Machine Learning

Lecture. 7.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Classification

UNIVERSITY
of
GLASGOW

- A large class of problems which Machine Learning techniques are applied to are classification problems

Classification

UNIVERSITY
of
GLASGOW

- A large class of problems which Machine Learning techniques are applied to are classification problems
- Object Location - Image Processing

Classification

UNIVERSITY *of* GLASGOW

- A large class of problems which Machine Learning techniques are applied to are classification problems
- Object Location - Image Processing
- Protein Fold Prediction - Bioinformatics

Classification

UNIVERSITY
of
GLASGOW

- A large class of problems which Machine Learning techniques are applied to are classification problems
- Object Location - Image Processing
- Protein Fold Prediction - Bioinformatics
- Gesture Recognition - HCI

Classification

UNIVERSITY
of
GLASGOW

- A large class of problems which Machine Learning techniques are applied to are classification problems
- Object Location - Image Processing
- Protein Fold Prediction - Bioinformatics
- Gesture Recognition - HCI
- Intrusion Detection - Networks & Systems

Classification

UNIVERSITY of GLASGOW

- A large class of problems which Machine Learning techniques are applied to are classification problems
- Object Location - Image Processing
- Protein Fold Prediction - Bioinformatics
- Gesture Recognition - HCI
- Intrusion Detection - Networks & Systems
- All are essentially classification problems

Example

As a simple example lets try and build a classifier which will predict whether a person is male or female based on their measured height alone.

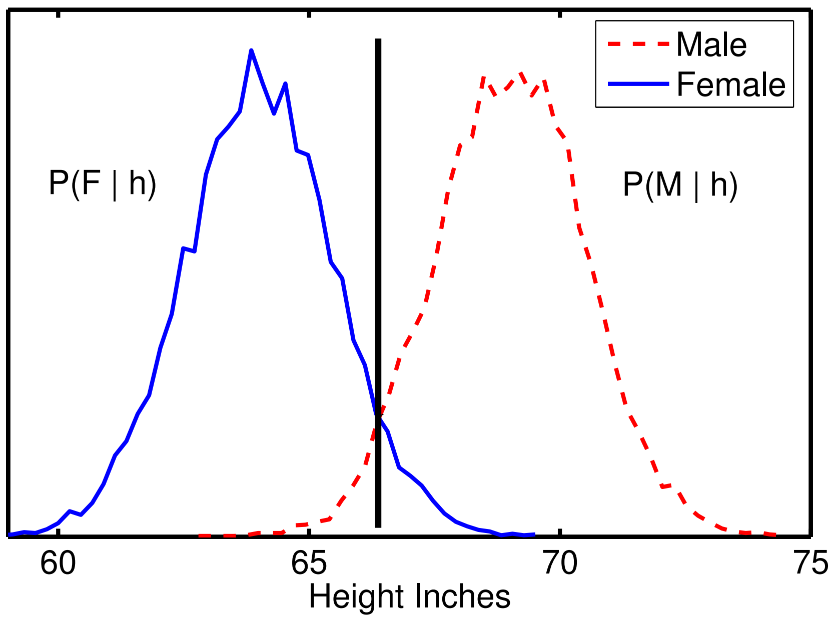

Figure 1: The distributions of measured height for both males and females in a population.

Class Priors

UNIVERSITY *of* GLASGOW

- The class variable C will take on two values so we can encode male by the value 1 and female by the value 0.

Class Priors

- The class variable C will take on two values so we can encode male by the value 1 and female by the value 0.
- Now within the general population there is an approximate equal number of male and females.

Class Priors

UNIVERSITY *of* GLASGOW

- The class variable C will take on two values so we can encode male by the value 1 and female by the value 0.
- Now within the general population there is an approximate equal number of male and females.
- In that case the probability of class male occurring will be defined simply as $P(C = 1)$ and the probability of class female occuring will be $P(C = 0)$.

Class Priors

- The class variable C will take on two values so we can encode male by the value 1 and female by the value 0.
- Now within the general population there is an approximate equal number of male and females.
- In that case the probability of class male occurring will be defined simply as $P(C = 1)$ and the probability of class female occuring will be $P(C = 0)$.
- Now these probabilities are set **prior** to making any measurements and hence are called the **prior probabilities** of class membership.

Class Priors

- If these are well balanced i.e. $P(C = 0) = P(C = 1) = 0.5$ then it is equally likely to observe either class.

Class Priors

- If these are well balanced i.e. $P(C = 0) = P(C = 1) = 0.5$ then it is equally likely to observe either class.
- However in applications such as medical diagnostics or intrusion detection the prior probabilities of one class e.g. network intrusion or cancer are much smaller than the other e.g. normal traffic or not cancer.

Class Priors

- If these are well balanced i.e. $P(C = 0) = P(C = 1) = 0.5$ then it is equally likely to observe either class.
- However in applications such as medical diagnostics or intrusion detection the prior probabilities of one class e.g. network intrusion or cancer are much smaller than the other e.g. normal traffic or not cancer.
- In this case then we can make a prediction before seeing any data that is more likely to be correct based on the prior probabilities alone.

