Machine Learning Module

Week 3

Lecture Notes 5 & 6

Probabilistic & Bayesian Methods

Mark Girolami
girolami@dcs.gla.ac.uk
Department of Computing Science
University of Glasgow

January 27, 2006

1 A Probabilistic View of Linear Regression

The data model which we have explored so far is of the form

$$t = f(x; \mathbf{w}) + \epsilon \tag{1}$$

where our observations or target values t are modeled by a deterministic function of our inputs, $f(x; \mathbf{w})$, which may be contaminated by noise or some error defined by ϵ. We are now going to present a probabilistic interpretation of the linear regression model which has far reaching consequences in terms of data modeling.

1.1 Distribution of Noise Component

The noise term can be assumed to be Gaussian or Normally distributed with mean zero and some variance σ i.e. $\epsilon \sim \mathcal{N}(0, \sigma)$. Now what this means is that the target value t given the value of x will then be a constant value $f(x; \mathbf{w})$ with an additive zero-mean Gaussian variable so that t given x will be a Gaussian distribution with mean $f(x; \mathbf{w})$ and variance σ. It might help to think of this as an information bearing signal, $f(x; \mathbf{w})$, which is corrupted with noise, ϵ, upon transmission so that the noisy signal received is the sum of both components.

This can be written as

$$t|x \sim \mathcal{N}(f(x; \mathbf{w}), \sigma)$$

which reads as t given x has a Gaussian distribution with mean $f(x; \mathbf{w})$ and variance σ. Likewise we can write

$$p(t|x) = \mathcal{N}(f(x; \mathbf{w}), \sigma)$$

which reads as the conditional probability distribution of t given x is Gaussian distribution with mean $f(x; \mathbf{w})$ and variance σ.

1.2 The Likelihood Principle

The question that we ask is *How likely is it that I would have observed the outputs given the inputs*, the likelihood of observing the outputs is the conditional probability of making all the observations. Now if we have made N observations $(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$ (where both vectors are $N \times 1$)

then we are interested in the joint probability of all the outputs conditioned on all the input values i.e. $p(t_1, t_2, \cdots, t_N|x_1, x_2, \cdots, x_N)$ which can be written in vector format more compactly as $p(\mathbf{t}|\mathbf{x})$.

We now make a further assumption, which is not always satisfied or justified, that we make our observations *independently* of each other so that the measurement we have just made does not affect the following measurement we make. This assumption essentially is assuming *statistical independence* between measurements.

The additional important assumption that we make is that the noise corrupting our measurements always comes from the same distribution and so our outputs will all be *identically distributed*.

Taken together these two assumptions can be stated as *we assume that the data is Independent and Identically Distributed* often denoted as IID for short.

Now with the IID assumption then the joint probability of our measurements takes a factored[1] or product form. In which case

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_{n=1}^{N} p(t_n|x_n) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \mathbf{w}), \sigma)$$

We see from our likelihood function above that the likelihood depends on the parameters of our model as the deterministic model response defines the mean of each univariate Gaussian. So to make this dependency explicit we have written the joint likelihood as $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma)$ indicating that the *targets*, $\mathbf{t}$, are dependent on the inputs $\mathbf{x}$, as well as the model parameters, $\mathbf{w}$, and the variance of the additive noise σ. This explicit conditioning will become very important in later work and it should become second nature to you as the course develops.

[1] Two random variables have a joint distribution $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$, now if the variables are independent of each other then the value which X or Y takes will have no influence on Y or X and so the conditional distributions $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$, which means that the joint disribution is simply the product of the marginal terms i.e. $P(X, Y) = P(X)P(Y)$. This extends to any arbitrary number of variables.

1.3 Maximum Likelihood

We now want to select the model parameters[2] $\mathbf{w}$ & σ which will make our observations most likely and so we need to maximise the likelihood function above with respect to all parameters. In actual fact we will maximise the logarithm of the likelihood function as the log-likelihood is more convenient to work with analytically[3] and as the logarithm is a convex function the estimated arguments $\widehat{\mathbf{w}}$ & $\widehat{\sigma}$ which maximise the log-likelihood will also maximise the likelihood. So lets do it.

