M5MS10
Machine Learning

Lecture 2

Spring 2018

Dr Ben Calderhead

b.calderhead@imperial.ac.uk

M5MS10
Machine Learning

Lecture 2

Spring 2018

Dr Ben Calderhead

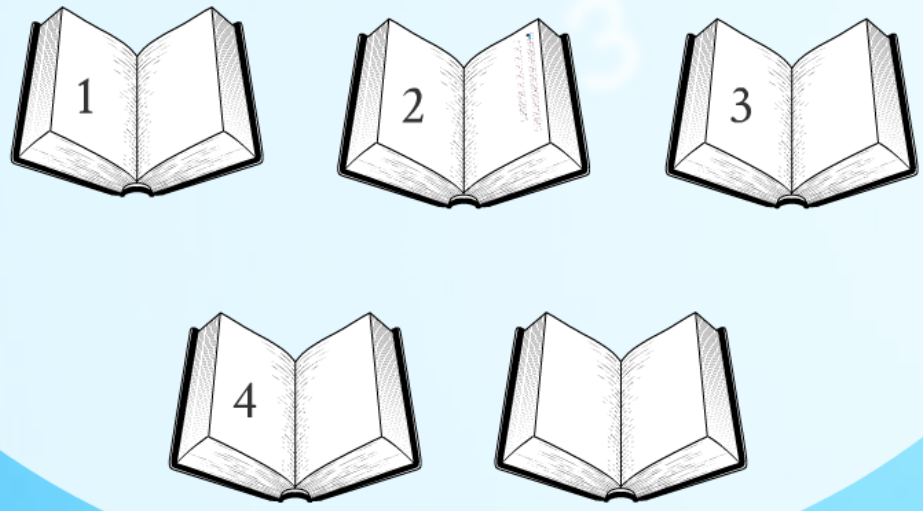
b.calderhead@imperial.ac.uk

MSc Projects

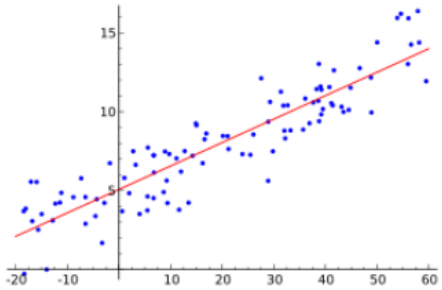
- **Quantitative Investment Strategies using Machine Learning**

- **Statistical Modelling in Formula 1**

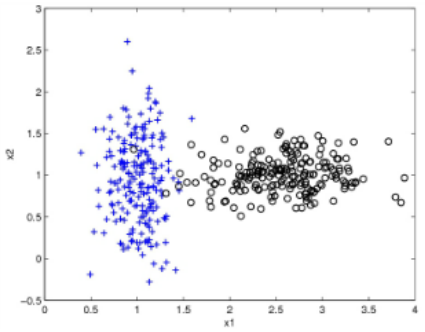
Parametric Methods



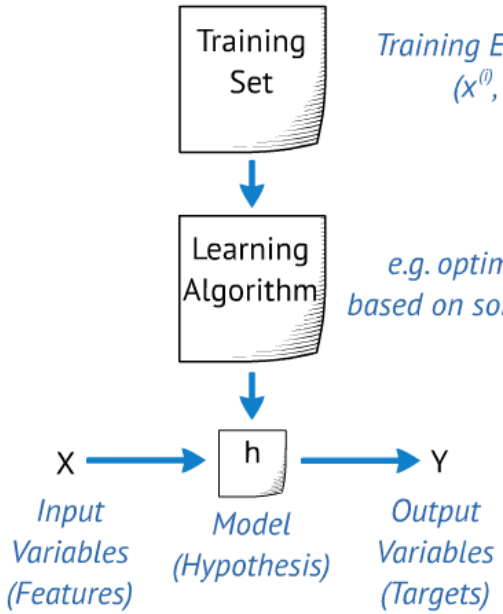
Supervised Learning Recap



Regression- supervised learning in which the labels are *continuous* values.



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



Given a model and some data, we can define a loss function measuring the mismatch between the model output and the observed labels.

We can learn the parameters of the model by minimising the loss function.

Cross validation gives us an approach for choosing the best predictive model.

The Previous Lecture

In the previous lecture we thought about the process of "learning from data". In other words, performing a statistical inference of the parameters of a model given some data.

We saw the general form of linear models and derived an expression for the optimal parameters based on minimising the mean squared error.

We can reduce the MSE by making the model more complex, however we may end up overfitting to the data - i.e. we begin describing the *noise* instead of the underlying *signal*.

We saw how cross validation can be used to avoid overfitting. i.e. optimise the parameters with respect to the average loss over different (training and) validation data sets.