Class Conditioned Likelihood

- There will be a natural distribution of the height of males and females, so in other words there will be a **class conditional distribution** of the measured features, in this case height.

Class Conditioned Likelihood

UNIVERSITY *of* GLASGOW

- There will be a natural distribution of the height of males and females, so in other words there will be a **class conditional distribution** of the measured features, in this case height.
- We can write these class conditional distributions as $p(h|C = 1)$ and $p(h|C = 0)$ form male and female classes respectively.

Class Conditioned Likelihood

- There will be a natural distribution of the height of males and females, so in other words there will be a **class conditional distribution** of the measured features, in this case height.
- We can write these class conditional distributions as $p(h|C = 1)$ and $p(h|C = 0)$ form male and female classes respectively.
- This likelihood can be used to obtain a posterior over the class variable.

Class Posterior

From Bayes rule can obtain posterior probability of class membership by noting

$$P(h, C = 1) = p(h|C = 1)P(C = 1) = P(C = 1|h)p(h)$$

and so

$$P(C = 1|h) = \frac{p(h|C = 1)P(C = 1)}{p(h)}$$

and the marginal likelihood of our measurement, $p(h)$, is the probability of measuring a height h irrespective of the class and so

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

which means that the class posteriors will also sum to one, $P(C = 1|h) + P(C = 1|h) = 1$.

Discriminant Functions

- The first thing to notice is that there is a distinct difference in the location of the distributions and they can be separated to a large extent (males are typically taller than females)

Discriminant Functions

UNIVERSITY *of* GLASGOW

- The first thing to notice is that there is a distinct difference in the location of the distributions and they can be separated to a large extent (males are typically taller than females)
- However there is a region where the two distributions overlap and it is here that classification errors can be made

Discriminant Functions

- The first thing to notice is that there is a distinct difference in the location of the distributions and they can be separated to a large extent (males are typically taller than females)
- However there is a region where the two distributions overlap and it is here that classification errors can be made
- The region of intersection where $P(C = 1|h) = P(C = 0|h)$ is important as it defines our decision boundary

Discriminant Functions

UNIVERSITY of GLASGOW

- The first thing to notice is that there is a distinct difference in the location of the distributions and they can be separated to a large extent (males are typically taller than females)
- However there is a region where the two distributions overlap and it is here that classification errors can be made
- The region of intersection where $P(C = 1|h) = P(C = 0|h)$ is important as it defines our decision boundary
- If we make a measurement of 69 inches then $P(C = 1|h) > P(C = 0|h)$ there is some probability that this is a rather tall female, to minimise unavoidable errors then decision should be based on the largest posterior probability.

Discriminant Functions

- We can then define a **discriminant function** based on our posterior probabilities

Discriminant Functions

UNIVERSITY *of* GLASGOW

- We can then define a **discriminant function** based on our posterior probabilities
- One such function could be the ratio of posterior probabilities for both classes

Discriminant Functions

- We can then define a **discriminant function** based on our posterior probabilities
- One such function could be the ratio of posterior probabilities for both classes
- If we take the logarithm of this ratio then the general discriminant function

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

would define the rules that h would be assigned to $C = 1$ (male) if $f(h) > 0$ and if $f(h) < 0$ the assignment would be to $C = 0$ (female)

Discriminative Classification

- Use general notation $\mathbf{x} = [x_1, \cdots, x_D]^\mathsf{T}$ representing D-dimensional vector of D features available for classification purposes.

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

Discriminative Classification

UNIVERSITY of GLASGOW

- Use general notation $\mathbf{x} = [x_1, \cdots, x_D]^\mathsf{T}$ representing D-dimensional vector of D features available for classification purposes.

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

- Ratio $P(C = 1|\mathbf{x})$ & $P(C = 0|\mathbf{x})$ lies on positive real line i.e. $[0 + \infty)$ so log-likelihood ratio will take values between $-\infty$ and $+\infty$.

Discriminative Classification

UNIVERSITY of GLASGOW

- Use general notation $\mathbf{x} = [x_1, \cdots, x_D]^\mathsf{T}$ representing D-dimensional vector of D features available for classification purposes.

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

- Ratio $P(C = 1|\mathbf{x})$ & $P(C = 0|\mathbf{x})$ lies on positive real line i.e. $[0 + \infty)$ so log-likelihood ratio will take values between $-\infty$ and $+\infty$.
- Model ratio using a linear-model, now employ explicit basis expansion of input $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}, \cdots, \phi_M(\mathbf{x})]^\mathsf{T}$, and each $\phi_m(\mathbf{x})$ defines the m'th basis function applied to the data vector $\mathbf{x}$.