$$\begin{aligned}
\mathcal{L} = \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) &= \sum_{n=1}^{N} \log p(t_n|x_n, \mathbf{w}, \sigma) \\
&= \sum_{n=1}^{N} \log \mathcal{N}(f(x_n; \mathbf{w}), \sigma) \\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}|t_n - f(x_n; \mathbf{w})|^2\right) \\
&= -\frac{N}{2}\log 2\pi - N \log \sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N} |t_n - f(x_n; \mathbf{w})|^2
\end{aligned}$$

We now only need to take derivatives and solve for the stationary points of the log-likelihood

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(t_n \mathbf{x}_n - \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} \mathbf{w}) = 0$$

Lets use the $N \times (K+1)$ dimensional matrix $\mathbf{X}$ which stacks all the column vectors $\mathbf{x}_n$ row-wise then

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathsf{T}}\mathbf{t} - \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w}) = 0$$

[2] Notice that the variance of the noise is also a model parameter which is not as obvious within the classical Least-Squares presentation.

[3] This is not the only reason as the log-likelihood has a direct connection to information theory and so the amount of information encoded by the model can be defined using the log-likelihood. We will not pursue these parallels further in this course.

and so the *maximum-likelihood* solution for $\mathbf{w}$ follows simply as $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$. So we see that the *Least-Squares* solution is also the maximum-likelihood solution when all distributional assumptions are made explicit.

The Hessian matrix of second-order partial derivatives follows as

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

which is strictly negative and so we have indeed obtained the maximum of the likelihood.

It is left to the student to show that the *maximum-likelihood* estimate for the noise variance follows as

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}}) \\ &= \frac{1}{N}\mathbf{t}^\mathsf{T}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\right)\mathbf{t} \\ &= \frac{1}{N}\left(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\hat{\mathbf{t}}\right)\end{aligned}$$

and that this is indeed a maximum by showing that the curvature of the likelihood at the stationary point $\widehat{\sigma}$ is always negative and equals

$$\frac{\partial \mathcal{L}^2}{\partial \sigma \partial \sigma} = -\frac{2N^2}{(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})} = -\frac{2N}{\widehat{\sigma^2}}$$

Show that the expression for the maximum of the log-likelihood is given by the following

$$-\frac{N}{2}(1 + \log 2\pi) - N \log \widehat{\sigma}$$

You should be able to see that the log-likelihood will monotonically increase in value as the reconstruction error decreases which means that the log-likelihood computed on the training data otherwise known as the *in-sample* likelihood will always favour more complex models.

1.4 Uncertainty in Estimates & Predictions

We can also describe the variability of our maximum-likelihood estimate $\widehat{\mathbf{w}}$ by writing out the covariance of the estimate

$$\mathsf{cov}\{\widehat{\mathbf{w}}\} = E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} - E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\}$$

where the expectations are taken with respect to the distribution of the data $\mathbf{t}$. Now remember that the Least-Squares and so the Maximum-Likelihood estimators are unbiased in which case the expected value of our estimate will be the *true* parameter value i.e. $E\{\widehat{\mathbf{w}}\} = \mathbf{w}$. We require to obtain

$$\begin{aligned} E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} &= \int p(\mathbf{t}|\mathbf{X}) \begin{bmatrix} \widehat{w_0}^2 & \widehat{w_0}\widehat{w_1} \\ \widehat{w_1}\widehat{w_0} & \widehat{w_1}^2 \end{bmatrix} d\mathbf{w} \\ &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathsf{T}} E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \end{aligned}$$

Now as the noise is assumed to have zero-mean and variance σ^2 then

$$\begin{aligned} E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\} &= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} - 2E\{\epsilon\}\mathbf{X}^{\mathsf{T}}\mathbf{w} + \sigma^2\mathbf{I} \\ &= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + \sigma^2\mathbf{I} \end{aligned}$$

and so using this result in the above we finally obtain

$$\mathsf{cov}\{\widehat{\mathbf{w}}\} = \mathbf{w}\mathbf{w}^{\mathsf{T}} + \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} - \mathbf{w}\mathbf{w}^{\mathsf{T}} = \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$$

