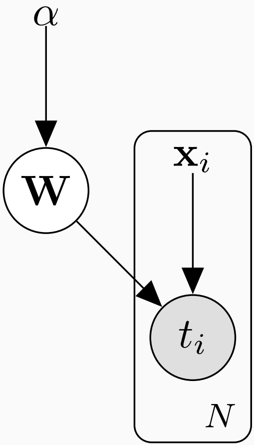
You can also do a feature mapping if you want and replace $\mathbf{x}$ with $\Phi(\mathbf{x})$

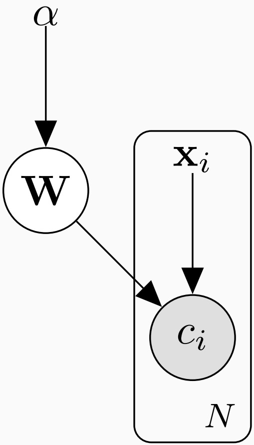
Bayesian Logistic Regression

$$\mathcal{N}(\boldsymbol{\mu}_N, \mathbf{S}_N) \neq \log\left(\prod_i^N \sigma(x_i)^{c_i} \cdot (1 - \sigma(x_i))^{1-c_i}\right) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathrm{T}} \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) - \log(Z)$$

- Previously we could use conjugacy to reach analytical posterior
- In this case we have changed the likelihood and the Gaussian prior is non-conjugate
- This posterior is intractable

Bayesian Linear Regression





Bayesian Logistic Regression

- want to reach posterior over weights

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

- we cannot use conjugacy anymore as likelihood is sigmoid

Laplace Approximation

Laplace Approximation

$$p(z|x) = \frac{1}{Z} f(z) = \frac{f(z)}{\int f(z) dz}$$

- Will use $p(z)$ to refer to the posterior
- $p(z)$ is unknown as we cannot compute Z
- $f(z)$ is possible to compute if we have likelihood and prior

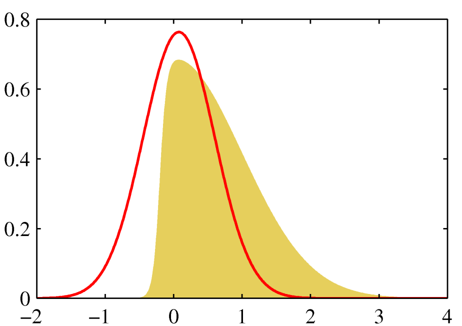
$$f(z) = p(x|z)p(z)$$

Laplace Approximation

$$\log\ p(z) = \log\ (\frac{1}{Z}f(z)) = \log(f(z)) + \text{const w.r.t. z}$$

- $p(z)$ and $f(z)$ will have the same modes
  - *is this always true?*
- Idea: we can approximate each mode with a distribution we can normalise

Laplace Approximation Ch. 4.4 [1]



- Find the mode of the posterior
- Fit Gaussian to this mode

Taylor Expansion

$$f(x) = f(x_0) + \frac{\partial}{\partial x} f(x_0)(x - x_0) + \frac{1}{2}\frac{\partial^2}{\partial x^2} f(x_0)(x - x_0)^2 + \mathcal{O}((x - x0)^3)$$

- A Taylor expansion is an approximation of a function around a specific value
- If we expand around a maxima x_0

$$\frac{\partial}{\partial x} f(x_0) = 0$$

- This leads to

$$f(x) = f(x_0) - \frac{1}{2}\left|\frac{\partial^2}{\partial x^2} f(x_0)\right|(x - x_0)^2 + \mathcal{O}((x - x0)^3)$$

Mode

$$f(\mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$$

- we want to find the mode of this, i.e. the maxima

$$\hat{\mathbf{w}} = \text{argmax}_{\mathbf{w}} p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$$

- This we know as the Maximum-a-Posterior (MAP) estimate

Laplace Approximation

1. Find mode of $p(z)$

$$\frac{\partial}{\partial z} p(z_0) = \frac{\partial}{\partial z} f(z_0) = 0$$

Laplace Approximation

1. Find mode of $p(z)$

$$\frac{\partial}{\partial z} p(z_0) = \frac{\partial}{\partial z} f(z_0) = 0$$

2. Make Taylor Expansion around mode

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} \frac{\partial^2}{\partial^2} \log(f(z_0))(z - z_0)^2$$