Discriminative Classification

UNIVERSITY
of
GLASGOW

- Back to the log-likelihood ratio and our linear model of it

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

Discriminative Classification

UNIVERSITY *of* GLASGOW

- Back to the log-likelihood ratio and our linear model of it

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

- As $P(C=1|\mathbf{x}) + P(C=0|\mathbf{x}) = 1$ then a tiny bit of algebra shows that

$$\begin{aligned} P(C=1|\mathbf{x}) &= \frac{1}{1+\exp\left(-\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})\right)} \\ &= \frac{\exp\left(\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})\right)}{1+\exp\left(\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})\right)} \end{aligned}$$

Discriminative Classification

UNIVERSITY *of* GLASGOW

- The likelihood for each data point (input-output pair) $(\mathbf{x}_n, t_n)$ will simply be the posterior probability $P(C = t_n | \mathbf{x}_n)$.

Discriminative Classification

UNIVERSITY of GLASGOW

- The likelihood for each data point (input-output pair) $(\mathbf{x}_n, t_n)$ will simply be the posterior probability $P(C = t_n|\mathbf{x}_n)$.
- Now we can write the likelihood component for each n as $P(C = t_n|\mathbf{x}_n, \mathbf{w})$ which equals

$$
\begin{aligned}
& P(C = 1|\mathbf{x}_n, \mathbf{w})^{t_n} \times (1 - P(C = 1|\mathbf{x}_n, \mathbf{w}))^{1-t_n} \\
= \quad & \left[\frac{1}{1 + \exp\left(-\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n)\right)}\right]^{t_n} \left[\frac{1}{1 + \exp\left(\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n)\right)}\right]^{1-t_n} \\
= \quad & \frac{\exp\left(\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n)\right)^{t_n}}{1 + \exp\left(\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n)\right)}
\end{aligned}
$$

Bayesian Classification

- Let us be bold and take a Bayesian viewpoint straightaway (you know it makes sense!) so we will place a Gaussian prior on our coefficients such that $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ and we assume that each t_n is sampled i.i.d (remember this from last week?) in which case our likelihood will be

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} P(C = t_n|\mathbf{x}_n, \mathbf{w})$$

Bayesian Classification

UNIVERSITY *of* GLASGOW

- Let us be bold and take a Bayesian viewpoint straightaway (you know it makes sense!) so we will place a Gaussian prior on our coefficients such that $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ and we assume that each t_n is sampled i.i.d (remember this from last week?) in which case our likelihood will be

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} P(C = t_n|\mathbf{x}_n, \mathbf{w})$$

Bayesian Classification

UNIVERSITY *of* GLASGOW

- Now that we are all good Bayesians we immediately want to define the posterior over the parameters and so we need the joint-likelihood formed by the likelihood and the prior

$$\begin{aligned} p(\mathbf{t}, \mathbf{w} | \mathbf{X}, \alpha) &= p(\mathbf{t} | \mathbf{X}, \mathbf{w}) p(\mathbf{w} | \alpha) \\ &= \prod_{n=1}^{N} \frac{\exp\left(\mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n)\right)^{t_n}}{1 + \exp\left(\mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n)\right)} \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \alpha^{-1} \mathbf{I}) \end{aligned}$$

Bayesian Classification

UNIVERSITY
of
GLASGOW

- Now that we are all good Bayesians we immediately want to define the posterior over the parameters and so we need the joint-likelihood formed by the likelihood and the prior

$$
\begin{aligned}
p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \alpha) &= p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha) \\
&= \prod_{n=1}^{N} \frac{\exp\left(\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x}_n)\right)^{t_n}}{1 + \exp\left(\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x}_n)\right)} \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \alpha^{-1}\mathbf{I})
\end{aligned}
$$

Bayesian Classification

- To obtain our posterior we require the following

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha)\frac{1}{p(\mathbf{t}|\mathbf{X}, \alpha)}$$

where the marginal likelihood

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

Bayesian Classification

UNIVERSITY of GLASGOW

- The multi-dimensional integral cannot be computed analytically.

Bayesian Classification

UNIVERSITY
of
GLASGOW

- The multi-dimensional integral cannot be computed analytically.
- Unlike the regression problem where a fully analytic expression for the posterior was available in the classification setting we run into some small degree of difficulty.

Bayesian Classification

UNIVERSITY
of
GLASGOW

- The multi-dimensional integral cannot be computed analytically.
- Unlike the regression problem where a fully analytic expression for the posterior was available in the classification setting we run into some small degree of difficulty.
- Compute integral numerically using MCMC

Bayesian Classification

UNIVERSITY *of* GLASGOW

- The multi-dimensional integral cannot be computed analytically.
- Unlike the regression problem where a fully analytic expression for the posterior was available in the classification setting we run into some small degree of difficulty.
- Compute integral numerically using MCMC
- Approximate the posterior with a tractable distribution - multivariate Gaussian

Laplace Approximation

UNIVERSITY
of
GLASGOW

- Assume the posterior is multivariate Gaussian.

Laplace Approximation

UNIVERSITY of GLASGOW

- Assume the posterior is multivariate Gaussian.
- With mean value at maximum of posterior.