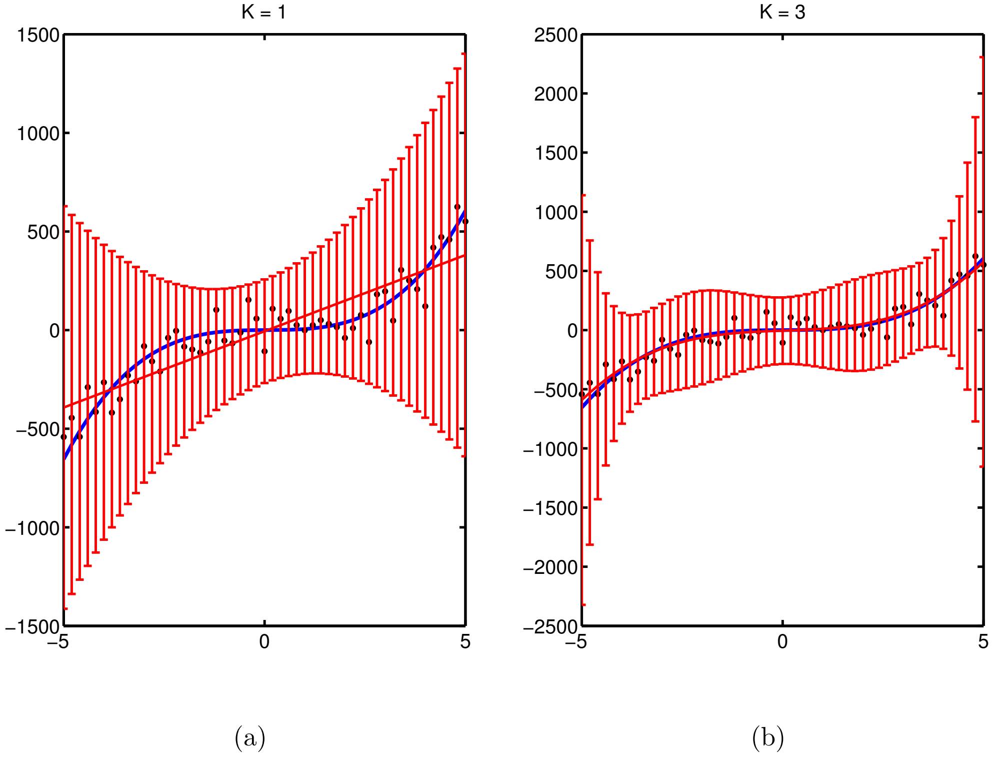
It is interesting to note that from the expression for the Hessian matrix of partial derivative that

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

where the negative Hessian of partial derivatives is defined as the *Information* matrix. We will not pursue this further but it is important to know that we now have both an estimator for the unknown parameters and a measure of the uncertainty (or spread or variability) in our estimate. This is particularly important when making prediction regarding unseen events as we can now say more than just what our predicted value is but what range of values may be expected. The tighter the range of values the more confident we can be of our predictions.

So to make a *new* prediction then our maximum-likelihood estimate and the associated variance around this estimate i.e. $\widehat{t}_{new} \pm \sigma^2_{new}$ where

$$\begin{aligned}\widehat{t}_{new} &= \mathbf{x}^{\mathsf{T}}_{new}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \\ \sigma^2_{new} &= \widehat{\sigma}^2\mathbf{x}^{\mathsf{T}}_{new}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{x}_{new}\end{aligned}$$



(a)

(b)

Figure 1: The blue solid line indicates the true noise free functions and the black dots are the actual observed noisy realisations of the data. The solid red line indicates the estimated function with the error-bars indicating the variance (uncertainty) in the estimated functional response at each of the data points ie $\widehat{t}_n \pm \sigma^2_n$.

Figure (1) shows a linear fit (K = 1) and a cubic fit (K = 3) to noisy realisations of the function $5x^3 - x^2 + x$ it is interesting to see how the error-bars decrease as the model becomes sufficiently flexible to model the underlying function. The following Matlab script (max_like_demo.m) will generate such plots.

```
clear
Range = 10;
Max_Model_Order = 10;
noise_var = 100;

L=[];
x = [-Range/2:0.2:Range/2]';
N=size(x,1);

f = 5*x.^3  - x.^2 + x;
f_n = f + noise_var*randn(size(x));

[i,j]=sort(x); X=x.^0;

for k=1:Max_Model_Order
    X=[X x.^k];
    w_hat = inv(X'*X)*X'*f_n;
    f_hat = X*w_hat;
    sigma_hat = mean((f_n - f_hat).^2);
    sigma = sigma_hat*diag(X*inv(X'*X)*X');

    L  = [L; -N*log(sqrt(sigma_hat)) - 0.5*N*(1 + log(2*pi))];

    plot(i,f(j),'b');
    hold on
    plot(i,f_n(j),'.k','MarkerSize',15)
    errorbar(i,f_hat(j),sigma(j),'-r.')
    hold off
    pause(1)
end

figure
plot(1:Max_Model_Order,L,'dr--');
```

It is also interesting to view the corresponding log-likelihood score for $K = 1$ to 10 as obtained from the Matlab script (Figure (2)).