Laplace Approximation

1. Find mode of $p(z)$

$$\frac{\partial}{\partial z}p(z_0) = \frac{\partial}{\partial z}f(z_0) = 0$$

2. Make Taylor Expansion around mode

$$\log f(z) \approx \log f(z_0) - \frac{1}{2}\frac{\partial^2}{\partial^2}\log(f(z_0))(z - z_0)^2$$

3. Take exponential to get function

$$f(z) \approx f(z_0)e^{-\frac{1}{2}\underbrace{\frac{\partial^2}{\partial^2}\log(f(z_0))}_{A}(z-z_0)^2} = f(z_0)e^{-\frac{1}{2}A(z-z_0)^2}$$

Laplace Approximation

$$f(z) \approx f(z_0) e^{-\frac{1}{2}A(z-z_0)^2}$$

- we want to find an approximation, to $p(z)$ so we need to normalise to a distribution

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

Laplace Approximation

$$f(z) \approx f(z_0)e^{-\frac{1}{2}A(z-z_0)^2}$$

- we want to find an approximation, to $p(z)$ so we need to normalise to a distribution

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

- assume that $q(z)$ is Gaussian, i.e. $f(z_0) = p(\text{mean})$

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

Laplace Approximation

- One dimensional

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

- D dimensional

$$q(\mathbf{z}) = \frac{|\mathbf{A}|}{(2\pi)^{\frac{D}{2}}} e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

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

- Where **A** is the Hessian matrix

Bayesian Logistic Regression

We want to use the same motivation as we did for normal regression

- Prior

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

- Likelihood

$$p(c_i|\mathbf{w}, \mathbf{x}_i) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$$

$$p(\mathbf{c}|\mathbf{w}, \mathbf{x}) = \prod_{i}^{N} \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^{c_i} \cdot (1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i))^{1-c_i}$$

You can also do a feature mapping if you want and replace $\mathbf{x}$ with $\Phi(\mathbf{x})$

Bayesian Logistic Regression

$$q(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N) \approx p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) = f(\mathbf{w})$$

- Compute $f(\mathbf{w})$

$$\begin{aligned}\log p(\mathbf{w}|\mathbf{t}) = &\log\left(\prod_i^N \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^{t_i} \cdot (1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i))^{1-t_i}\right) \\ &- \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) - \log(Z)\end{aligned}$$

- The stationary point is the MAP estimate

Bayesian Logistic Regression

$$\mathbf{S}_N^{-1} = -\nabla\nabla \log p(\mathbf{w}|\mathbf{t})|_{\mathbf{w}=\mathbf{w}_{MAP}} = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}))\mathbf{x}\mathbf{x}^{\mathrm{T}}$$

- we can compute the Hessian around $\mathbf{w}_{\mathrm{MAP}}$
- this leads to the final approximation

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

Summary

- Compute a mode of the posterior distribution, i.e MAP estimate

Laplace Approximation

Summary

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode

Laplace Approximation

Summary

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode
    - this gives us only a quadratic term left

Summary

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode
  - this gives us only a quadratic term left
- Identify elements in expansion as parameters of a Gaussian

Summary

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode
    - this gives us only a quadratic term left
- Identify elements in expansion as parameters of a Gaussian
- Normalise to a distribution

Laplace Approximation

Summary

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode
  - this gives us only a quadratic term left
- Identify elements in expansion as parameters of a Gaussian
- Normalise to a distribution
- *You can do exactly the same thing with a GP*

Prediction

$$p(\mathcal{C}_1|\mathbf{x},\mathbf{t}) = \int p(\mathcal{C}_1|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathbf{t})\mathrm{d}\mathbf{w}$$
$$\approx \int p(\mathcal{C}_1|\mathbf{x},\mathbf{w})q(\mathbf{w})\mathrm{d}\mathbf{w}$$

- To compute predictions we can use our new approximate posterior in place of the true posterior

Summary

Summary

- Classification often means non-conjugate prior
- Laplace Approximation
  - match modes with the true posterior
- As we often know the MAP estimate of different models we can often apply this method relatively easy
- Can be really bad if the posterior is far from Gaussian
- We can fit several modes and make a mixture

eof

References

Christopher M. Bishop.
Pattern Recognition and Machine Learning (Information Science and Statistics).
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.