Laplace Approximation

- Assume the posterior is multivariate Gaussian.
- With mean value at maximum of posterior.
- With covariance proportional to curvature of posterior around mean

Laplace Approximation

UNIVERSITY *of* GLASGOW

- Assume the posterior is multivariate Gaussian.
- With mean value at maximum of posterior.
- With covariance proportional to curvature of posterior around mean
- In other words if we define the parameters at the maximum of the posterior as $\mathbf{w}_{MAP}$ and the covariance of the approximation as $\mathbf{C}$, where

$$\mathbf{C} = -\left(\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}^{\mathsf{T}}} \log p(\mathbf{t}, \mathbf{w} | \mathbf{X}, \alpha)\right)^{-1}$$

where the right-hand side is computed at the *MAP* value $\mathbf{w}_{MAP}$

Laplace Approximation

- In which case we can write

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha)\frac{1}{p(\mathbf{t}|\mathbf{X}, \alpha)} \approx \mathcal{N}_{\mathbf{w}}(\mathbf{w}_{MAP}, \mathbf{C})$$

Laplace Approximation

UNIVERSITY *of* GLASGOW

- In which case we can write

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\alpha) = p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\alpha)\frac{1}{p(\mathbf{t}|\mathbf{X},\alpha)} \approx \mathcal{N}_{\mathbf{w}}(\mathbf{w}_{MAP},\mathbf{C})$$

- Need to somehow estimate the *Maximum a Posteriori* parameter value as well as compute the curvature of the posterior at that point

Laplace Approximation

- In which case we can write

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha)\frac{1}{p(\mathbf{t}|\mathbf{X}, \alpha)} \approx \mathcal{N}_{\mathbf{w}}(\mathbf{w}_{MAP}, \mathbf{C})$$

- Need to somehow estimate the *Maximum a Posteriori* parameter value as well as compute the curvature of the posterior at that point
- Note that we need to find the parameter values which maximise the posterior and we can do this by maximising the logarithm of the joint likelihood as the normalising term (the marginal) does not depend on the parameters

Laplace Approximation

UNIVERSITY of GLASGOW

- So as before let us write out the logarithm of the joint likelihood which follows as

$$
\begin{aligned}
\mathcal{L} = \log p(\mathbf{t}, \mathbf{w} | \mathbf{X}, \alpha) &= \sum_{n=1}^{N} t_n \mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n) \\
&- \log\left(1 + \exp\left(\mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n)\right)\right) \\
&- \frac{1}{\alpha} \mathbf{w}^{\mathsf{T}} \mathbf{w} - \frac{D}{2} \log(2\pi\alpha^2)
\end{aligned}
$$

this is clearly not as nice an expression as we had for the linear regression models we have already met

Laplace Approximation

UNIVERSITY *of* GLASGOW

- Take first and second derivatives with respect to all the parameter values $\mathbf{w}$.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n) - P(C=1|\mathbf{x}_n)\phi(\mathbf{x}_n) - \frac{1}{\alpha}\mathbf{w}$$
$$= \mathbf{\Phi}^{\mathsf{T}}\mathbf{t} - \mathbf{\Phi}^{\mathsf{T}}\mathbf{p} - \frac{1}{\alpha}\mathbf{w}$$

where the $N \times 1$ vector of class-membership probabilities is defined as $\mathbf{p} = [P(C=1|\mathbf{x}_1), \cdots, P(C=1|\mathbf{x}_N)]^{\mathsf{T}}$ and the $N \times M$ matrix $\mathbf{\Phi}$ composed of basis functions

Laplace Approximation

- The second-derivatives follows as before $\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}}$

$$
\begin{aligned}
&= -\sum_{n=1}^{N} \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^\mathsf{T} P(C=1|\mathbf{x}_n)\left(1 - P(C=1|\mathbf{x}_n)\right) - \\
&= -\mathbf{\Phi}^\mathsf{T}\mathbf{V}\mathbf{\Phi} - \frac{1}{\alpha}\mathbf{I}
\end{aligned}
$$

where $\mathbf{V}$ is an $N \times N$ dimensional diagonal matrix defined as $diag(v_{11}, \cdots, v_{NN})$

Laplace Approximation

UNIVERSITY *of* GLASGOW

- The second-derivatives follows as before $\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^{\mathsf{T}}}$

$$
\begin{aligned}
&= -\sum_{n=1}^{N} \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\mathsf{T}} P(C=1|\mathbf{x}_n)\left(1 - P(C=1|\mathbf{x}_n)\right) - \\
&= -\mathbf{\Phi}^{\mathsf{T}}\mathbf{V}\mathbf{\Phi} - \frac{1}{\alpha}\mathbf{I}
\end{aligned}
$$

  where $\mathbf{V}$ is an $N \times N$ dimensional diagonal matrix defined as $diag(v_{11}, \cdots, v_{NN})$

- Now then we can define the covariance matrix of the *approximate* posterior as

$$
\mathbf{C} = \left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{V}\mathbf{\Phi} + \frac{1}{\alpha}\mathbf{I}\right)^{-1}
$$

Newton Optimisation

- The *MAP* value for the parameters does not follow in the nice closed form by setting the gradients to zero and solving for $\mathbf{w}$ as in the standard linear regression model as each element of the vector $\mathbf{p}$ i.e. $P(C = 1|\mathbf{x}_n)$ is itself a nonlinear function of $\mathbf{w}$. We now need to resort to optimisation techniques.