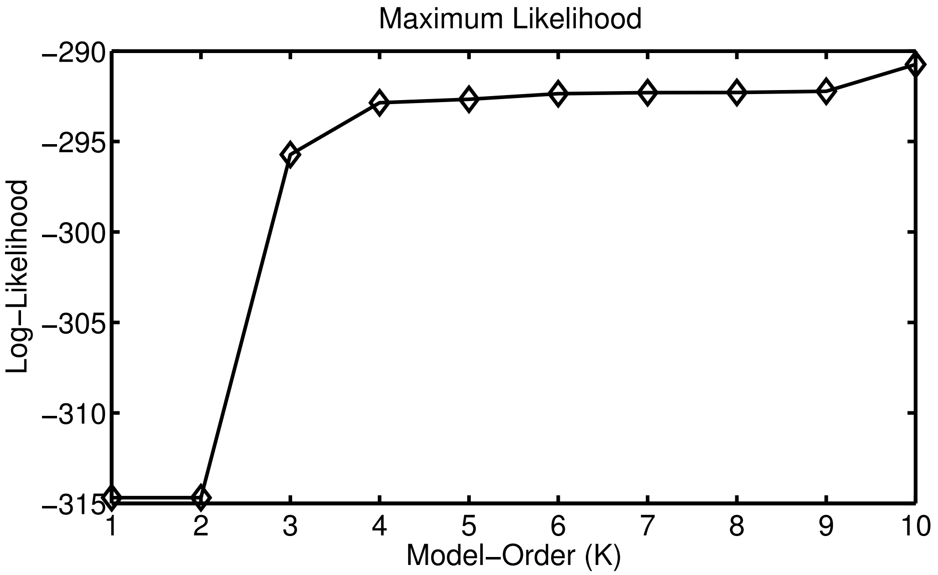


Figure 2: The Maximum Likelihood score for polynomial models from $K = 1$ to $K = 10$. Perhaps unsurprisingly the likelihood score monotonically increases with K.

The covariance of the parameter estimates can also be used to assess how relevant certain parameters are but we shall look at this later in the course.

1.5 Conclusion

This section has introduced the usage of the likelihood principle in devising linear regression models. However, the Maximum-Likelihood framework is entirely general and can be employed whenever a probabilistic representation of a model can be defined. We have seen that the introduction of probabilistic semantics provides a far richer *tool-set* than simple least-squares in that uncertainty in estimates can be given. However maximum-likelihood is not without its problems as we have seen in the previous experiment maximising likelihood will lead to overfitting unless predictive-likelihoods using CV are monitored in the model fitting process.

The following section now introduces the Bayesian method applied to linear regression modeling.

2 The Bayesian Approach

Now we have come some way by presenting our model identification problem within a likelihood based framework and this is particulary powerful. However, at the end of the day the question we really would like to answer is not so much how likely the data is given our model but what is the likelihood of the model parameters given the data. This is essentially an inversion of the probabilities associated with the likelihood principle and relies on Bayes inversion rule to achieve this.

Bayes rule is simple to understand but it has wide implications as to what sort of questions potentially could be posed and answered which led to an enormous amount of controversy amongst statisticians. Today however it is widely accepted that Bayesian methods of inference are particularly powerful and the development and standard usage of Bayesian methods are widespread within the Machine Learning community.

2.1 Posterior Inference

So in Likelihood methods we are interested in how likely the data is given our model and associated parameters $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma)$ but as we really want to know about the model given the data then the quantity we should concerned with is $p(\mathbf{w}, \sigma|\mathbf{x}, \mathbf{t})$.