Today's Lecture

In today's lecture, we will consider a couple of other approaches for avoiding overfitting. These are general principles, not only applicable to linear models, but most other statistical models we might be interested in!

We'll look briefly at the use of regularisation to prevent overfitting, before consider probabilistic approaches to parameter estimation.

The MSE gives us a point estimate of the parameters - we'll see that this is equivalent to the Maximum Likelihood Estimator and we remind ourselves how to quantify the uncertainty in our prediction.

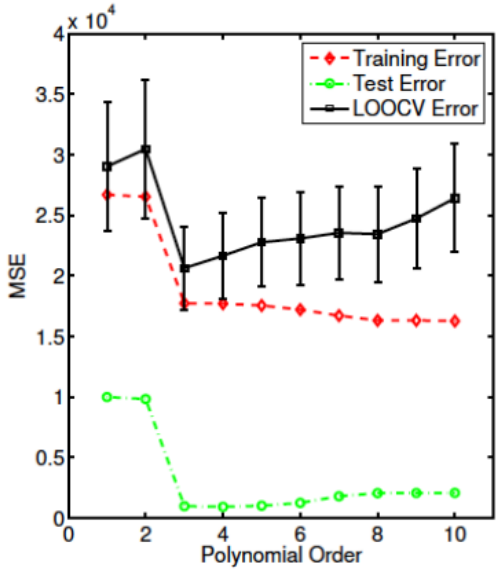
Finally, we'll consider the Bayesian approach to this problem and think about the use of marginal likelihood as yet another method for preventing overfitting.

Computer Lab 1

Employing too simple a model results in poor predictions.

However, employing too complex a model may also result in poor predictions due to overfitting.

We should take care about what performance measure we optimise for finding the appropriate parameters for our model, so that the model can generalise its performance beyond the data used for training.



Cross validation provides a better measure of performance by holding back subsets of the data to be used for testing.

Average Loss Functions

In lecture 1, we defined the *sample average loss function*,

$$\frac{1}{N}\sum_{i=1}^{N} J\left[y^{(i)}, h_{\theta}\left(x^{(i)}\right)\right]$$

We can derive it more formally in the following manner:

- Assume that each input-output pair (x,y) is drawn from some underlying distribution p(x,y).
- Ideally we want to minimise the loss over all possible (x,y) pairs that could be observed, i.e. minimise the *Expected Loss*.

$$\int\int J\left(y, h_{\theta}\left(x\right)\right) p(x, y) dx, dy$$

We therefore see that the sample average loss function is simply a finite sample estimate of the above expectation.

Regularised Loss Functions

An alternative approach to prevent overfitting is to add a regularisation term to the loss function. This will penalise models with very large parameter values.

- Ridge Regression employs L2 regularisation:

$$\frac{1}{N}\sum_{i=1}^{N} J\left[y^{(i)}, h_\theta\left(x^{(i)}\right)\right] + \sum_{j=1}^{K} ||\theta_j||_2$$

Advantages: easy to optimise
Disadvantages: solution strongly affected by outliers

- Lasso Regression employs L1 regularisation:

$$\frac{1}{N}\sum_{i=1}^{N} J\left[y^{(i)}, h_\theta\left(x^{(i)}\right)\right] + \sum_{j=1}^{K} ||\theta_j||_1$$

Advantages: encourages sparse solutions, robust to outliers in the data, performs *feature selection*
Disadvantages: L1 norm not differentiable, requires more sophisticated methods of optimisation.

Probabilistic Interpretation

In order to learn parameters for a linear model, we previously defined a mean squared error loss function.

We shall now consider loss functions from a probabilistic view. Let us consider the following relationship between the input and output variables,

$$y = h(x; \theta) + \epsilon$$

The right hand side describes a mean and some error term that captures noise or other unmodelled effects.

Our model, h, is a *deterministic* function of the inputs.

We can choose to model the noise however we want, but a common choice is to use a Normal distribution.

Probabilistic Interpretation

Make some reasonable assumptions:

- Observations made independently of each other
- Noise corrupting measurements come from same distribution i.e. iid data

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \sigma) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}, \boldsymbol{\theta}, \sigma) = \prod_{i=1}^{N} \mathcal{N}_{y^{(i)}}\left(h(x^{(i)}, \boldsymbol{\theta}), \sigma\right)$$

This is our likelihood function, which we can maximise in order to estimate the parameters. i.e. consider when first order derivative is equal to zero and solve for parameters.

The maximum likelihood solution is $\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}$ which is the same expression as the MSE previously.

Estimating Uncertainty

We note that we could also calculate the stationary point with respect to sigma, and obtain an analytic expression for that parameter too.

