M5MS10
Machine Learning

Spring 2018

Lecture 3

Dr Ben Calderhead

b.calderhead@imperial.ac.uk

M5MS10
Machine Learning

Spring 2018

Lecture 3

Dr Ben Calderhead

b.calderhead@imperial.ac.uk

IS10
Learning

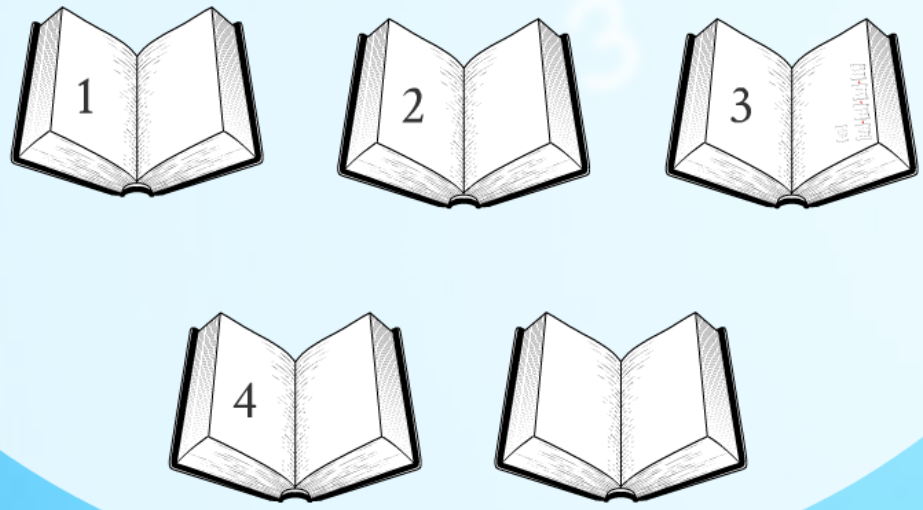
2018

re 3

lderhead

mperial.ac.uk

Parametric Methods



Introduction to Classification

What is Classification?

Classification - supervised learning in which the labels are from a *discrete* set of values.

Automatic methods for deciding in which group a new object should be categorised.

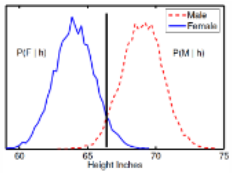
Very important class of problems in the fields of Machine Learning and Statistics.

Examples include:

- Spam filtering
- Fraud detection (e.g. credit card transaction fraud)
- Character/face recognition
- Medical diagnosis

Bayesian Classification Example

As an example let's try to make a classifier that makes predictions about the gender of a person based only on knowledge of their height.



Class Priors

The class variable C will take on two values:

- Male will be encoded by the value 1
- Female will be encoded by the value 0

We denote the probability of class "male" occurring as $p(C = 1)$, and the probability of class "female" occurring as $p(C = 0)$.

This is our prior distribution, since it defines what we might expect *before* we see the data.

Class Priors

We note that within a general population there is approximately an even number of males and females, and so we might reasonably set the prior to be $p(C = 1) = 0.5$ and $p(C = 0) = 0.5$.

For other applications, e.g. medical diagnostics, we might have a much smaller prior probability for one of our classes, e.g. probability of a particular disease within a general population. This allows the model to make more accurate predictions in the absence of data, or with very little data.

Class Conditioned Likelihood

For each class, we then have a different distribution for the height variable, h.

For example, we would expect males on average to be taller, and so the distribution of the heights of males will have a different mean from the distribution of the heights of females.

This is known as a class conditional distribution, which in this case we denote as $p(h|C = 1)$ and $p(h|C = 0)$, for males and females respectively.

We can use this as a likelihood to obtain the posterior distribution over the class variable, C.

Class Posterior

We can use Bayes' theorem to obtain the posterior distribution,

$$p(C|h) = \frac{p(h|C)p(C)}{p(h)}$$

We see that the marginal likelihood is the probability of measuring a height, $p(h)$, regardless of the class (i.e. integrating out the class variable). Since our class variable is binary, it follows straightforwardly that

$$\begin{aligned} p(h) &= \sum_C p(h|C)p(C) \\ &= p(h|C = 1)p(C = 1) + p(h|C = 0)p(C = 0) \end{aligned}$$

and so the posterior of the class variable is well-defined since it sums to one.

$$p(C = 1|h) + p(C = 0|h) = 1$$

Discriminant Functions

We notice that we can generally distinguish between males and females using this model, since the means of the two class conditional distributions are quite well separated.

However, there is a region in the middle where the two distributions overlap, and when we make an observation of h that lies in this area we might make classification errors.

In particular the decision boundary is given by the intersection, where $p(C = 1|h) = p(C = 0|h)$.

For any given measurement, there will be some probability that it belongs to each class. We can make decisions that minimise the error by using the *larger* of the two class posterior probabilities.

e.g. if $p(C = 1|h) > p(C = 0|h)$, then we can make the prediction that h belongs to class $C = 1$.

Discriminant Functions

We can simplify this decision making process by introducing a discriminant function based on the posterior probabilities.

One approach is to consider a function based on the *ratio* of posterior probabilities. In particular, taking the log of such a ratio gives us,

$$f(h) = \log \frac{P(C = 1|h)}{P(C = 0|h)}$$

which allows us to use the decision rule that if $f(h) > 0$ then h would be assigned to the male class, $C = 1$, otherwise if $f(h) < 0$ then h would be assigned to the female class, $C = 0$

Discriminative vs Generative

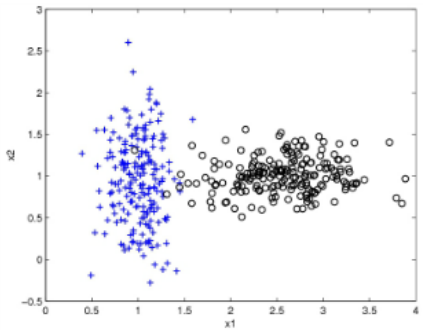
We considered the posterior distribution over the class variable given some class conditional distribution, and come up with a discriminant function for making decisions about class membership.

There are two main approaches for doing classification using this discriminant function.

In the generative approach, we focus on defining the class conditional distributions, $p(h|C = 1)$ and $p(h|C = 0)$, and then use Bayes' theorem to obtain the discriminant function. We can use this approach to generate typical data from the model by drawing samples from $p(h|C)$.

An alternative is the discriminative approach, which involves modelling the discriminant function *directly*, for example using a linear model like the one we used for regression. We can consider this by forming a likelihood based on the discriminant function.

What is Classification?



Classification- supervised learning in which the labels are from a*discrete*set of values.

Automatic methods for deciding in which group a new object should be categorised.