Now for two random variables X & Y the joint probability $p(X, Y)$ can be decomposed in two ways such that $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$. So this then allows us to make the inversion $P(Y|X) = P(X|Y)\frac{P(Y)}{P(X)}$ which allows to to take a prior belief about the truth of Y obtain some evidence X about Y which will have a probability $P(X|Y)$ and so we can now update our belief in the face of the new evidence such that our *prior* $P(Y)$ can be updated to $P(Y|X)$ refered to as our *posterior* belief.

Let's then dive in and look at our linear regression model within the Bayesian formalism.

Firstly we will assume that we know what the noise variance, σ^2, is for the sake of clarity in presentation of the main ideas, a full Bayesian analysis would also do inference on both the regression coefficients and noise variance but the analysis becomes cluttered with details which do not help in getting over the important concepts.

Now remember that we know the value of σ and the input data $\mathbf{X}$ is given to us so there is no uncertainty in these and as such we will only reason about the target values $\mathbf{t}$ and the parameters $\mathbf{w}$ so the joint probability of everything associated with our model can be written and decomposed as below.

$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)p(\mathbf{w}) = p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)p(\mathbf{t}|\mathbf{X}, \sigma)$$

So using the expressions above we can invert our probabilities to obtain

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

So our posterior distribution over the parameter values can be seen to be taking the data likelihood (which we maximised to obtain our MLE) $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ and weighting with it the prior parameter probability distribution $p(\mathbf{w})$[4] and normalising with $p(\mathbf{t}|\mathbf{X}, \sigma)$ which is called the marginal likelihood[5]. So our posterior distribution for the parameters can be seen as the prior belief being updated after we observe our data so in other words.

$$\mathsf{posterior} = \frac{\mathsf{likelihood} \times \mathsf{prior}}{\mathsf{marginal\ likelihood}}$$

2.2 Defining the Prior

Let us say that our model parameters are simply $\mathbf{w} = w_0, w_1$ now what range of values would we expect the parameters to reasonably take *prior* to seeing any data?, are there any values or ranges of values that would be more desirable than others? What we are now doing is *defining our prior* for the model parameters.

Let's say that before seeing any data we would prefer some parameter values to be small, this is a sensible strategy especially when there are many possibly redundant parameter values. We are perfectly free to make whatever assumptions are most appropriate at this point and in this instance we will assume that all our parameter values will follow a Gaussian distribution with a mean of zero and a standard deviation of α. This encodes that we would prefer small parameter values a priori, we also assume that the parameters are *a priori* independent of each other so $w_0 \sim \mathcal{N}(0, \alpha)$ and likewise $w_1 \sim \mathcal{N}(0, \alpha)$. So $p(\mathbf{w}|\alpha) = \mathcal{N}_{w_0}(0, \alpha)\mathcal{N}_{w_1}(0, \alpha) = \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda})$ where $\mathbf{\Lambda} = \alpha\mathbf{I}$.

[4] Note in this case there is no conditioning on $\mathbf{X}$ or σ as we set the prior before seeing any data.

[5] This term arises from $p(\mathbf{t}|\mathbf{X}, \sigma) = \int p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma)p(\mathbf{w})d\mathbf{w}$ where we integrate out or *marginalise* the model parameters

2.3 From the Prior to the Posterior

Now we know[6] that the likelihood is an N-dimensional multivariate Gaussian $\prod_{n=1}^{N} \mathcal{N}_{t_n}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma) = \mathcal{N}_\mathbf{t}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})$ and so we can write the posterior as

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha) = \frac{\mathcal{N}_\mathbf{t}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_\mathbf{w}(\mathbf{0}, \mathbf{\Lambda})}{\int \mathcal{N}_\mathbf{t}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_\mathbf{w}(\mathbf{0}, \mathbf{\Lambda})d\mathbf{w}}$$

There are many standard results for combinations and manipulations of multivariate Gaussians (refer to the course website for references) which can be applied in cases such as this. But in this case (and only in this case) we will work through this long-hand so that students get a feel for the mechanics of such manipulations.