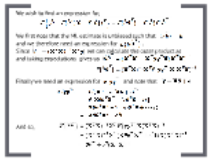
Furthermore we can estimate the uncertainty in our inferred parameter values, by calculating the covariance of our maximum likelihood estimate.

Remembering that the covariance is defined as,

$$E\left\{(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})^T\right\} = E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\} - E\{\hat{\boldsymbol{\theta}}\}E\{\hat{\boldsymbol{\theta}}^T\}$$

we can calculate an estimate for the uncertainty in our maximum likelihood solution,

$$\begin{aligned} E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\} - E\{\hat{\boldsymbol{\theta}}\}E\{\hat{\boldsymbol{\theta}}^T\} &= \boldsymbol{\theta}\boldsymbol{\theta}^T + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} - \boldsymbol{\theta}\boldsymbol{\theta}^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$



We wish to find an expression for,

$$E\left\{(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})^T\right\} = E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\} - E\{\hat{\boldsymbol{\theta}}\}E\{\hat{\boldsymbol{\theta}}^T\}$$

We first note that the ML estimate is unbiased such that $E\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}$ and we therefore need an expression for $E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\}$.
Since $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ we can calculate the outer product as and taking expectations gives us $\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$

$$E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\{\mathbf{y}\mathbf{y}^T\}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

Finally we need an expression for $E\{\mathbf{y}\mathbf{y}^T\}$ and note that $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$

$$\begin{aligned}
E\{\mathbf{y}\mathbf{y}^T\} &= E\{(\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon})(\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon})^T\} \\
&= E\{\mathbf{X}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{X}^T + 2\boldsymbol{\epsilon}\boldsymbol{\theta}^T\mathbf{X} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}\} \\
&= \mathbf{X}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{X}^T + 2E\{\boldsymbol{\epsilon}\}\boldsymbol{\theta}^T\mathbf{X} + E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}\} \\
&= \mathbf{X}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{X}^T + \sigma^2\mathbf{I}
\end{aligned}$$

And so,

$$\begin{aligned}
E\{\theta\theta^T\} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\{\mathbf{y}\mathbf{y}^T\}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{X}^T + \sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \boldsymbol{\theta}\boldsymbol{\theta}^T + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

Estimating Uncertainty

We note that we could also calculate the stationary point with respect to sigma, and obtain an analytic expression for that parameter too.

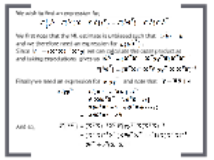
Furthermore we can estimate the uncertainty in our inferred parameter values, by calculating the covariance of our maximum likelihood estimate.

Remembering that the covariance is defined as,

$$E\left\{(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})^T\right\} = E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\} - E\{\hat{\boldsymbol{\theta}}\}E\{\hat{\boldsymbol{\theta}}^T\}$$

we can calculate an estimate for the uncertainty in our maximum likelihood solution,

$$\begin{aligned} E\{\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\} - E\{\hat{\boldsymbol{\theta}}\}E\{\hat{\boldsymbol{\theta}}^T\} &= \boldsymbol{\theta}\boldsymbol{\theta}^T + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} - \boldsymbol{\theta}\boldsymbol{\theta}^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$



Estimating Uncertainty

This is an important result as it allows us to estimate the uncertainty in our maximum likelihood estimate, and in fact this uncertainty can be written in terms of the second partial derivatives of the log-likelihood:

$$E\left\{(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})^T\right\} = -\left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}\right)^{-1}$$

Notes:

- We can interpret this as a Gaussian approximation of the uncertainty of the maximum likelihood.
- Based on a measure of the curvature of the likelihood surface - equal to the inverse of the Expected Fisher Information (minimum variance of an unbiased estimator).
- The expected Fisher Information describes how quickly the gradient changes with respect to the model parameters (think about geometry of likelihood function).
- A small curvature implies a high variance estimate (since taking the inverse).
- Provides a fascinating link with Riemannian geometry!

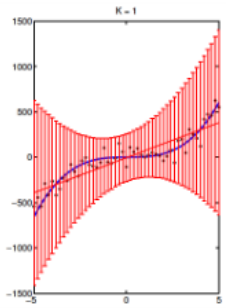
Revisiting Polynomial Models

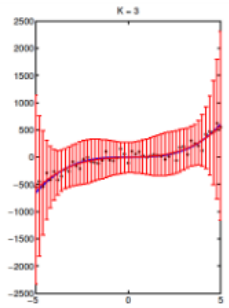
- Let's have a look at learning all parameters of the linear model using third order polynomial data - including the noise parameter - by maximising the likelihood (which is an equivalent way of looking at a loss function).