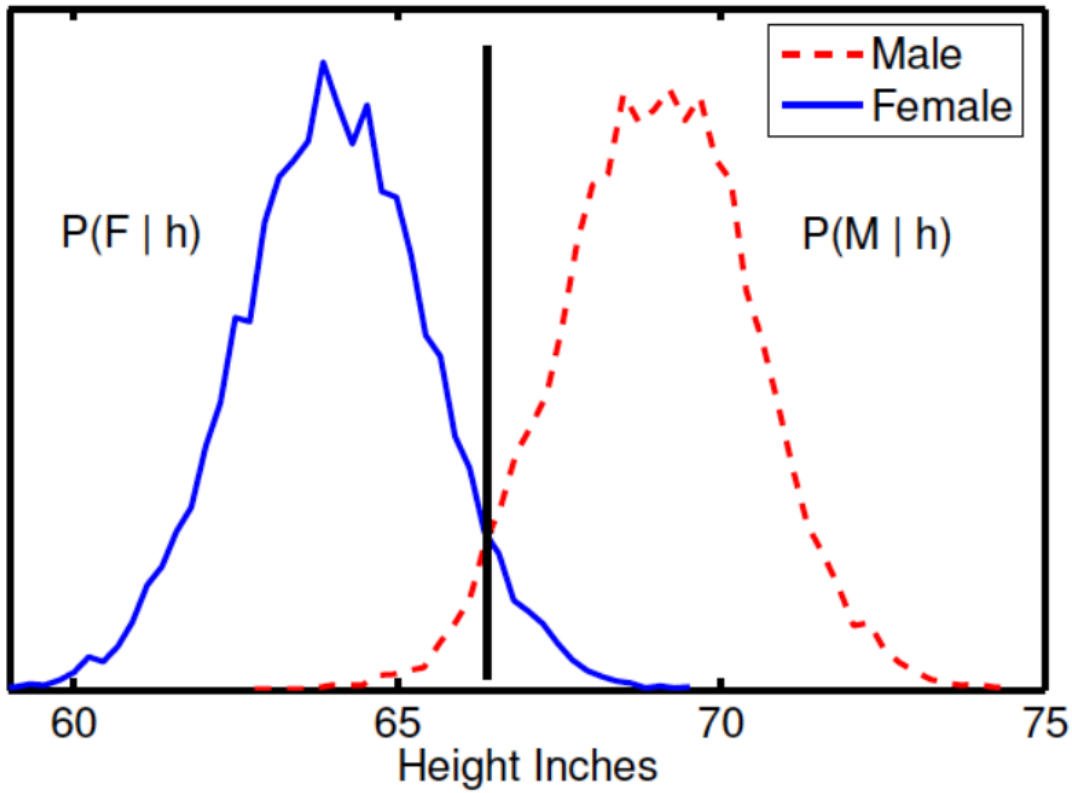
Very important class of problems in the fields of Machine Learning and Statistics.

Examples include:

- Spam filtering
- Fraud detection (e.g. credit card transaction fraud)
- Character/face recognition
- Medical diagnosis

Bayesian Classification Example

As an example let's try to make a classifier that makes predictions about the gender of a person based only on knowledge of their height.



Class Priors

The class variable C will take on two values:

- Male will be encoded by the value 1
- Female will be encoded by the value 0

We denote the probability of class "male" occurring as $p(C = 1)$, and the probability of class "female" occurring as $p(C = 0)$.

This is our prior distribution, since it defines what we might expect *before* we see the data.

Class Priors

We note that within a general population there is approximately an even number of males and females, and so we might reasonably set the prior to be $p(C = 1) = 0.5$ and $p(C = 0) = 0.5$.

For other applications, e.g. medical diagnostics, we might have a much smaller prior probability for one of our classes, e.g. probability of a particular disease within a general population. This allows the model to make more accurate predictions in the absence of data, or with very little data.

Class Conditioned Likelihood

For each class, we then have a different distribution for the height variable, h.

For example, we would expect males on average to be taller, and so the distribution of the heights of males will have a different mean from the distribution of the heights of females.

This is known as a class conditional distribution, which in this case we denote as $p(h|C = 1)$ and $p(h|C = 0)$, for males and females respectively.

We can use this as a likelihood to obtain the posterior distribution over the class variable, C.

Class Posterior

We can use Bayes' theorem to obtain the posterior distribution,

$$p(C|h) = \frac{p(h|C)p(C)}{p(h)}$$

We see that the marginal likelihood is the probability of measuring a height, $p(h)$, regardless of the class (i.e. integrating out the class variable). Since our class variable is binary, it follows straightforwardly that

$$\begin{aligned} p(h) &= \sum_C p(h|C)p(C) \\ &= p(h|C=1)p(C=1) + p(h|C=0)p(C=0) \end{aligned}$$

and so the posterior of the class variable is well-defined since it sums to one.

$$p(C=1|h) + p(C=0|h) = 1$$

Discriminant Functions

We notice that we can generally distinguish between males and females using this model, since the means of the two class conditional distributions are quite well separated.

However, there is a region in the middle where the two distributions overlap, and when we make an observation of h that lies in this area we might make classification errors.

In particular the decision boundary is given by the intersection, where $p(C = 1|h) = p(C = 0|h)$.

For any given measurement, there will be some probability that it belongs to each class. We can make decisions that minimise the error by using the *larger* of the two class posterior probabilities.

e.g. if $p(C = 1|h) > p(C = 0|h)$, then we can make the prediction that h belongs to class $C = 1$.

Discriminant Functions

We can simplify this decision making process by introducing a discriminant function based on the posterior probabilities.

One approach is to consider a function based on the *ratio* of posterior probabilities. In particular, taking the log of such a ratio gives us,

$$f(h) = \log \frac{P(C = 1|h)}{P(C = 0|h)}$$

which allows us to use the decision rule that if $f(h) > 0$ then h would be assigned to the male class, $C = 1$, otherwise if $f(h) < 0$ then h would be assigned to the female class, $C = 0$

Discriminative vs Generative

We considered the posterior distribution over the class variable given some class conditional distribution, and come up with a discriminant function for making decisions about class membership.

There are two main approaches for doing classification using this discriminant function.

In the generative approach, we focus on defining the class conditional distributions, $p(h|C = 1)$ and $p(h|C = 0)$, and then use Bayes' theorem to obtain the discriminant function. We can use this approach to generate typical data from the model by drawing samples from $p(h|C)$.

An alternative is the discriminative approach, which involves modelling the discriminant function *directly*, for example using a linear model like the one we used for regression. We can consider this by forming a likelihood based on the discriminant function.

Introduction to Classification

What is Classification?

Classification - supervised learning in which the labels are from a *discrete* set of values.

Automatic methods for deciding in which group a new object should be categorised.

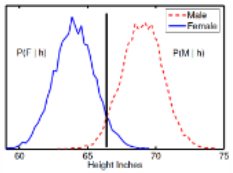
Very important class of problems in the fields of Machine Learning and Statistics.

Examples include:

- Spam filtering
- Fraud detection (e.g. credit card transaction fraud)
- Character/face recognition
- Medical diagnosis

Bayesian Classification Example

As an example let's try to make a classifier that makes predictions about the gender of a person based only on knowledge of their height.



Class Priors

The class variable C will take on two values:

- Male will be encoded by the value 1
- Female will be encoded by the value 0

We denote the probability of class "male" occurring as $p(C = 1)$, and the probability of class "female" occurring as $p(C = 0)$.

This is our prior distribution, since it defines what we might expect *before* we see the data.

Class Priors

We note that within a general population there is approximately an even number of males and females, and so we might reasonably set the prior to be $p(C = 1) = 0.5$ and $p(C = 0) = 0.5$.

For other applications, e.g. medical diagnostics, we might have a much smaller prior probability for one of our classes, e.g. probability of a particular disease within a general population. This allows the model to make more accurate predictions in the absence of data, or with very little data.

Class Conditioned Likelihood

For each class, we then have a different distribution for the height variable, h.

For example, we would expect males on average to be taller, and so the distribution of the heights of males will have a different mean from the distribution of the heights of females.