Newton Optimisation

UNIVERSITY *of* GLASGOW

- The *MAP* value for the parameters does not follow in the nice closed form by setting the gradients to zero and solving for $\mathbf{w}$ as in the standard linear regression model as each element of the vector $\mathbf{p}$ i.e. $P(C = 1|\mathbf{x}_n)$ is itself a nonlinear function of $\mathbf{w}$. We now need to resort to optimisation techniques.
- We need to find the parameter values $\mathbf{w}_{MAP}$ which will yield the maximum so make moves in parameter space which will yield the largest change in the criterion to be maximised, in this case the joint likelihood.

Newton Optimisation

UNIVERSITY *of* GLASGOW

- To find the roots of functions $f(x) = 0$ from an initial guess of x_0. The next guess is given as

$$x_{k+1} \leftarrow x_k - \frac{f(x_k)}{f'(x_k)}$$

Newton Optimisation

UNIVERSITY *of* GLASGOW

- To find the roots of functions $f(x) = 0$ from an initial guess of x_0. The next guess is given as

$$x_{k+1} \leftarrow x_k - \frac{f(x_k)}{f'(x_k)}$$

- Seek stationary points $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0}$ so take Newton method to find roots of a single variable function and extend to multiple variables

$$\mathbf{w} \leftarrow \mathbf{w} - \left(\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^{\mathsf{T}}} \right)^{-1} \frac{\partial \mathcal{L}}{\partial \mathbf{w}}$$

Newton Optimisation

UNIVERSITY of GLASGOW

- Employing our expressions for the 1st & 2nd derivatives then

$$
\begin{aligned}
\mathbf{w} &\leftarrow \mathbf{w} + \mathbf{C}\left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{t} - \mathbf{\Phi}^{\mathsf{T}}\mathbf{p} - \frac{1}{\alpha}\mathbf{w}\right) \\
&= \mathbf{C}\left(\mathbf{C}^{-1}\mathbf{w} + \mathbf{\Phi}^{\mathsf{T}}\mathbf{t} - \mathbf{\Phi}^{\mathsf{T}}\mathbf{p} - \frac{1}{\alpha}\mathbf{w}\right) \\
&= \left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{V}\mathbf{\Phi} + \frac{1}{\alpha}\mathbf{I}\right)^{-1}\mathbf{\Phi}^{\mathsf{T}}\left(\mathbf{V}\mathbf{\Phi}\mathbf{w} + \mathbf{t} - \mathbf{p}\right)
\end{aligned}
$$

Newton Optimisation

UNIVERSITY *of* GLASGOW

- Employing our expressions for the 1st & 2nd derivatives then

$$
\begin{aligned}
\mathbf{w} \quad &\leftarrow \quad \mathbf{w} + \mathbf{C}\left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{t} - \mathbf{\Phi}^{\mathsf{T}}\mathbf{p} - \frac{1}{\alpha}\mathbf{w}\right) \\
&= \quad \mathbf{C}\left(\mathbf{C}^{-1}\mathbf{w} + \mathbf{\Phi}^{\mathsf{T}}\mathbf{t} - \mathbf{\Phi}^{\mathsf{T}}\mathbf{p} - \frac{1}{\alpha}\mathbf{w}\right) \\
&= \quad \left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{V}\mathbf{\Phi} + \frac{1}{\alpha}\mathbf{I}\right)^{-1}\mathbf{\Phi}^{\mathsf{T}}\left(\mathbf{V}\mathbf{\Phi}\mathbf{w} + \mathbf{t} - \mathbf{p}\right)
\end{aligned}
$$

- At each step $\mathbf{w}$ is updated, using new values $\mathbf{w}$ elements of both $\mathbf{p}$ and $\mathbf{V}$ are updated then next Newton step re-applied until convergence

Laplace Demo

UNIVERSITY *of* GLASGOW

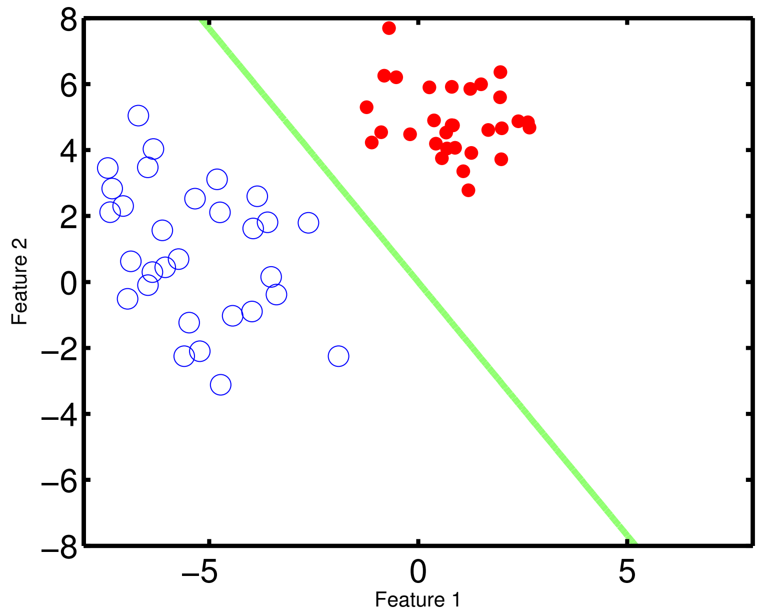

Figure 2: The blue circles are examples from class $C = 0$ and the solid red dots are examples from class $C = 1$. The green line shows the decision boundary $P(C = 1|\mathbf{x}) = 0.5$ obtained from the estimated $\mathbf{w}_{MAP}$ using the Newton routine described above.