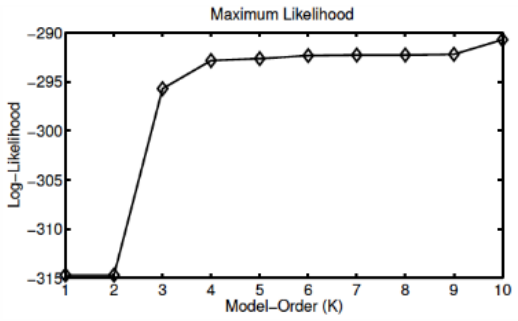
$$\hat{\mathbf{y}}_{\text{new}} = \mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$\sigma^2_{\text{new}} = \hat{\sigma}^2\mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$$

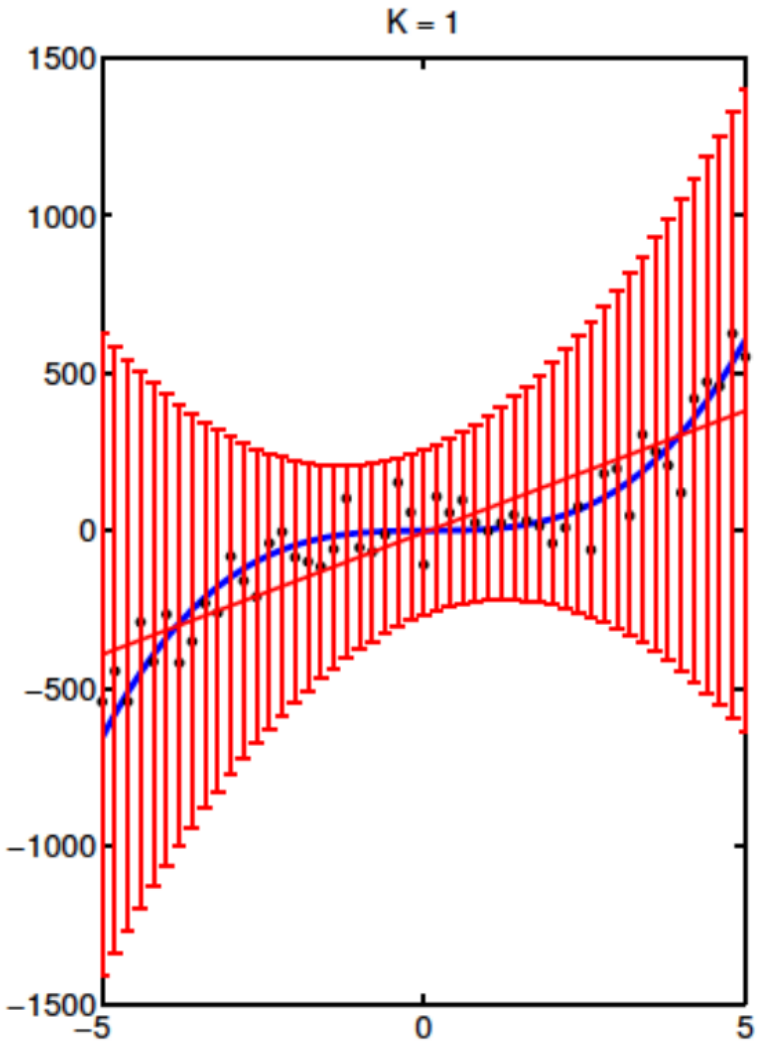
where $\hat{\sigma}^2 = \frac{1}{N}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\hat{\mathbf{y}}_{\text{new}})$

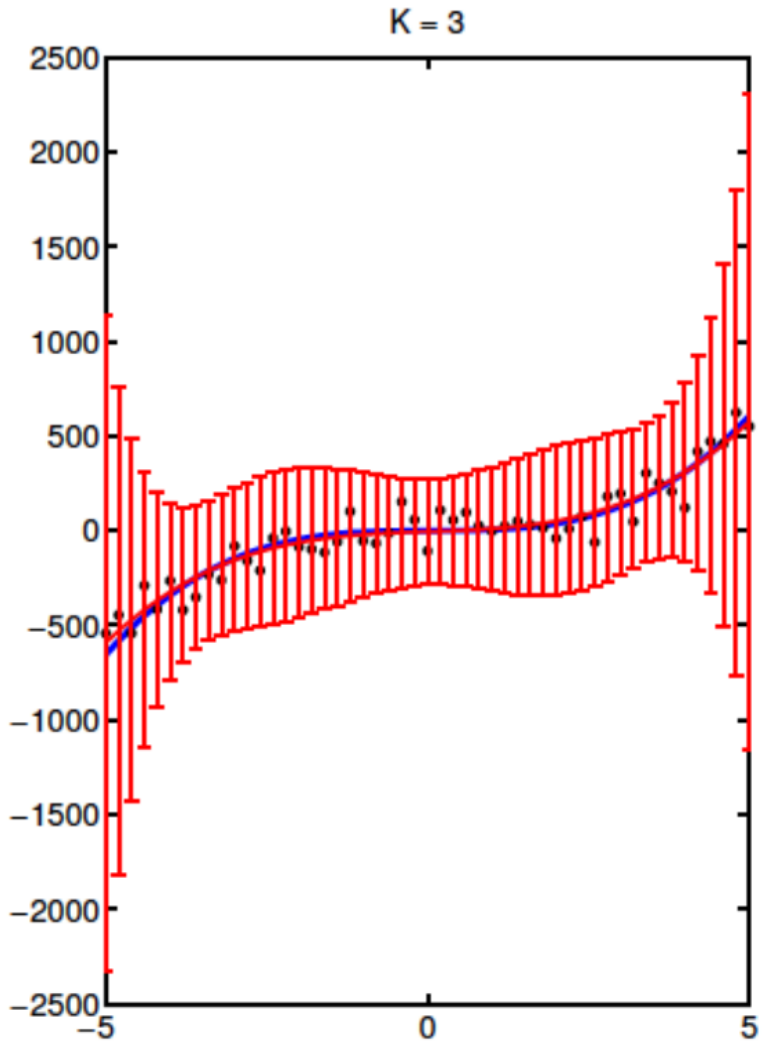
See if you can derive this yourself in an analogous manner.