This is known as a class conditional distribution, which in this case we denote as $p(h|C = 1)$ and $p(h|C = 0)$, for males and females respectively.

We can use this as a likelihood to obtain the posterior distribution over the class variable, C.

Class Posterior

We can use Bayes' theorem to obtain the posterior distribution,

$$p(C|h) = \frac{p(h|C)p(C)}{p(h)}$$

We see that the marginal likelihood is the probability of measuring a height, $p(h)$, regardless of the class (i.e. integrating out the class variable). Since our class variable is binary, it follows straightforwardly that

$$\begin{aligned} p(h) &= \sum_C p(h|C)p(C) \\ &= p(h|C = 1)p(C = 1) + p(h|C = 0)p(C = 0) \end{aligned}$$

and so the posterior of the class variable is well-defined since it sums to one.

$$p(C = 1|h) + p(C = 0|h) = 1$$

Discriminant Functions

We notice that we can generally distinguish between males and females using this model, since the means of the two class conditional distributions are quite well separated.

However, there is a region in the middle where the two distributions overlap, and when we make an observation of h that lies in this area we might make classification errors.

In particular the decision boundary is given by the intersection, where $p(C = 1|h) = p(C = 0|h)$.

For any given measurement, there will be some probability that it belongs to each class. We can make decisions that minimise the error by using the *larger* of the two class posterior probabilities.

e.g. if $p(C = 1|h) > p(C = 0|h)$, then we can make the prediction that h belongs to class $C = 1$.

Discriminant Functions

We can simplify this decision making process by introducing a discriminant function based on the posterior probabilities.

One approach is to consider a function based on the *ratio* of posterior probabilities. In particular, taking the log of such a ratio gives us,

$$f(h) = \log \frac{P(C = 1|h)}{P(C = 0|h)}$$

which allows us to use the decision rule that if $f(h) > 0$ then h would be assigned to the male class, $C = 1$, otherwise if $f(h) < 0$ then h would be assigned to the female class, $C = 0$

Discriminative vs Generative

We considered the posterior distribution over the class variable given some class conditional distribution, and come up with a discriminant function for making decisions about class membership.

There are two main approaches for doing classification using this discriminant function.

In the generative approach, we focus on defining the class conditional distributions, $p(h|C = 1)$ and $p(h|C = 0)$, and then use Bayes' theorem to obtain the discriminant function. We can use this approach to generate typical data from the model by drawing samples from $p(h|C)$.

An alternative is the discriminative approach, which involves modelling the discriminant function *directly*, for example using a linear model like the one we used for regression. We can consider this by forming a likelihood based on the discriminant function.

Discriminative Classification

Discriminative Classification

Let's first consider the discriminative approach, using the notation $\mathbf{x} = [x_1, \ldots, x_D]^T$ to represent the D dimensional vector of input variables or *features* that we can use in our classifier.

We note that the discriminant function

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})}$$

produces values that lie on the real line, between $-\infty$ and ∞.

We can model this using a linear function like the one we saw in lecture 1.

Discriminative Classification

Let's consider a model that is linear in the parameters but possibly nonlinear in the features,

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \theta^T \phi(\mathbf{x})$$

where we have defined a nonlinear function of the inputs $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x})]^T$, and the m^{th} basis function applied to the data is denoted $\phi_m(\mathbf{x})$.

For the regression example, we considered polynomial basis functions, e.g. $\phi_m(\mathbf{x}) = \mathbf{x}^m$.

Since $p(C=1|\mathbf{x}) + p(C=1|\mathbf{x}) = 1$, we can show with a wee bit of algebra that

$$p(C=1|\mathbf{x}) = \frac{1}{1+\exp\left(-\theta^T\phi(\mathbf{x})\right)} = \frac{\exp\left(\theta^T\phi(\mathbf{x})\right)}{1+\exp\left(\theta^T\phi(\mathbf{x})\right)}$$

Discriminative Classification

We can now define the likelihood for each input-output pair, $(\mathbf{x}_n, y_n)$ as

$$\begin{aligned} p(C=y_n|\mathbf{x}_n,\theta) &= p(C=1|\mathbf{x}_n,\theta)^{y_n} \times (1-p(C=1|\mathbf{x}_n,\theta))^{1-y_n} \\ &= \left[\frac{1}{1+\exp\left(-\theta^T\phi(\mathbf{x}_n)\right)}\right]^{y_n}\left[\frac{1}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)}\right]^{1-y_n} \\ &= \frac{\exp\left(\theta^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)} \end{aligned}$$

Given this likelihood function, we could employ the methods we've seen so far to learn the parameters, e.g. maximum-likelihood, cross-validation, Bayesian approach.

Bayesian Approach

Let's consider the Bayesian approach for learning the parameters using this discriminative model.

We may place a Gaussian prior on the coefficients of our linear model, such that $p(\theta|\alpha) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$.

Assuming that our data is i.i.d. the likelihood follows as,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\theta) &= \prod_{n=1}^{N} p(C=y_n|\mathbf{x}_n,\theta) \\ &= \prod_{n=1}^{N} \frac{\exp\left(\theta^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)} \end{aligned}$$

Bayesian Approach

Finally, we can construct the posterior distribution over the parameters.

$$p(\theta|\mathbf{y},\mathbf{X},\alpha) = \frac{p(\mathbf{y}|\mathbf{X},\theta)p(\theta|\alpha)}{p(\mathbf{y}|\mathbf{X},\alpha)}$$

where the marginal likelihood follows as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\alpha) &= \int p(\mathbf{y}|\mathbf{X},\theta)p(\theta|\alpha)d\theta \\ &= \int \prod_{n=1}^{N} \frac{\exp\left(\theta^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})d\theta \end{aligned}$$

Bayesian Approach

Unfortunately, we cannot calculate an analytic form for this potentially high dimensional integral. Some possibilities for estimating this quantity are:

- Use numerical quadrature methods (only good for very low dimensions, <5)
- Use Markov chain Monte Carlo to approximate using a Monte Carlo estimator
- Approximate the posterior with a tractable distribution, such as a Gaussian

Discriminative Classification

Let's first consider the discriminative approach, using the notation $\mathbf{x} = [x_1, \ldots, x_D]^T$ to represent the D dimensional vector of input variables or *features* that we can use in our classifier.

We note that the discriminant function

$$\log \frac{P(C = 1|\mathbf{x})}{P(C = 0|\mathbf{x})}$$

produces values that lie on the real line, between $-\infty$ and ∞.

We can model this using a linear function like the one we saw in lecture 1.

Discriminative Classification

Let's consider a model that is linear in the parameters but possibly nonlinear in the features,

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \boldsymbol{\theta}^T \phi(\mathbf{x})$$

where we have defined a nonlinear function of the inputs $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x})]^T$, and the m'th basis function applied to the data is denoted $\phi_m(\mathbf{x})$.

For the regression example, we considered polynomial basis functions, e.g. $\phi_m(\mathbf{x}) = \mathbf{x}^m$.