The first thing to note is that the product of two Gaussians is also a Gaussian and the marginal form of the product of two Gaussians is also Gaussian so we can then see that the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha)$ will take the form of a Gaussian as well. We can drop the dependence on the denominator as it is not a function of $\mathbf{w}$ and collecting the terms dependent on $\mathbf{w}$ then we can write

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha) &= \frac{\mathcal{N}_\mathbf{t}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_\mathbf{w}(\mathbf{0}, \mathbf{\Lambda})}{p(\mathbf{t}|\mathbf{X}, \sigma)} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t}-\mathbf{X}\mathbf{w}) - \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{\Lambda}^{-1}\mathbf{w}\right) \\
&\propto \exp\left(-\frac{1}{2}\mathbf{w}^\mathsf{T}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Lambda}^{-1}\right)\mathbf{w} + \frac{1}{\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{t}\right)
\end{aligned}$$

Now the exponential term of a multivariate Gaussian can be written as

$$-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{\Sigma}^{-1}\mathbf{w} + \mathbf{w}^\mathsf{T}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$$

Comparing the components which depend on $\mathbf{w}$ then

$$\begin{aligned}
\mathbf{\Sigma}^{-1} &= \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Lambda}^{-1}\right) \Rightarrow \mathbf{\Sigma} = \sigma^2\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \sigma^2\mathbf{\Lambda}^{-1}\right)^{-1} \\
\mathbf{\Sigma}^{-1}\boldsymbol{\mu} &= \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} \Rightarrow \boldsymbol{\mu} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \sigma^2\mathbf{\Lambda}^{-1}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}
\end{aligned}$$

[6] If you are not convinced do not take my word for it work it out long-hand

and as $\mathbf{\Lambda} = \alpha \mathbf{I}$ then we see that the required posterior over the parameters is a multivariate Gaussian such that

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

where

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

This is a lovely result as we can now see that our posterior uncertainty about the model parameters $\mathbf{w}$ is fully defined by this multivariate Gaussian distribution.

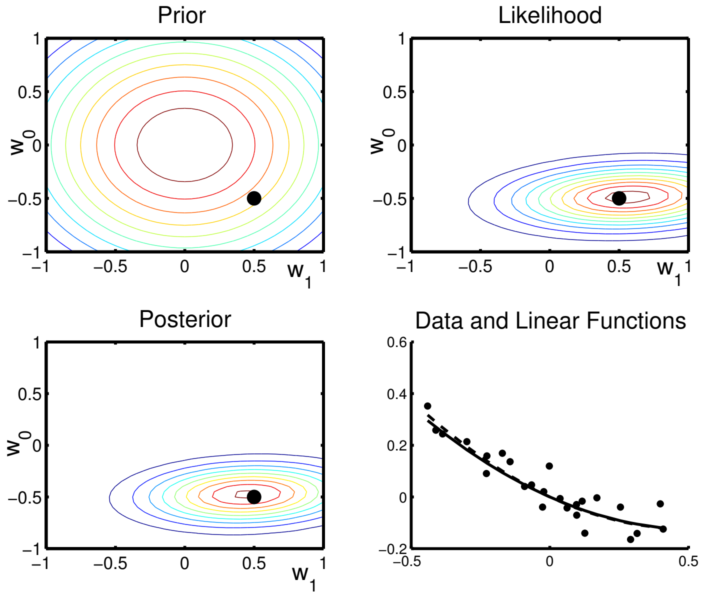


Figure 3: Top Left shows the prior distribution with the black-spot highlighting the *true* parameter values. The top right plot shows the likelihood and we can see that it is concentrated around the true values. The bottom left shows the corresponding posterior and finally the bottom right shows the data the true function and the estimated one when σ is known and α, the prior variance, is set to unity.

Figure (3) demonstrates the updating of the prior to the posterior via the likelihood for a simple model. The Matlab code to generate the above plots is available in the Week 3, Laboratory Folder, brdemo.m.

2.4 Bayesian Predictive Distributions

Now whilst in the Maximum Likelihood framework the MLE is plugged in to obtain predicted target values for a new data point in the Bayesian framework we can use our posterior distribution to average (or integrate) over our uncertainty in the possible parameter values.

$$\begin{aligned}
E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} &= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{\mathbf{x}_{new}^{\mathsf{T}}\mathbf{w}\right\} \\
&= \mathbf{x}_{new}^{\mathsf{T}}\int \mathbf{w}p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)d\mathbf{w} \\
&= \mathbf{x}_{new}^{\mathsf{T}}\boldsymbol{\mu} = \mathbf{x}_{new}^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}
\end{aligned}$$