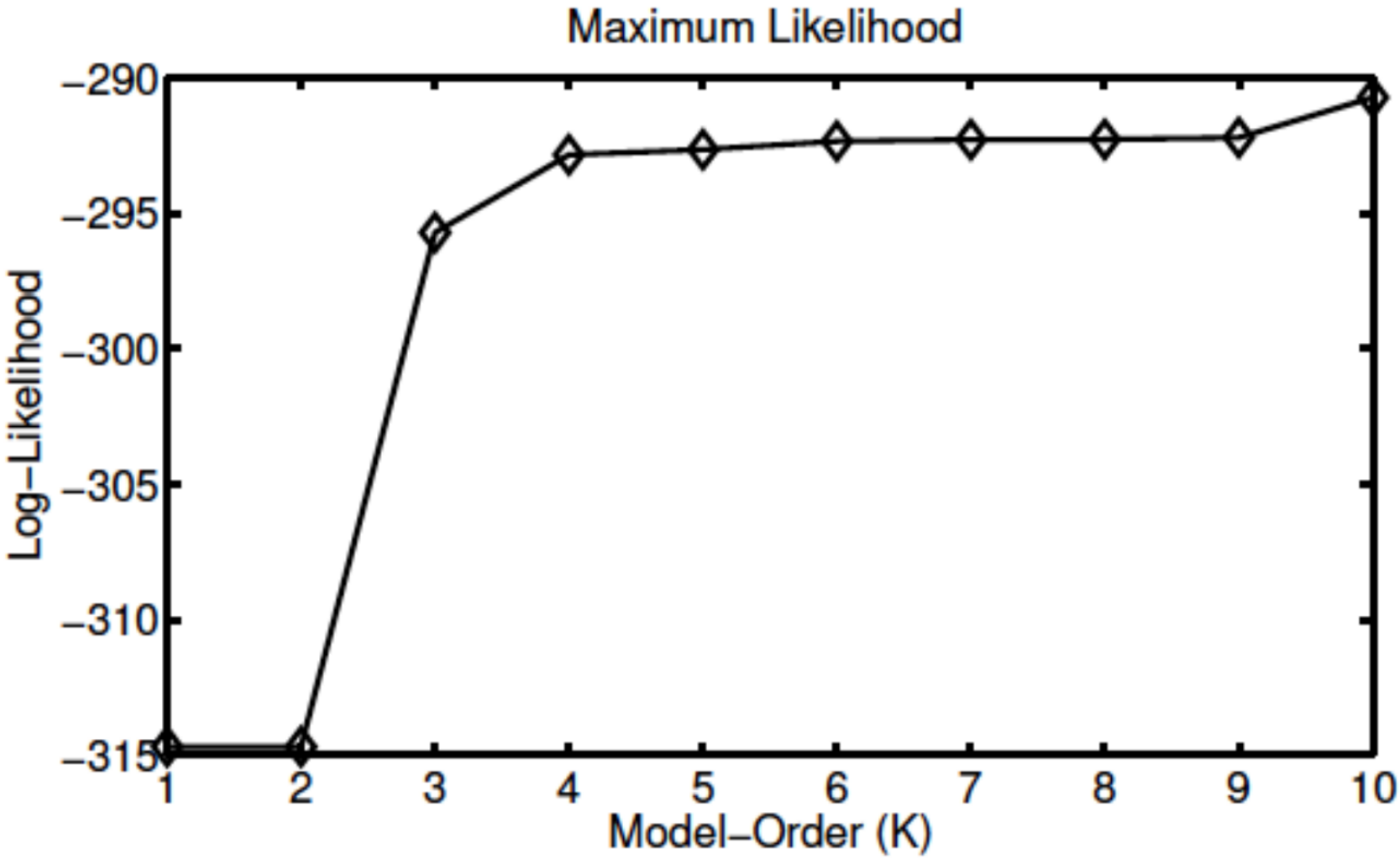







K = 1
1500
1000
500
0
−500
−1000
−1500
−5
0
5


K = 3
2500
2000
1500
1000
500
0
−500
−1000
−1500
−2000
−2500
−5
0
5


Maximum Likelihood
−290
−295
−300
−305
−310
−315
Log−Likelihood
1
2
3
4
5
6
7
8
9
10
Model−Order (K)

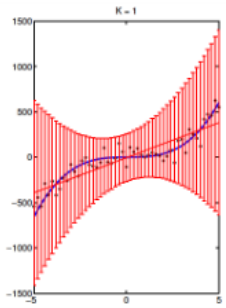
Revisiting Polynomial Models

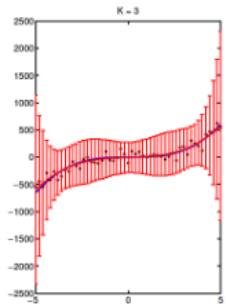
- Let's have a look at learning all parameters of the linear model using third order polynomial data - including the noise parameter - by maximising the likelihood (which is an equivalent way of looking at a loss function).

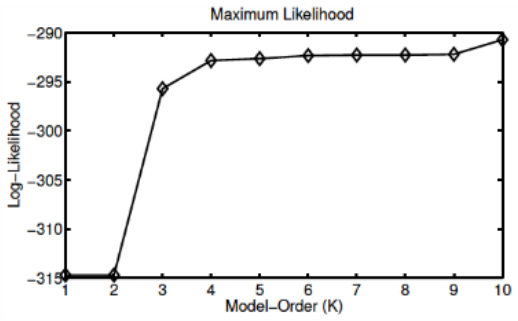
$$\hat{\mathbf{y}}_{\text{new}} = \mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$\sigma^2_{\text{new}} = \hat{\sigma}^2\mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$$

where $\hat{\sigma}^2 = \frac{1}{N}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\hat{\mathbf{y}}_{\text{new}})$

See if you can derive this yourself in an analogous manner.







The Bayesian Approach

So far we have "learned" the parameters for our linear models by minimising the loss or, equivalently, maximising the likelihood, $p(\mathbf{y}|\boldsymbol{\theta}, \sigma, \mathbf{X})$

Likelihood methods assume that there is a "true" underlying set of parameter values to solve the regression problem, and we optimise the parameters to find this most likely set.

The Bayesian approach makes it explicit that it is the probability distribution defined over the parameters that we are most interested in, i.e. $p(\boldsymbol{\theta}, \sigma|\mathbf{X}, \mathbf{y})$

We can obtain the posterior distribution over the parameters $p(\boldsymbol{\theta}, \sigma|\mathbf{X}, \mathbf{y})$ from the likelihood by using Bayes theorem, and it turns out that prior distributions play the role of the regularisation terms we saw earlier.

The Bayesian Approach

Let's make the simplifying assumption that we know the noise level parameter σ.

We are therefore interested in the posterior distribution over the parameters $\boldsymbol{\theta}$,

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

where the marginal likelihood can be written as,

$$p(\mathbf{y}|\mathbf{X}, \sigma) = \int p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}, \sigma)p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

The likelihood has already been defined in terms of a Gaussian error model. We now need to define the prior.

The Prior Distribution

We can choose our prior however we want. Assuming that the parameters are *a priori* independent, choosing a Gaussian prior is equivalent to having an L2 regularisation term - this encodes our belief regarding the possible parameter values, before we see any of the data.

$$p(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{0}, \boldsymbol{\Lambda}) \tag{1}$$

where $\boldsymbol{\Lambda} = \alpha\mathbf{I}$, i.e. a multivariate normal with scaled identity covariance, encoding the belief that the parameters should be small and centred around zero.