Since $p(C=1|\mathbf{x}) + p(C=1|\mathbf{x}) = 1$, we can show with a wee bit of algebra that

$$p(C=1|\mathbf{x}) = \frac{1}{1+\exp\left(-\boldsymbol{\theta}^T \phi(\mathbf{x})\right)} = \frac{\exp\left(\boldsymbol{\theta}^T \phi(\mathbf{x})\right)}{1+\exp\left(\boldsymbol{\theta}^T \phi(\mathbf{x})\right)}$$

Discriminative Classification

We can now define the likelihood for each input-output pair, $(\mathbf{x}_n, y_n)$ as

$$
\begin{aligned}
p(C = y_n|\mathbf{x}_n, \boldsymbol{\theta}) &= p(C = 1|\mathbf{x}_n, \boldsymbol{\theta})^{y_n} \times (1 - p(C = 1|\mathbf{x}_n, \boldsymbol{\theta}))^{1-y_n} \\
&= \left[\frac{1}{1 + \exp\left(-\boldsymbol{\theta}^T\phi(\mathbf{x}_n)\right)}\right]^{y_n} \left[\frac{1}{1 + \exp\left(\boldsymbol{\theta}^T\phi(\mathbf{x}_n)\right)}\right]^{1-y} \\
&= \frac{\exp\left(\boldsymbol{\theta}^T\phi(\mathbf{x}_n)\right)^{y_n}}{1 + \exp\left(\boldsymbol{\theta}^T\phi(\mathbf{x}_n)\right)}
\end{aligned}
$$

Given this likelihood function, we could employ the methods we've seen so far to learn the parameters, e.g. maximum-likelihood, cross-validation, Bayesian approach.

Bayesian Approach

Let's consider the Bayesian approach for learning the parameters using this discriminative model.

We may place a Gaussian prior on the coefficients of our linear model, such that $p(\boldsymbol{\theta}|\alpha) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$.

Assuming that our data is i.i.d. the likelihood follows as,

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) &= \prod_{n=1}^{N} p(C = y_n|\mathbf{x}_n, \boldsymbol{\theta}) \\
&= \prod_{n=1}^{N} \frac{\exp\left(\boldsymbol{\theta}^T \phi(\mathbf{x}_n)\right)^{y_n}}{1 + \exp\left(\boldsymbol{\theta}^T \phi(\mathbf{x}_n)\right)}
\end{aligned}
$$

Bayesian Approach

Finally, we can construct the posterior distribution over the parameters.

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \alpha) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)}$$

where the marginal likelihood follows as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \alpha) &= \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta} \\ &= \int \prod_{n=1}^{N} \frac{\exp\left(\boldsymbol{\theta}^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\boldsymbol{\theta}^T\phi(\mathbf{x}_n)\right)} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})d\boldsymbol{\theta} \end{aligned}$$

Bayesian Approach

Unfortunately, we cannot calculate an analytic form for this potentially high dimensional integral. Some possibilities for estimating this quantity are:

- Use numerical quadrature methods (only good for very low dimensions, <5)
- Use Markov chain Monte Carlo to approximate using a Monte Carlo estimator
- Approximate the posterior with a tractable distribution, such as a Gaussian

Discriminative Classification

Discriminative Classification

Let's first consider the discriminative approach, using the notation $\mathbf{x} = [x_1, \ldots, x_D]^T$ to represent the D dimensional vector of input variables or *features* that we can use in our classifier.

We note that the discriminant function

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})}$$

produces values that lie on the real line, between $-\infty$ and ∞.

We can model this using a linear function like the one we saw in lecture 1.

Discriminative Classification

Let's consider a model that is linear in the parameters but possibly nonlinear in the features,

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \theta^T \phi(\mathbf{x})$$

where we have defined a nonlinear function of the inputs $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x})]^T$, and the m^{th} basis function applied to the data is denoted $\phi_m(\mathbf{x})$.

For the regression example, we considered polynomial basis functions, e.g. $\phi_m(\mathbf{x}) = \mathbf{x}^m$.

Since $p(C=1|\mathbf{x}) + p(C=1|\mathbf{x}) = 1$, we can show with a wee bit of algebra that

$$p(C=1|\mathbf{x}) = \frac{1}{1+\exp\left(-\theta^T\phi(\mathbf{x})\right)} = \frac{\exp\left(\theta^T\phi(\mathbf{x})\right)}{1+\exp\left(\theta^T\phi(\mathbf{x})\right)}$$

Discriminative Classification

We can now define the likelihood for each input-output pair, $(\mathbf{x}_n, y_n)$ as

$$\begin{aligned} p(C=y_n|\mathbf{x}_n,\theta) &= p(C=1|\mathbf{x}_n,\theta)^{y_n} \times (1-p(C=1|\mathbf{x}_n,\theta))^{1-y_n} \\ &= \left[\frac{1}{1+\exp\left(-\theta^T\phi(\mathbf{x}_n)\right)}\right]^{y_n}\left[\frac{1}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)}\right]^{1-y_n} \\ &= \frac{\exp\left(\theta^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)} \end{aligned}$$

Given this likelihood function, we could employ the methods we've seen so far to learn the parameters, e.g. maximum-likelihood, cross-validation, Bayesian approach.

Bayesian Approach

Let's consider the Bayesian approach for learning the parameters using this discriminative model.

We may place a Gaussian prior on the coefficients of our linear model, such that $p(\theta|\alpha) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$.

Assuming that our data is i.i.d. the likelihood follows as,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\theta) &= \prod_{n=1}^{N} p(C=y_n|\mathbf{x}_n,\theta) \\ &= \prod_{n=1}^{N} \frac{\exp\left(\theta^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)} \end{aligned}$$

Bayesian Approach

Finally, we can construct the posterior distribution over the parameters.

$$p(\theta|\mathbf{y},\mathbf{X},\alpha) = \frac{p(\mathbf{y}|\mathbf{X},\theta)p(\theta|\alpha)}{p(\mathbf{y}|\mathbf{X},\alpha)}$$

where the marginal likelihood follows as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\alpha) &= \int p(\mathbf{y}|\mathbf{X},\theta)p(\theta|\alpha)d\theta \\ &= \int \prod_{n=1}^{N} \frac{\exp\left(\theta^T\phi(\mathbf{x}_n)\right)^{y_n}}{1+\exp\left(\theta^T\phi(\mathbf{x}_n)\right)} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})d\theta \end{aligned}$$

Bayesian Approach

Unfortunately, we cannot calculate an analytic form for this potentially high dimensional integral. Some possibilities for estimating this quantity are:

- Use numerical quadrature methods (only good for very low dimensions, <5)
- Use Markov chain Monte Carlo to approximate using a Monte Carlo estimator
- Approximate the posterior with a tractable distribution, such as a Gaussian

Implementing Classification

Laplace Approximation

We may employ a Laplace approximation by assuming that the posterior may be well approximated by a multivariate Gaussian distribution.

The mean is given by the maximum a-posteriori (MAP) estimate (denoted $\boldsymbol{\theta}_{\text{MAP}}$) and the covariance (denoted C_{MAP}) is inversely proportional to the curvature of the posterior around the mean.

$$C_{\text{MAP}} = -\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log p(\mathbf{y}, \boldsymbol{\theta}_{\text{MAP}}|\mathbf{X}, \alpha)\right)^{-1}$$

Such that,

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \alpha) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)} \approx \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, C_{\text{MAP}})$$

Note that the derivatives do not depend on the marginal likelihood, since it is a constant. We can therefore optimise the MAP and calculate the curvature using only the likelihood and prior.

Laplace Approximation

Writing out the joint log likelihood and prior terms,

$$\begin{aligned}\mathcal{L} = \log p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{X}, \alpha) &= \sum_{n=1}^{N} y_n \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_n) \\ &\quad - \log(1 + \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_n))) \\ &\quad - \frac{1}{\alpha}\boldsymbol{\theta}^T\boldsymbol{\theta} - \frac{D}{2}\log(2\pi\alpha^2)\end{aligned}$$

This isn't quite as nice as the expression we encountered for linear regression!