Laplace Demo

UNIVERSITY *of* GLASGOW

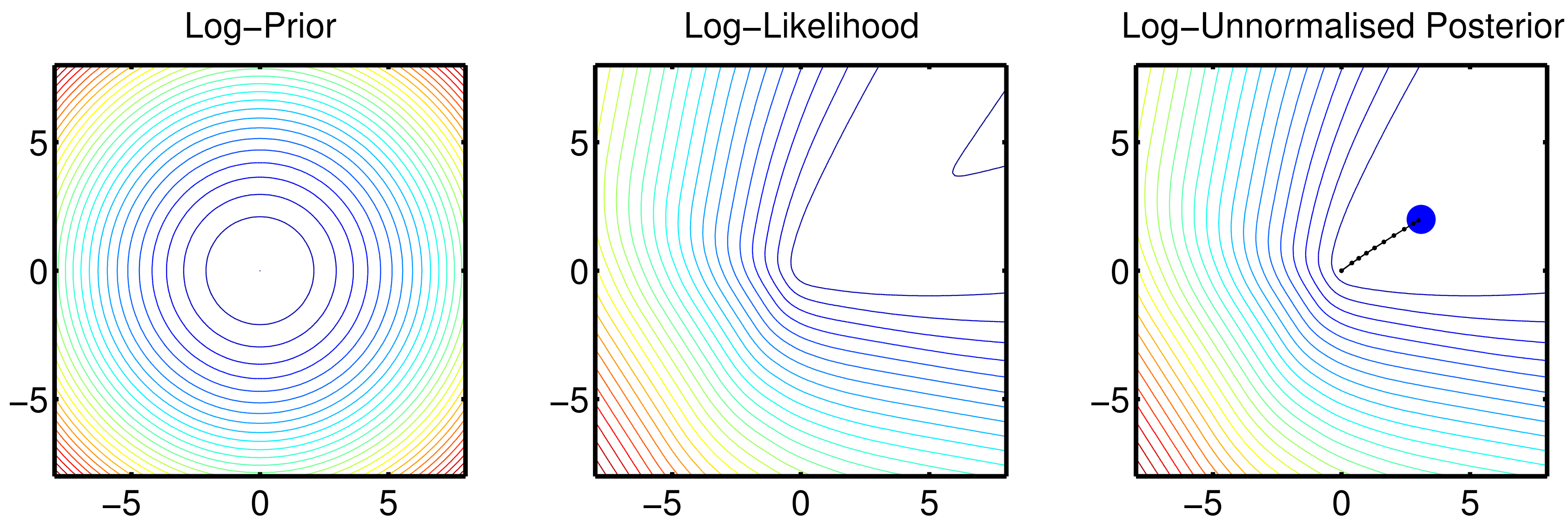

Figure 3: The three contour plots above show the negative logarithm of parameter probability distributions where the left-hand plot shows the distribution of the parameter values $\mathbf{w} = [w_1 \quad w_2]^\mathsf{T}$ under the defined prior. The middle plot shows the negative log-likelihood which is distinctly non-Gaussian and the right-hand plot shows the joint likelihood (un-normalised posterior). The large solid blue dot shows the point in parameter space where the posterior is a maximum and the lines of small dark dots shows the evolution of the Newton algorithm towards this point starting from and initial point of $\mathbf{w} = [0 \ \ 0]$, ten steps are required to achieve this optimum.