The Posterior Distribution

In this case, since the likelihood is Gaussian distributed and the prior is Gaussian distributed, we can actually compute the posterior analytically,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \sigma) &= \frac{\mathcal{N}_{\mathbf{y}}(\mathbf{X}\boldsymbol{\theta}, \sigma\mathbf{I})\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{0}, \boldsymbol{\Lambda})}{\int \mathcal{N}_{\mathbf{y}}(\mathbf{X}\boldsymbol{\theta}, \sigma\mathbf{I})\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{0}, \boldsymbol{\Lambda}) d\boldsymbol{\theta}} \\ &= \mathcal{N}_{\boldsymbol{\theta}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

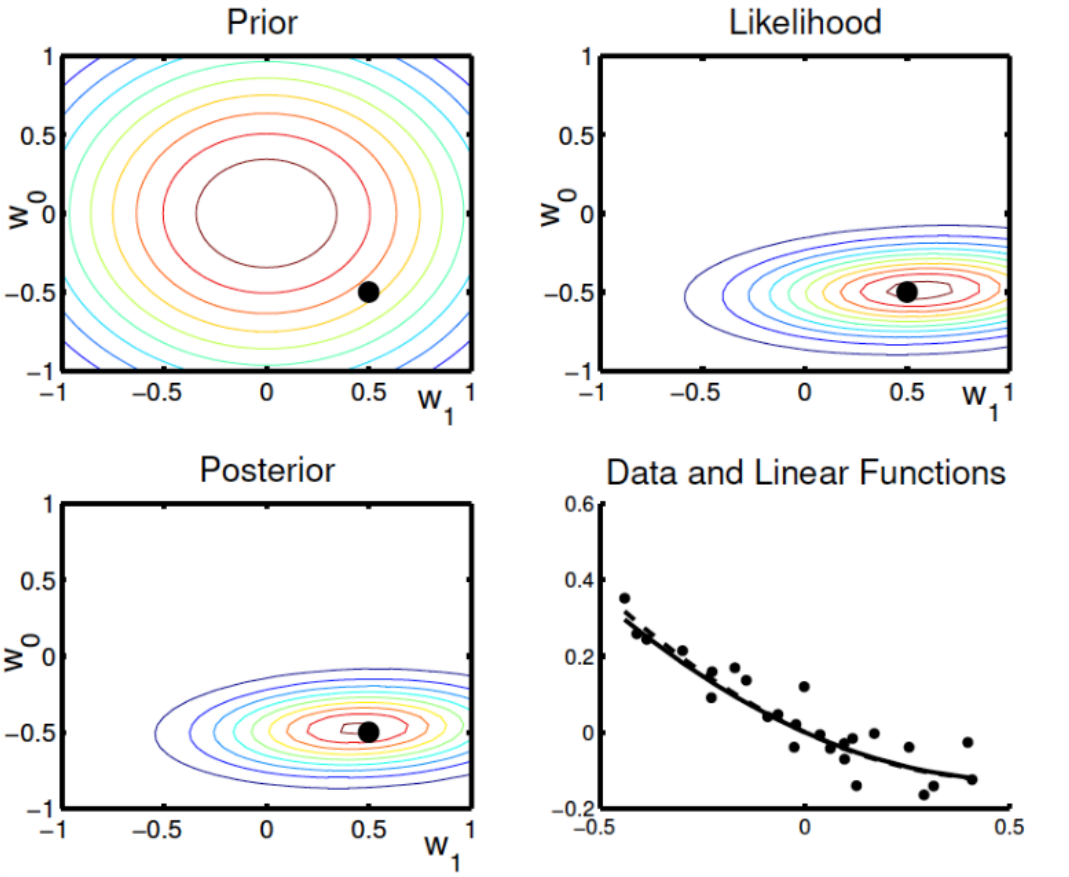
where

$$\boldsymbol{\mu} = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \qquad \boldsymbol{\Sigma} = \sigma^2\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}$$

For doing this maths, the Matrix Cookbook comes in extremely useful!