Optimisation

The first order derivatives with respect to the parameters follow as,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} &= \sum_{n=1}^{N} y_n\boldsymbol{\phi}(\mathbf{x}_n) - p(C=1|\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n) - \frac{1}{\alpha}\boldsymbol{\theta} \\ &= \boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p} - \frac{1}{\alpha}\boldsymbol{\theta}\end{aligned}$$

where we define $\mathbf{p} = [p(C=1|\mathbf{x}_1), \ldots, p(C=1|\mathbf{x}_N)]^T$, i.e. the $N \times 1$ vector of class-membership probabilities, and $\boldsymbol{\Phi}$ is the $N \times M$ design matrix of basis functions.

Optimisation

The second order derivatives with respect to the parameters then follow as,

$$\begin{aligned}\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} &= \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T p(C=1|\mathbf{x}_n)(1 - p(C=1|\mathbf{x}_n)) - \frac{1}{\alpha}\mathbf{I} \\ &= -\boldsymbol{\Phi}^T\mathbf{V}\boldsymbol{\Phi} - \frac{1}{\alpha}\mathbf{I}\end{aligned}$$

where $\mathbf{V}$ is the $N \times N$ *diagonal* matrix defined with

$$v_{n,n} = p(C=1|\mathbf{x}_n)(1 - p(C=1|\mathbf{x}_n))$$

The covariance of the Laplace approximation is therefore,

$$C_{\text{MAP}} = \left(\boldsymbol{\Phi}^T\mathbf{V}\boldsymbol{\Phi} - \frac{1}{\alpha}\mathbf{I}\right)^{-1}$$

Newton's Method

The complex nature of the log joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ means that we cannot simply set the derivative equal to zero and solve analytically since it's a nonlinear function.

We must therefore resort to optimisation techniques, such as Newton's method.

Based on the simple idea of using a first order Taylor expansion of a function whose roots we want to find, i.e. we want to solve some function $g(\boldsymbol{\theta}) = 0$.

Truncating the Taylor expansion at 1st order and rearranging gives us the following update for the i'th iteration:

$$x_{i+1} = x_i - \frac{g(x_i)}{g'(x_i)}$$

Newton's Method

For the multivariate function we are interested in, we can employ the same approach to obtain the following update for the i'th iteration:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left(\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)^{-1}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}$$

If we plug in the expressions we just calculated for the first and second order partial derivatives, then we obtain the following optimisation algorithm,

$$\begin{aligned}\boldsymbol{\theta}_{i+1} &= \boldsymbol{\theta}_i + C_i\left(\boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p}_i - \frac{1}{\alpha}\boldsymbol{\theta}_i\right) \\ &= C_i\left(C_i^{-1}\boldsymbol{\theta}_i + \boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p}_i - \frac{1}{\alpha}\boldsymbol{\theta}_i\right) \\ &= \left(\boldsymbol{\Phi}^T\mathbf{V}_i\boldsymbol{\Phi} - \frac{1}{\alpha}\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^T(\mathbf{V}_i\boldsymbol{\Phi}\boldsymbol{\theta}_i + \mathbf{t} - \mathbf{p}_i)\end{aligned}$$

Bayesian Prediction

Once we have found the MAP estimate we can use our model to make predictions about the class of a new observation,

$$p(C=1|\mathbf{x}_{\text{new}}, \alpha, \mathbf{X}, \mathbf{y}) = \int p(C=1|\mathbf{x}_{\text{new}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \alpha)d\boldsymbol{\theta}$$

Once again, we could either approximate this integral using a Monte Carlo estimator with samples from the posterior or Laplace approximation, or we can assume the the posterior is sharply peaked around the MAP estimate, in which case,

$$\begin{aligned}p(C=1|\mathbf{x}_{\text{new}}, \alpha, \mathbf{X}, \mathbf{y}) &\approx p(C=1|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}_{\text{MAP}}, \alpha, \mathbf{X}, \mathbf{y}) \\ &= \frac{1}{1 + \exp(-\boldsymbol{\theta}_{\text{MAP}}^T\boldsymbol{\phi}(\mathbf{x}_{\text{new}}))}\end{aligned}$$

So if this discriminant function is greater than 0.5, we assign $\mathbf{x}_{\text{new}}$ to class 1. Otherwise we assign it to class 0.

Laplace Approximation

We may employ a Laplace approximation by assuming that the posterior may be well approximated by a multivariate Gaussian distribution.

The mean is given by the maximum a-posteriori (MAP) estimate (denoted $\boldsymbol{\theta}_{\mathsf{MAP}}$) and the covariance (denoted C_{MAP}) is inversely proportional to the curvature of the posterior around the mean.

$$C_{\mathsf{MAP}} = -\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}} \log p(\mathbf{y}, \boldsymbol{\theta}_{\mathsf{MAP}}|\mathbf{X}, \alpha)\right)^{-1}$$

Such that,

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \alpha) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)} \approx \mathcal{N}\left(\boldsymbol{\theta}_{\mathsf{MAP}}, C_{\mathsf{MAP}}\right)$$

Note that the derivatives do not depend on the marginal likelihood, since it is a constant. We can therefore optimise the MAP and calculate the curvature using only the likelihood and prior.

Laplace Approximation

Writing out the joint log likelihood and prior terms,

$$\begin{aligned}\mathcal{L} = \log p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{X}, \alpha) &= \sum_{n=1}^{N} y_n \boldsymbol{\theta}^T \phi(\mathbf{x}_n) \\ &\quad - \log(1 + \exp(\boldsymbol{\theta}^T \phi(\mathbf{x}_n))) \\ &\quad - \frac{1}{\alpha}\boldsymbol{\theta}^T\boldsymbol{\theta} - \frac{D}{2}\log(2\pi\alpha^2)\end{aligned}$$

This isn't quite as nice as the expression we encountered for linear regression!

Optimisation

The first order derivatives with respect to the parameters follow as,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} &= \sum_{n=1}^{N} y_n \phi(\mathbf{x}_n) - p(C=1|\mathbf{x}_n)\phi(\mathbf{x}_n) - \frac{1}{\alpha}\boldsymbol{\theta} \\ &= \boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p} - \frac{1}{\alpha}\boldsymbol{\theta}\end{aligned}$$

where we define $\mathbf{p} = [p(C=1|\mathbf{x}_1), \ldots, p(C=1|\mathbf{x}_N)]^T$, i.e. the $N \times 1$ vector of class-membership probabilities, and $\boldsymbol{\Phi}$ is the $N \times M$ design matrix of basis functions.

Optimisation

The second order derivatives with respect to the parameters then follow as,

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} &= \sum_{n=1}^{N} \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T p(C=1|\mathbf{x}_n)(1-p(C=1|\mathbf{x}_n)) - \frac{1}{\alpha}\mathbf{I} \\ &= -\mathbf{\Phi}^T \mathbf{V} \mathbf{\Phi} - \frac{1}{\alpha}\mathbf{I}\end{aligned}$$

where $\mathbf{V}$ is the $N \times N$ *diagonal* matrix defined with

$$v_{n,n} = p(C=1|\mathbf{x}_n)(1-p(C=1|\mathbf{x}_n))$$

The covariance of the Laplace approximation is therefore,

$$C_{\text{MAP}} = \left(\mathbf{\Phi}^T \mathbf{V} \mathbf{\Phi} - \frac{1}{\alpha}\mathbf{I} \right)^{-1}$$

Newton's Method

The complex nature of the log joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ means that we cannot simply set the derivative equal to zero and solve analytically since it's a nonlinear function.