Laplace Demo

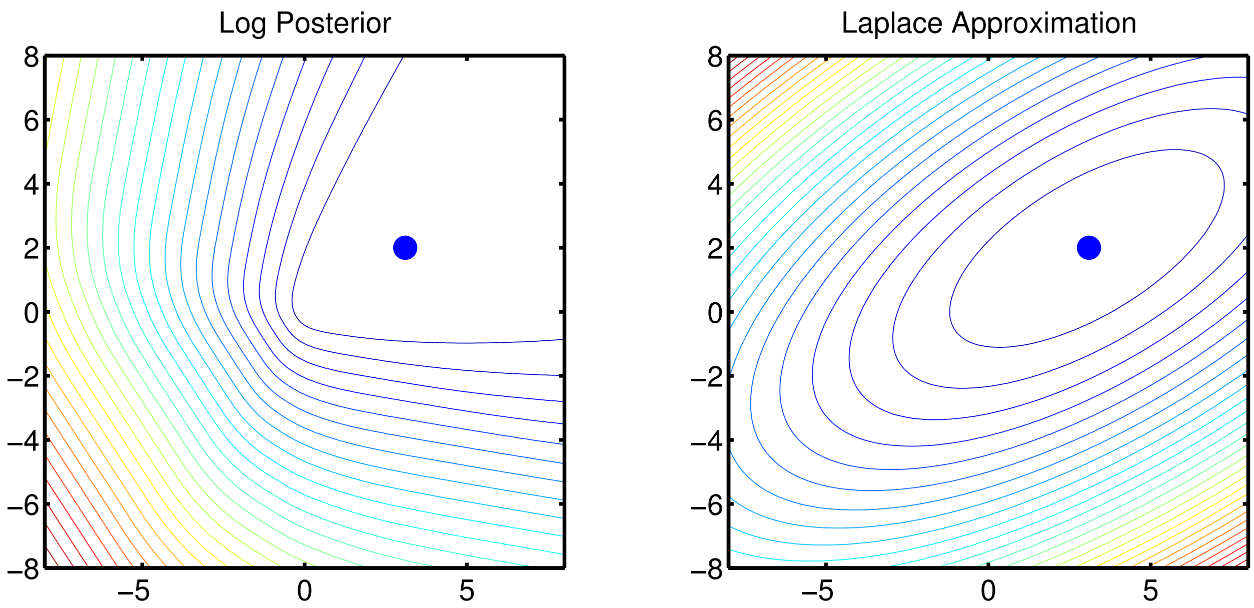

Figure 4: The left-plot shows the negative log-posterior whilst the right-plot shows the Laplace approximation. The first thing to note is that the location of the maximum has been reasonably well identified. The second point is to note that the positive curvature of the posterior (as both parameter values increase they become *a posteriori* more probable. We can observe this curvature in our Laplace approximation, however, note that as we move away from the *MAP* value the approximation is not so good.)

Bayesian Classification

- Now to make predictions we want the following distribution $P(C = 1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{t})$ which is

$$\int P(C = 1|\mathbf{x}_{new}, \mathbf{w})p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha)d\mathbf{w}$$

Bayesian Classification

UNIVERSITY of GLASGOW

- Now to make predictions we want the following distribution $P(C = 1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{t})$ which is

$$\int P(C = 1|\mathbf{x}_{new}, \mathbf{w})p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha)d\mathbf{w}$$

- Assume posterior is sharply peaked around *MAP* value $\Rightarrow$ class predictions made using the approximate predictive posterior probability

$$\begin{aligned} P(C = 1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{t}) &\approx P(C = 1|\mathbf{x}_{new}, \mathbf{w}_{MAP}, \alpha, \mathbf{X}, \mathbf{t}) \\ &= \frac{1}{1 + \exp(-\mathbf{w}_{MAP}^{\mathsf{T}}\phi(\mathbf{x}_{new}))} \end{aligned}$$

Bayesian Classification

UNIVERSITY *of* GLASGOW

- Now to make predictions we want the following distribution $P(C=1|\mathbf{x}_{new},\alpha,\mathbf{X},\mathbf{t})$ which is

$$\int P(C=1|\mathbf{x}_{new},\mathbf{w})p(\mathbf{w}|\mathbf{X},\mathbf{t},\alpha)d\mathbf{w}$$

- Assume posterior is sharply peaked around *MAP* value $\Rightarrow$ class predictions made using the approximate predictive posterior probability

$$\begin{aligned} P(C=1|\mathbf{x}_{new},\alpha,\mathbf{X},\mathbf{t}) &\approx P(C=1|\mathbf{x}_{new},\mathbf{w}_{MAP},\alpha,\mathbf{X},\mathbf{t}) \\ &= \frac{1}{1+\exp(-\mathbf{w}_{MAP}^{\mathsf{T}}\phi(\mathbf{x}_{new}))} \end{aligned}$$

- So the discriminant function is $P(C=1|\mathbf{x}_{new},\alpha,\mathbf{X},\mathbf{t})>0.5$ then $\mathbf{x}_{new}$ is assigned to Class $C=1$ and $C=0$ otherwise.