The posterior variance in our prediction can be obtained as

$$\begin{aligned}
\mathsf{var}(t_{new}|\mathbf{x}_{new}) &= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}^2|\mathbf{x}_{new}\right\} - E^2_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} \\
&= \mathbf{x}_{new}^{\mathsf{T}}E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma)}\left\{\mathbf{w}\mathbf{w}^{\mathsf{T}}\right\}\mathbf{x}_{new} - \left(\mathbf{x}_{new}^{\mathsf{T}}E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma)}\left\{\mathbf{w}\right\}\right)^2 \\
&= \mathbf{x}_{new}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{x}_{new} = \sigma^2\mathbf{x}_{new}^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{x}_{new}
\end{aligned}$$

2.5 The Prior Provides Regularised Solutions

We should compare this with the maximum likelihood predictions and we find that the effect of the prior probability over the parameters enters into the solution via the $\frac{\sigma^2}{\alpha}\mathbf{I}$ term. Now as $\alpha \to \infty$ then we will recover the MLE prediction and this makes sense because the width of our Gaussian prior $p(\mathbf{w}|\alpha)$ will increase as α increases which means that we will become less precise about the prior values which the parameters should take and in the limit they will all become equally likely *a priori*. This has what is referred to as having a *regularising* effect on the solution.

The Matlab script regdemo.m generates noisy data which has a simple underlying linear function. The prior variance α is varied from the value 10

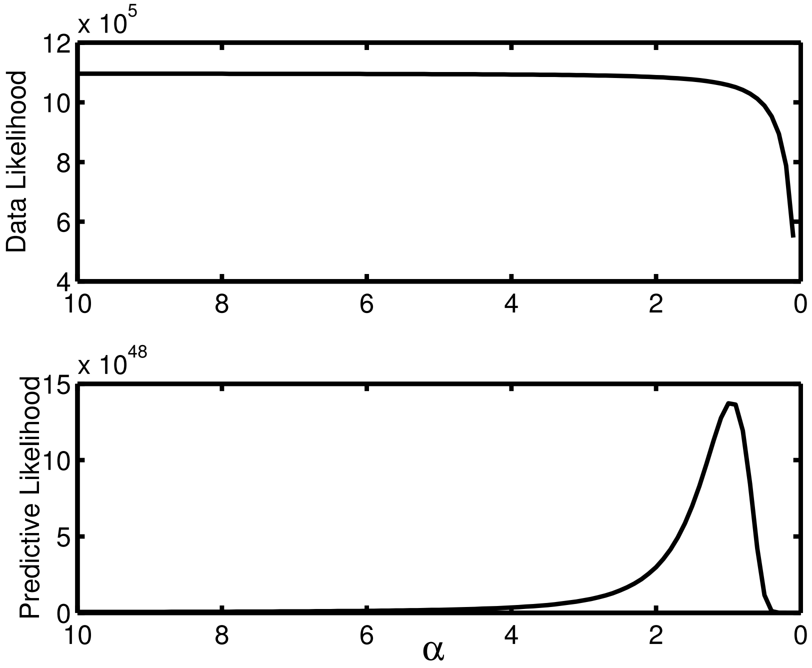


Figure 4: The top chart shows the in-sample likelihood as a function of the prior variance and we can see a drop in likelihood as the regularising effect of the prior becomes significant. The bottom chart shows how the out-of-sample predictive likelihood varies with α with a significant increase in performance at a specific α value. This is a nice example of the effect that bias & variance has on a predictive model.

to 0.1, giving a small to large range of regularisation effect. The likelihood of the *training* data is computed as is the likelihood of independent *test* data (the predictive likelihood) for each value of prior variance. The results are shown in Figure (4) and it is clear that for large values of variance we have the maximum likelihood solutions performance however as the prior variance decreases and the regularization takes effect whilst we see a drop in the data likelihood, as we are introducing a bias, the predictive likelihood increases sharply at a particular value of α. So although we are sacrificing some bias here as can be seen by the drop in data likelihood we gain by seeing a drop in the variance as can be seen by the increase in predictive likelihood. Wonderful!.

3 Conclusion

Introducing the Bayesian methodology has been an important part of our module. The linear regression examples here are nice and illustrative as the posterior and marginal distributions all take the a nice analytic form i.e. a multi-variate Gaussian. However the majority of realistic applications in Machine Learning do not permit such nice analytic forms for the posteriors, as the following weeks lectures will show. Lots of fun awaits us when we start to look at other Machine Learning methods where nice closed form Bayesian analysis is not possible.