We must therefore resort to optimisation techniques, such as Newton's method.

Based on the simple idea of using a first order Taylor expansion of a function whose roots we want to find, i.e. we want to solve some function $g(\boldsymbol{\theta}) = 0$.

Truncating the Taylor expansion at 1st order and rearranging gives us the following update for the i'th iteration:

$$x_{i+1} = x_i - \frac{g(x_i)}{g'(x_i)}$$

Newton's Method

For the multivariate function we are interested in, we can employ the same approach to obtain the following update for the i'th iteration:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right)^{-1} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$$

If we plug in the expressions we just calculated for the first and second order partial derivatives, then we obtain the following optimisation algorithm,

$$\begin{aligned}
\boldsymbol{\theta}_{i+1} &= \boldsymbol{\theta}_i + C_i \left(\boldsymbol{\Phi}^T \mathbf{y} - \boldsymbol{\Phi}^T \mathbf{p}_i - \frac{1}{\alpha} \boldsymbol{\theta}_i \right) \\
&= C_i \left(C_i^{-1} \boldsymbol{\theta}_i + \boldsymbol{\Phi}^T \mathbf{y} - \boldsymbol{\Phi}^T \mathbf{p}_i - \frac{1}{\alpha} \boldsymbol{\theta}_i \right) \\
&= \left(\boldsymbol{\Phi}^T \mathbf{V}_i \boldsymbol{\Phi} - \frac{1}{\alpha} \mathbf{I} \right)^{-1} \boldsymbol{\Phi}^T (\mathbf{V}_i \boldsymbol{\Phi} \boldsymbol{\theta}_i + \mathbf{t} - \mathbf{p}_i)
\end{aligned}$$

Bayesian Prediction

Once we have found the MAP estimate we can use our model to make predictions about the class of a new observation,

$$p(C=1|\mathbf{x}_{\text{new}}, \alpha, \mathbf{X}, \mathbf{y}) = \int p(C=1|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \alpha) d\boldsymbol{\theta}$$

Once again, we could either approximate this integral using a Monte Carlo estimator with samples from the posterior or Laplace approximation, or we can assume the the posterior is sharply peaked around the MAP estimate, in which case,

$$\begin{aligned} p(C=1|\mathbf{x}_{\text{new}}, \alpha, \mathbf{X}, \mathbf{y}) &\approx p(C=1|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}_{\text{MAP}}, \alpha, \mathbf{X}, \mathbf{y}) \\ &= \frac{1}{1+\exp(-\boldsymbol{\theta}_{\text{MAP}}^T \phi(\mathbf{x}_{\text{new}}))} \end{aligned}$$

So if this discriminant function is greater than 0.5, we assign $\mathbf{x}_{\text{new}}$ to class 1. Otherwise we assign it to class 0.

Implementing Classification

Laplace Approximation

We may employ a Laplace approximation by assuming that the posterior may be well approximated by a multivariate Gaussian distribution.

The mean is given by the maximum a-posteriori (MAP) estimate (denoted $\boldsymbol{\theta}_{\text{MAP}}$) and the covariance (denoted C_{MAP}) is inversely proportional to the curvature of the posterior around the mean.

$$C_{\text{MAP}} = -\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log p(\mathbf{y}, \boldsymbol{\theta}_{\text{MAP}}|\mathbf{X}, \alpha)\right)^{-1}$$

Such that,

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \alpha) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)} \approx \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, C_{\text{MAP}})$$

Note that the derivatives do not depend on the marginal likelihood, since it is a constant. We can therefore optimise the MAP and calculate the curvature using only the likelihood and prior.

Laplace Approximation

Writing out the joint log likelihood and prior terms,

$$\begin{aligned}\mathcal{L} = \log p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{X}, \alpha) &= \sum_{n=1}^{N} y_n \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_n) \\ &\quad - \log(1 + \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_n))) \\ &\quad - \frac{1}{\alpha}\boldsymbol{\theta}^T\boldsymbol{\theta} - \frac{D}{2}\log(2\pi\alpha^2)\end{aligned}$$

This isn't quite as nice as the expression we encountered for linear regression!

Optimisation

The first order derivatives with respect to the parameters follow as,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} &= \sum_{n=1}^{N} y_n\boldsymbol{\phi}(\mathbf{x}_n) - p(C=1|\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n) - \frac{1}{\alpha}\boldsymbol{\theta} \\ &= \boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p} - \frac{1}{\alpha}\boldsymbol{\theta}\end{aligned}$$

where we define $\mathbf{p} = [p(C=1|\mathbf{x}_1), \ldots, p(C=1|\mathbf{x}_N)]^T$, i.e. the $N \times 1$ vector of class-membership probabilities, and $\boldsymbol{\Phi}$ is the $N \times M$ design matrix of basis functions.

Optimisation

The second order derivatives with respect to the parameters then follow as,

$$\begin{aligned}\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} &= \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T p(C=1|\mathbf{x}_n)(1 - p(C=1|\mathbf{x}_n)) - \frac{1}{\alpha}\mathbf{I} \\ &= -\boldsymbol{\Phi}^T\mathbf{V}\boldsymbol{\Phi} - \frac{1}{\alpha}\mathbf{I}\end{aligned}$$

where $\mathbf{V}$ is the $N \times N$ *diagonal* matrix defined with

$$v_{n,n} = p(C=1|\mathbf{x}_n)(1 - p(C=1|\mathbf{x}_n))$$

The covariance of the Laplace approximation is therefore,

$$C_{\text{MAP}} = \left(\boldsymbol{\Phi}^T\mathbf{V}\boldsymbol{\Phi} - \frac{1}{\alpha}\mathbf{I}\right)^{-1}$$

Newton's Method

The complex nature of the log joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ means that we cannot simply set the derivative equal to zero and solve analytically since it's a nonlinear function.

We must therefore resort to optimisation techniques, such as Newton's method.

Based on the simple idea of using a first order Taylor expansion of a function whose roots we want to find, i.e. we want to solve some function $g(\boldsymbol{\theta}) = 0$.