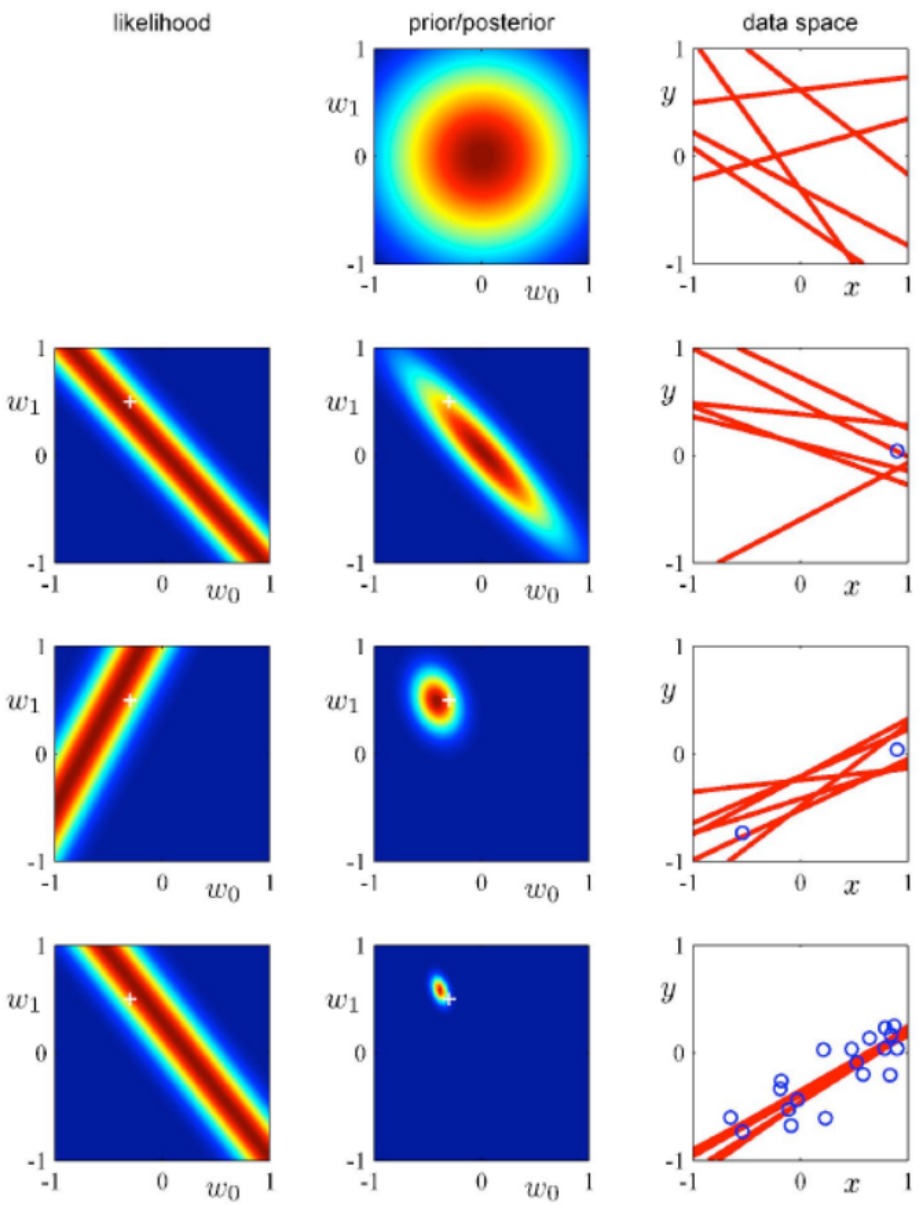
(Although working through this example long hand is good for the soul...)

Example Distributions



Here we plot the posterior distribution over 2 parameters of a 2nd order linear model fit to some data. The black spot shows the true values used to generate the data. The true value of σ was used and the prior variance was set to $\alpha = 1$.

Increasing Data



No data points

1 data point

2 data points

Many data points

The Predictive Posterior

Finally, we want to make predictions. With the likelihood approach we simply use the single (maximum) value for $\boldsymbol{\theta}$ that we obtained. In the Bayesian approach, we get a posterior distribution over the parameters and we can average over this to obtain the distribution of our predictions, called the predictive posterior, as follows.

$$\begin{aligned} p(\mathbf{y}_{\text{new}}|\mathbf{X}, \mathbf{y}, \sigma) &= \int p(\mathbf{y}_{\text{new}}|\mathbf{X}, \boldsymbol{\theta}, \sigma) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \sigma) d\boldsymbol{\theta} \\ &= \mathcal{N}(\mathbf{x}^T \boldsymbol{\mu}, \mathbf{x}_{\text{new}}^T \Sigma \mathbf{x}_{\text{new}}) \end{aligned} \tag{2}$$

In addition, we note that the marginal likelihood denotes the probability of the model given the data, with the parameters integrated out. This quantity can be used for model comparison, since the integral automatically (!) penalises higher dimensional models.

(Think about what happens to the probability mass assigned to a unit of volume as the dimensionality increases...)

Summary

- We have examined the Bayesian approach to learning parameters for linear models.
- The prior can be considered as a regularisation term, which favours simpler solutions. And the marginal likelihood offers an alternative to cross validation for model selection.
- We obtain a probability distribution as our answer, which gives us a measure of how confident we can be in our answer.
- In the case of linear models with Gaussian noise, the distributions of interest may all be calculated analytically. In contrast, the ODE models from the previous lecture are not analytic and require more sophisticated methods such as Markov chain Monte Carlo to sample from the posterior.

Further Reading on Linear Modelling in Machine Learning

- Chapters 17 & 18
  Bayesian Reasoning and Machine Learning
  by David Barber

- Chapter 7
  Machine Learning: A Probabilistic Perspective
  by Kevin Murphy

  See Blackboard for a few interesting papers!

M5MS10
Machine Learning

Lecture 2

Spring 2018

Dr Ben Calderhead

b.calderhead@imperial.ac.uk