1 Curve Fitting and Data Modelling

1.1 Slide 4

Polynomial curve fitting is a simple regression problem which we can use to motivate and illustrate a number of key concepts in pattern analysis. We observe a real-valued input variable x and wish to use this to predict the value of a real-valued target variable t. In this synthetic example, we have generated data from the function $\sin(2\pi x)$ with some random noise added to the target values. The green curve represents the 'true' function that the data was generated from. Our goal is to use the training set to make predictions $\hat{t}$ from previously unseen input values $\hat{x}$. This means that we try to discover the underlying function. This inherently difficult because we only have a finite data set (10 points in this example). The addition of noise means that there is an inherent uncertainty of the value of $\hat{t}$. Our starting point is a simple curve-fitting approach using a polynomial in x. M is the order of the polynomial, and the coefficients w_i are gathered into a vector $\mathbf{w}$.

1.2 Slide 5

How do we fit a polynomial function to the data? We create an error function that measures the misfit between the true function $y(x, \mathbf{w})$ and the training set data points. A common choice is the sum-of-squares error function shown on the slide, where the factor 1/2 is included for later convenience. It is a non-negative quantity that is zero if and only if the function $y(x, \mathbf{w})$ were to pass exactly through each training data point.

Because the error function is a *quadratic* function of the weights $\mathbf{w}$, its derivatives with respect to the coefficients are linear in the elements of $\mathbf{w}$ and so the minimisation of the error function has a unique solution $\mathbf{w}^*$ which can be found in closed form.

1.3 Slides 6–9

This group of slides illustrate the problem of choosing the order M of the polynomial. This is an example of the important concept called *model comparison* or *model selection.*

We notice that the constant and first order polynomials fit the data poorly. The third order polynomial seems to give the best fit. The higher order polynomial ($M = 9$) gives a perfect fit to the data: it passes exactly through every data point and $E(\mathbf{w}^*) = 0$. However, the fitted curve oscillates wildly and gives a very poor representation of the function $\sin(2\pi x)$. This phenomenon is known as *over-fitting.*

1.4 Slide 10

The RMS error is more convenient to use for comparison since the division by N, the number of data points, allows us to compare different sizes of data sets on an equal footing, while the square root ensures that E_{RMS} is measured on the same scale and in the same units as the target variable t.

The results shown in the graph may appear paradoxical because a polynomial of a given order contains all lower order polynomials as special cases. The $M = 9$ polynomial is capable of generating just as good results as the $M = 3$ polynomial. This shows that the limited nature of the training set and the decision to fit it as closely as possible are at the root of the problem.

1.5 Slide 11

Further insight can be obtained by examining the values of the coefficients $\mathbf{w}^*$ obtained from polynomials of various order. As M increases, the magnitude of the coefficients gets larger.

1.6 Slides 12–13

It is interesting to see how the behaviour of the model changes with the size of the dataset. For a given model complexity, the over-fitting problem becomes less severe as the size of the dataset increases. Another view of this is that as the dataset gets larger, the more complex (flexible) the model that we can afford to fit to the data.

1.7 Slides 15–17

Before we go into the complexities of Bayesian inference, we will consider the simpler (and perhaps more general) technique for control over-fitting which is *regularisation.* This involves adding an penalty term to the error function in order to discourage the coefficients from reaching large values. In this equation

$$\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 = \cdots + w_M^2$$

is the squared length/magnitude of the vector $\mathbf{w}$ and the coefficient λ governs the relative importance of the regularisation term compared with the sum-of-squares error term. The coefficient w_0 is often left out since its inclusion causes the result to depend on the choice of origin for the target variable (it is the mean of the target variable). An alternative approach is to give it its own regularisation parameter. You will also note that this discussion leaves out the question of how to determine the best value of λ. Often this is done by cross-validation: a third dataset (distinct from both training and test sets) used to optimise the model complexity.

2 Likelihood and Model Fitting

In this section we explore model fitting in greater depth and introduce the key idea of *likelihood.*

2.1 Slides 20–21

The *Gaussian* or *normal* distribution plays a fundamental role in probability statistics. One reason for this is the Central Limit Theorem. In its simplest form it states that a random variable that is the average of independent identically distributed (i.i.d. for short) random variables each of which has a well-defined finite mean and variance will tend to a normal distribution as the number of component random variables tends to infinity. This result holds no matter what the distribution of the original random variables was. (In passing, note that the commonly used normal approximation to the binomial distribution is based on this theorem, since the binomial is a sum of independent Bernoulli random variables).

2.2 Slide 22

In this context, an obvious question to ask is "Given a set of data, what is the best fitting Gaussian?"

Suppose that we have a dataset of observations $\mathbf{x} = (x_1, \ldots, x_N)^T$ representing N observations of the scalar variable x. We suppose that these observations are drawn from a Gaussian distribution whose mean μ and variance σ^2 are unknown. Because the dataset is i.i.d., we can write the probability of the dataset given μ and σ^2 as shown on the slide.

N.B. This is a function of μ and σ.

2.3 Slide 23

What we are going to do now is to find the parameters that maximise the likelihood function. It might seem more natural to maximise the probability of the parameters given the data, not the probability of the data given the parameters. It turns out that these two concepts are related.

It turns out that the calculations are easier by considering the log of the maximum likelihood. Differentiating this with respect to μ and σ in turn, and setting the values to zero, leads easily to the closed form solutions.