Truncating the Taylor expansion at 1st order and rearranging gives us the following update for the i'th iteration:

$$x_{i+1} = x_i - \frac{g(x_i)}{g'(x_i)}$$

Newton's Method

For the multivariate function we are interested in, we can employ the same approach to obtain the following update for the i'th iteration:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left(\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)^{-1}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}$$

If we plug in the expressions we just calculated for the first and second order partial derivatives, then we obtain the following optimisation algorithm,

$$\begin{aligned}\boldsymbol{\theta}_{i+1} &= \boldsymbol{\theta}_i + C_i\left(\boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p}_i - \frac{1}{\alpha}\boldsymbol{\theta}_i\right) \\ &= C_i\left(C_i^{-1}\boldsymbol{\theta}_i + \boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\mathbf{p}_i - \frac{1}{\alpha}\boldsymbol{\theta}_i\right) \\ &= \left(\boldsymbol{\Phi}^T\mathbf{V}_i\boldsymbol{\Phi} - \frac{1}{\alpha}\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^T(\mathbf{V}_i\boldsymbol{\Phi}\boldsymbol{\theta}_i + \mathbf{t} - \mathbf{p}_i)\end{aligned}$$

Bayesian Prediction

Once we have found the MAP estimate we can use our model to make predictions about the class of a new observation,

$$p(C=1|\mathbf{x}_{\text{new}}, \alpha, \mathbf{X}, \mathbf{y}) = \int p(C=1|\mathbf{x}_{\text{new}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \alpha)d\boldsymbol{\theta}$$

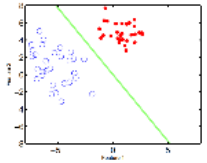
Once again, we could either approximate this integral using a Monte Carlo estimator with samples from the posterior or Laplace approximation, or we can assume the the posterior is sharply peaked around the MAP estimate, in which case,

$$\begin{aligned}p(C=1|\mathbf{x}_{\text{new}}, \alpha, \mathbf{X}, \mathbf{y}) &\approx p(C=1|\mathbf{x}_{\text{new}}, \boldsymbol{\theta}_{\text{MAP}}, \alpha, \mathbf{X}, \mathbf{y}) \\ &= \frac{1}{1 + \exp(-\boldsymbol{\theta}_{\text{MAP}}^T\boldsymbol{\phi}(\mathbf{x}_{\text{new}}))}\end{aligned}$$

So if this discriminant function is greater than 0.5, we assign $\mathbf{x}_{\text{new}}$ to class 1. Otherwise we assign it to class 0.

Bayesian Discriminative Classification Example

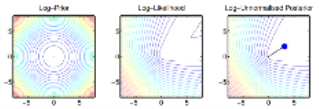
Classification Example



The blue dots represent class C = 1 and the red dots class C = 0.

The decision boundary for a 1st order polynomial linear model is shown in green, with the MAP estimate calculated using Newton's method.

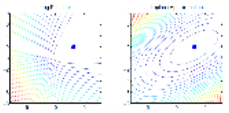
Bayesian Distribution



The prior, log-likelihood and unnormalised posterior distributions. Note that the log-likelihood is noticeably non-Gaussian.

The blue dot shows the MAP estimate using Newton's method, which was initialise at the point (0,0).

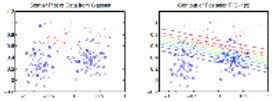
Laplace Approximation



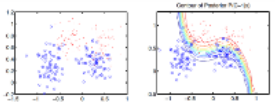
Here we observe the difference between the true log posterior and the optimal Laplace approximation.

In this case it is not a great fit - we might therefore get better results using Markov chain Monte Carlo and a Monte Carlo estimator!

Posterior Decisions

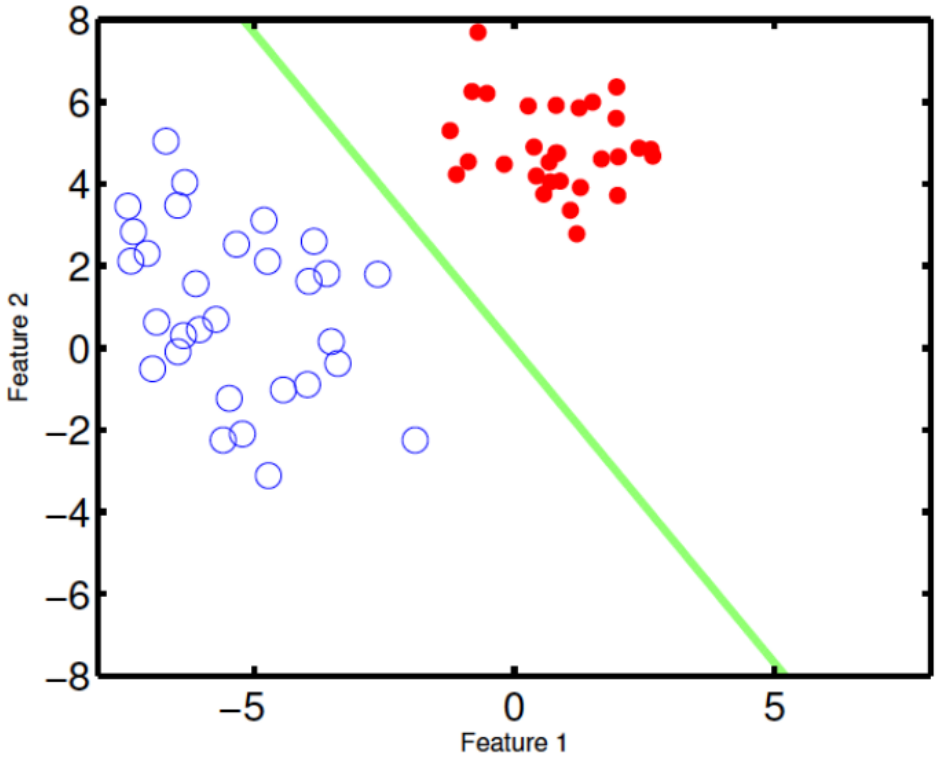


The posterior distribution using a 1st order polynomial model.



The posterior distribution using a 3rd order polynomial model.

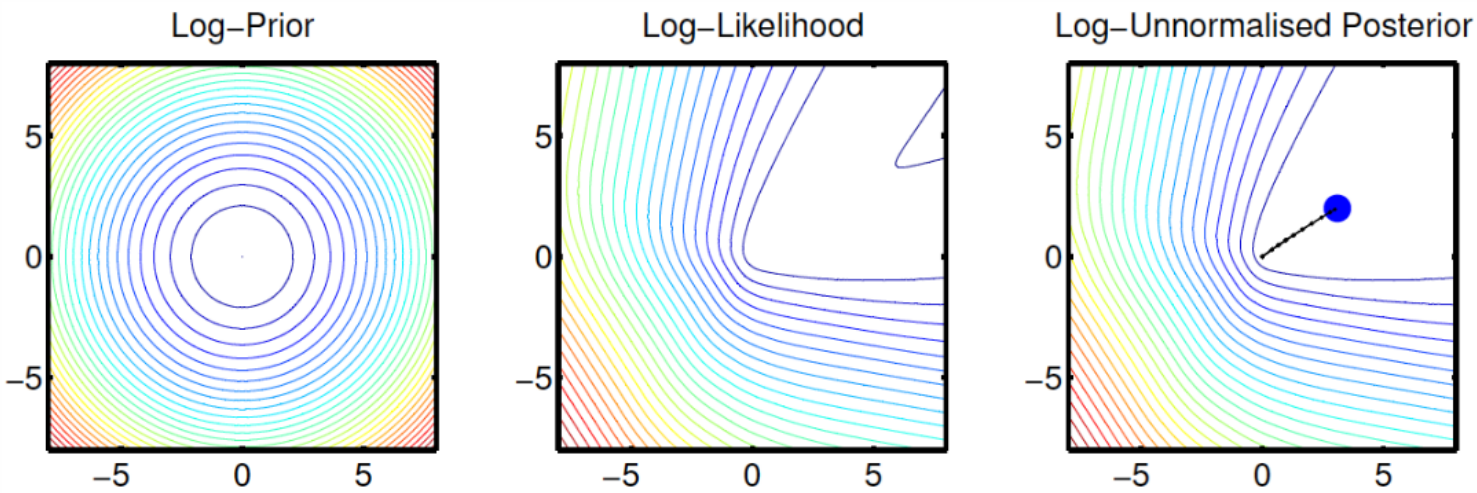
Classification Example



The blue dots represent class C = 1 and the red dots class C = 0.