Bayesian Classification

UNIVERSITY of GLASGOW

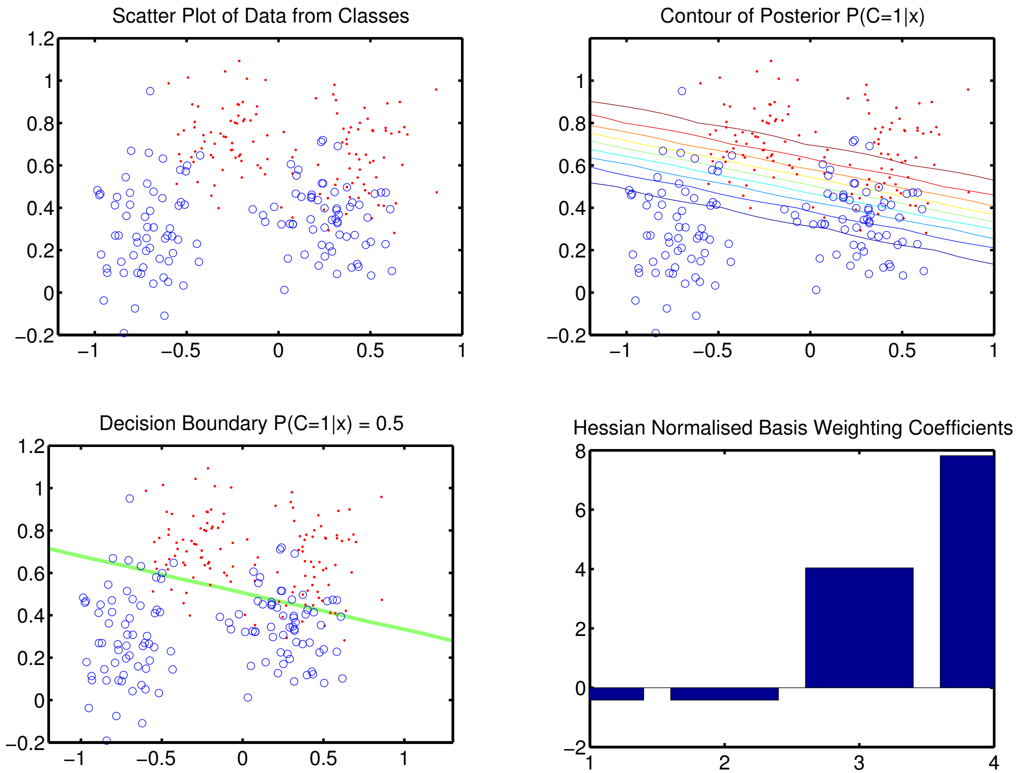

Figure 5: Top-left: two-dimensional data. Right hand plot shows posterior probability of class membership for linear model and decision boundary $P(C = 1|\mathbf{x}) = 0.5$ is shown in the bottom left plot. Magnitude of weighting coefficients normalised by the square-root of the Hessian matrix in bottom right plot, small values indicate irrelevant weights.

Bayesian Classification

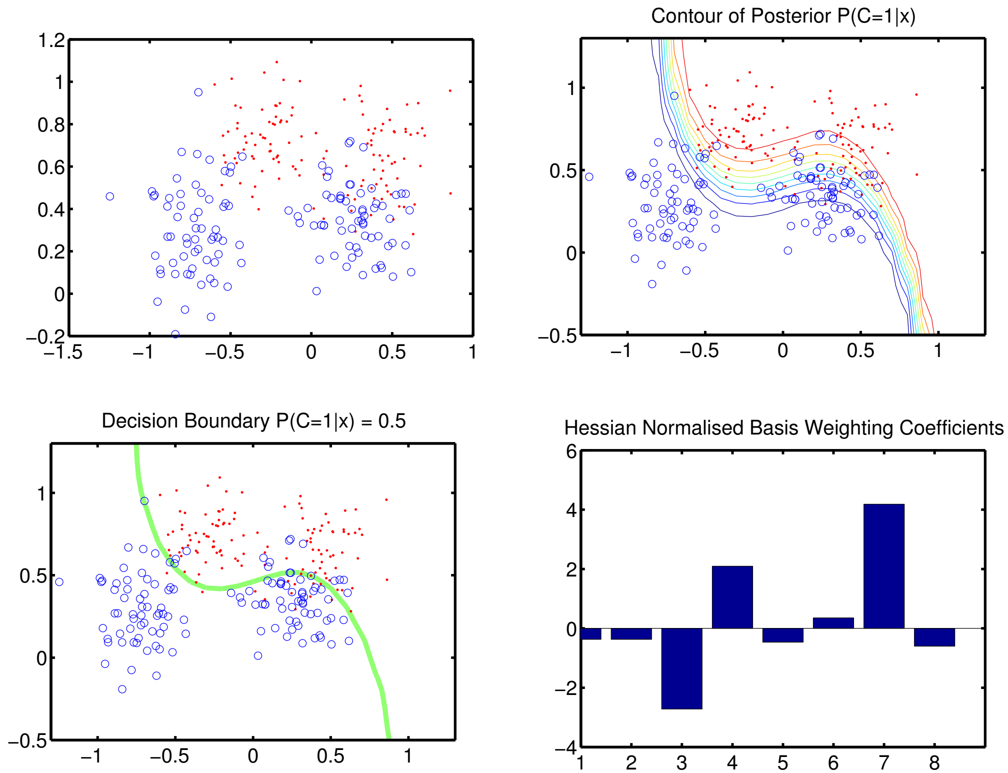

Figure 6: Top-left: two-dimensional data. Right hand plot shows posterior probability of class membership for cubic model and decision boundary $P(C = 1|\mathbf{x}) = 0.5$ is shown in the bottom left plot. Magnitude of weighting coefficients normalised by the square-root of the Hessian matrix in bottom right plot, small values indicate irrelevant weights.