2.4 Slide 24

There are significant limitations to the maximum likelihood approach, as we shall see later. For now, we can look at some issues that arise for the Gaussian distribution. In particular, the maximum likelihood approach systematically underestimates the variance of the distribution. This is an example of the phenomenon called *bias* which is related to over-fitting.

We note that μ_{ML} and σ^2_{ML} are both functions of the dataset values. Consider the expectations of these quantities with respect to the dataset values, and we get the equations on the slide. On average, the true variance is underestimated by $(N-1)/N$.

In the graph, the green curve shows the true Gaussian distribution from which data is generated while the red curves show the Gaussian distribution obtained by fitting to three datasets, each consisting of two data points shown in blue, using maximum likelihood. Averaged across the three datasets, the mean is correct, but the variance is systematically under-estimated because it is measured relative to the sample mean and not relative to the true mean.

In this case, the bias is relatively small, and becomes less significant as $N \to \infty$. But in this case, we are only fitting two parameters to the data: later we shall be using models with many more parameters!

2.5 Slide 25

In the curve fitting task, we can express our uncertainty over the value of the target variable using a probability distribution. For this purpose, we will use a conditional Gaussian distribution. Here β^{-1} is the *precision* of the distribution which is the inverse variance (for consistency with later discussions).

2.6 Slide 26

We now use the training data to determine the parameters $\mathbf{w}$ and β by maximum likelihood. Note that we are doing more than in the original example, because we are also trying to estimate the variance/precision of the noise model. We assume that the data is drawn independently from the distribution.

When maximising with respect to $\mathbf{w}$ we can omit the last two terms (since they don't depend on $\mathbf{w}$) and the scaling of the function does not change where the maximum is found, so we can replace $\beta/2$ by $1/2$. This is then directly equivalent to the sum-of-squares error function. Another way of saying this is that sum-of-squares is equivalent to maximum likelihood with a Gaussian noise model.

2.7 Slide 27

Having determined the parameters $\mathbf{w}$ and β, we can now make predictions for new values of x. Because we have a probabilistic model, we obtain a *predictive distribution* that gives the probability distribution for t by substituting in the maximum likelihood estimates.

2.8 Slide 28

We can take a step towards a more Bayesian approach by placing a prior distribution over $\mathbf{w}$. For simplicity, we choose a Gaussian distribution with zero mean and precision α. Recall that $M+1$ is the number of elements in the vector $\mathbf{w}$ for a polynomial of order M. Variables such as α which control the distribution of model parameters are called *hyperparameters*. Using Bayes' theorem, the posterior distribution for $\mathbf{w}$ is proportional to the product of the prior distribution and the likelihood function.

We can now determine $\mathbf{w}$ by finding the most probable value of $\mathbf{w}$ given the data, in other words by maximising the posterior distribution. This is called *maximum a posterior* or MAP. Taking the negative logarithm of the equation, we find that the MAP solution is found by the minimum of the third equation: this is equivalent to the regularised sum-of-squares error function with regularisation parameter $\lambda = \alpha/\beta$.

2.9 Slides 29–30

A fully Bayesian treatment would integrate over all possible values of $\mathbf{w}$ weighted by the corresponding probability. This *marginalisation* is at the heart of the Bayesian approach to pattern analysis.

This is represented in the first equation on the slide (omitting the dependence on α and β for simplicity). It turns out that this can be evaluated analytically (that is not usually the case: Gaussians make our life a lot easier here!) and the predictive distribution is also Gaussian with mean $m(x)$ and variance $s^2(x)$.

In the equation for $s^2(x)$, the first term is the noise on the target variables but the second term arises from the uncertainty in the parameters $\mathbf{w}$ and is a consequence of the Bayesian treatment. The consequence of this additional uncertainty can be seen on the graph in slide 30. Here the hyperparameters $\alpha = 5 \times 10^{-3}$ and $\beta = 1$ (which is cheating a bit since it is the known noise variance). The red curve denotes the mean of the predictive distribution and the red region corresponds to ±1 standard deviation around the mean.

2.10 Slide 31

The curve fitting approach we have described works pretty well with a single input variable x. In fact, there probably wouldn't be such a subject as machine learning.

But in practice, we usually have to deal with spaces of high dimensionality comprising many input variables. We can see the problem that arises by considering a region of space divided into regular cells. The number of such cells grows exponentially with the dimensionality of the space. The problem with an exponential number of cells is that would need an exponentially large quantity of training data to ensure that the cells are not empty.

2.11 Slide 32

A general polynomial with coefficients up to order 3 has the form given on the slide. As D increases, the number of independent coefficients grows proportionally to D^3. In practice, we might need to use a higher-order polynomial. For a polynomial of order M, the growth in the number of coefficients as like D^M. Although this is a power law rather than exponential, as D gets larger, we might have to increase M as well. At any rate, the method will rapidly become unwieldy and impractical.

Consider the behaviour of a Gaussian distribution in a high-dimensional space. If we transform from Cartesian to polar coordinates and integrate out the directional variables, we obtain an expression for the density $p(r)$ as a function of radius r from the origin. Thus $p(r)\delta r$ is the probability mass in a thin shell of thickness δr located at radius r. This distribution is plotted for various values of D in the figure. We see that for large D the probability mass of the Gaussian is concentrated in a thin shell quite a long way from the origin.