The decision boundary for a 1st order polynomial linear model is shown in green, with the MAP estimate calculated using Newton's method.

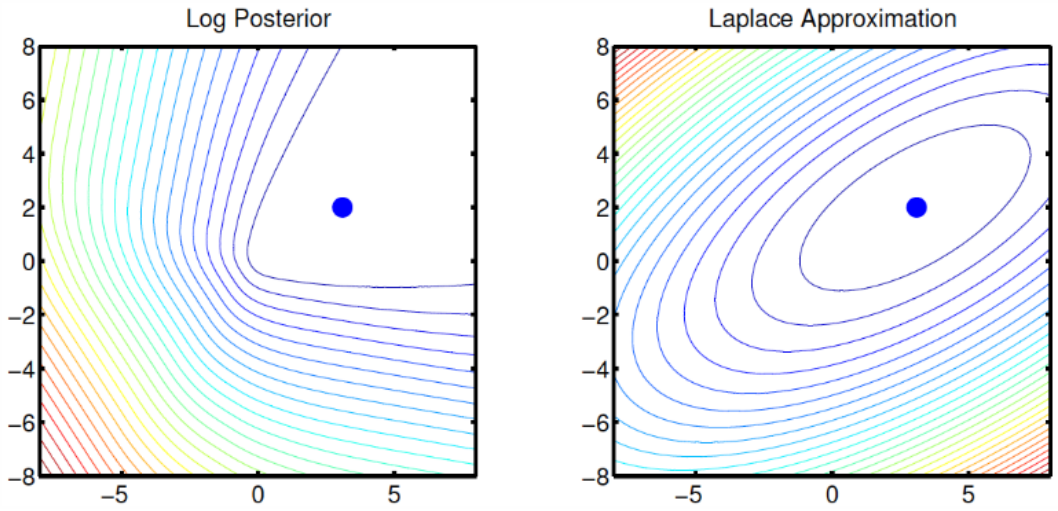
Bayesian Distribution



The prior, log-likelihood and unnormalised posterior distributions. Note that the log-likelihood is noticeably non-Gaussian.

The blue dot shows the MAP estimate using Newton's method, which was initialise at the point (0,0).

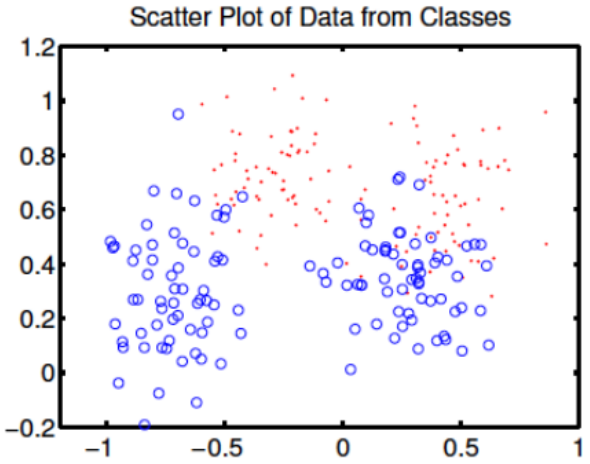
Laplace Approximation

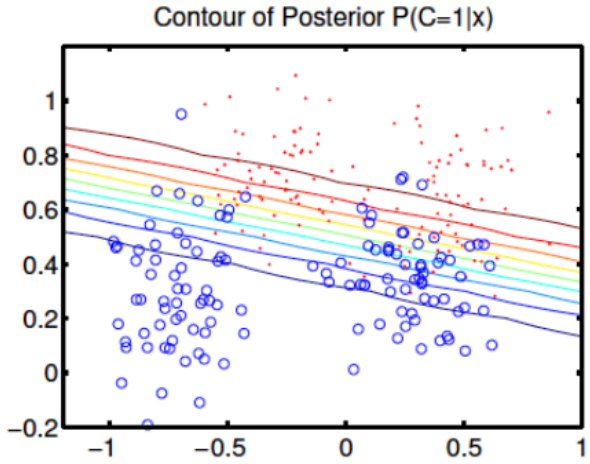


Here we observe the difference between the true log posterior and the optimal Laplace approximation.

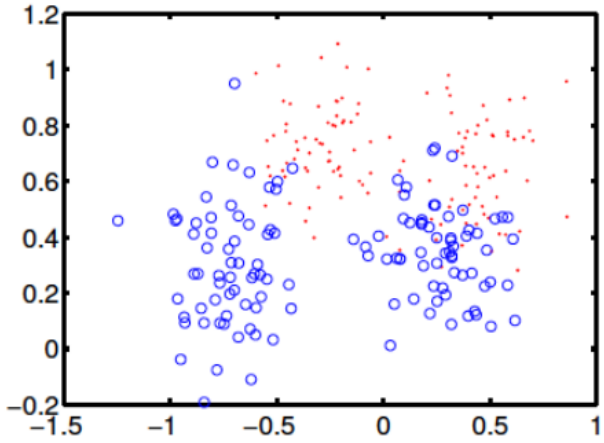
In this case it is not a great fit - we might therefore get better results using Markov chain Monte Carlo and a Monte Carlo estimator!

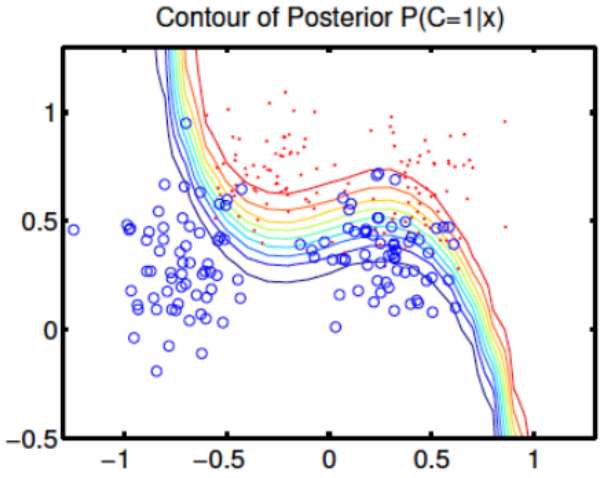
Posterior Decisions





The posterior distribution using a 1st order polynomial model.





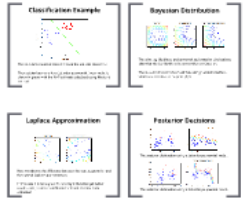
The posterior distribution using a 3rd order polynomial model.

Introduction to Classification

Discriminative Classification

Implementing Classification

Bayesian Discriminative Classification Example



Further Reading

- Chapter 18 - Bayesian Linear Models
- Chapter 18.2 - Classification

Bayesian Reasoning and Machine Learning
by David Barber

- Chapter 8 - Logistic Regression

Machine Learning: A Probabilistic Perspective
by Kevin Murphy

Further Reading

- Chapter 18 - Bayesian Linear Models
- Chapter 18.2 - Classification

Bayesian Reasoning and Machine Learning
by David Barber

- Chapter 8 - Logistic Regression

Machine Learning: A Probabilistic Perspective
by Kevin Murphy

M5MS10
Machine Learning

Spring 2018

Lecture 3

Dr Ben Calderhead

b.calderhead@imperial.ac.uk