Machine Learning Module

Week 1

Lecture Notes 1 & 2

Introduction & Linear Regression

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

January 8, 2006

1 Introduction

Machine Learning is fast becoming one of the most important areas of research & development activity in Computing Science[1] with companies such as *Microsoft, Google, Yahoo & Amazon* as well as major international banks & financial institutions actively recruiting Machine Learning specialists in an ongoing basis and supporting major research & development groups dedicated to Machine Learning research and its application in a number of domains.

Machine Learning is a massively interesting area to study as the subject is being developed by research questions and applications with diverse backgrounds. For example **Computing Science** contributes to Machine Learning from the field of *Artificial Intelligence* and specifically *Neural Computing*; **Statistics** has an important role to play in that *Likelihood Based Inference* and *Bayesian Statistical Inference* have become the cornerstone of many Machine Learning methods; **Physics** has contributed *Monte Carlo* methods which have been adopted in Machine Learning to enable large-scale application of *Bayesian Inference* in *Information Retrieval* and *Image Processing* to name but two; **Engineering** many engineers have contributed to the area of Machine Learning by providing challenging applications in *Control* engineering where nonlinear systems have been modeled by *Artificial Neural Networks* and *Gaussian Processes*.

A large amount of Machine Learning focuses on important theoretical mathematical and statistical issues but this course will concentrate on introducing concepts and methods which have direct applications of interest.

Above all else studying Machine Learning is an awful lot of fun.

1.1 Important Applications of Machine Learning Methods

The number of practical examples of Machine Learning applications is huge and if we just restrict ourselves to learning problems associated with the main research groups within the Department of Computing Science then an interesting and exciting list emerges.

[1] For example *The Journal of Machine Learning Research* has the highest rating (2004 & 2005) for a journal in artificial intelligence, automation and control, or statistics and probability. It is the second highest rating of any computer science journal.

- Bioinformatics
  - Predicting the interaction of genes within an organism
  - Inferring gene & protein network structures
  - Predicting protein function from sequence
- Computer Vision & Graphics
  - Image reconstruction from degraded images
  - Object detection and localisation
  - Visual tracking
- Networked Systems Measurement & Control
  - Autonomic network management systems
  - Detecting network level packet patterns
  - Intrusion detection systems
- Human Computer Interaction
  - Speech recognition
  - System control via auditory feedback
  - Gesture recognition
- Information Retrieval
  - New topic identification in news feeds
  - Language Models for *ad hoc* retrieval
  - Image & video retrieval
- Software Engineering & Technology
  - Compilers that learn to optimise (Edinburgh)
  - Automatic classification of software behaviour
- Formal Analysis, Theory & Algorithms An area that Machine Learning has not broken into yet.

1.2 Structure of Module

The two areas of Machine Learning which this introductory course will focus on are **Supervised Learning** and **Unsupervised Learning**. The first five weeks will concentrate on **Supervised Learning** where the *machine* will *learn* under the supervision of a *teacher* providing reward when the machine gets the assigned task correct and punishment when it does not. The main areas of supervised learning which we will study will be methods for *Regression* and **Classification**. The second half of the course will focus on *Unsupervised* learning techniques where the machine will learn in the absence of a teacher.

There will be two lectures each week and one laboratory session.

2 Linear Regression

An important and general problem in Machine Learning, which has wide application, is *learning* or *inferring* a functional relationship between a set of attribute variables and associated response or target variables.

To begin with we will consider the most straightforward of *learning* problems, Linear Regression[2]. We will use a practical example to introduce the main concepts of linear regression modeling. The plot of Figure 1 shows the gold medal winning distance in the long jump event at each of the Olympic Games held since 1896. Our aim is to use the data available to *learn* a model of the functional dependence (if one exists) between the time elapsed since the first modern games were held and the distance which would win gold in the long jump and use this model to make predictions about the winning distances in future games. Clearly if we were going to use these predictions to do something serious e.g. try and make some money through betting on the winning distances, then there is much more information available that can be taken into account when devising such a predictive model for making predictions (the recent form of the main competitors for example). But we take this simple example to introduce and develop the main ideas of linear regression models.

[2]The term *regression* was originally used in the context of genetics by Francis Galton (1877) when studying how intelligence is passed on (or not as the case may be) from generation to generation. The term was then adopted by statisticians who developed Galton's work within a statistical context.

2.1 Defining the Model

We can begin by defining our model as a function which maps our *input* attributes, in this case length of time elapsed, to our *output* or target values. A class of functionals which maps integers ($\mathbb{Z}$) to the Real line ($\mathbb{R}$) has to be considered such that

$$f : \mathbb{Z} \rightarrow \mathbb{R}$$

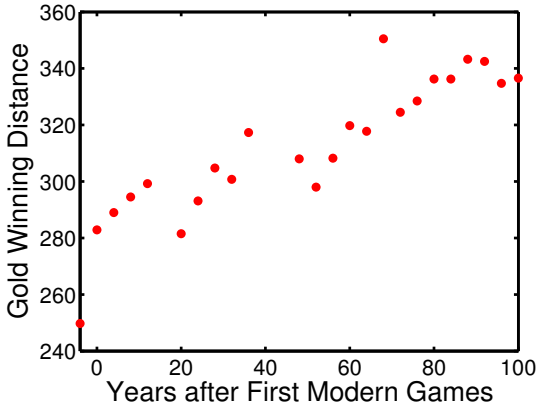

Figure 1: Gold Medal Distance for the long jump from 1896 to 2004 plotted against the number of years since the first modern games were held with 1900 being 0 and 1896 being -4. Note that the two world wars interrupt the games in 1914, 1940 & 1944.

2.1.1 Modeling Assumption

Now we make our first *modeling assumption* by assuming that a straight line can adequately model the functional relationship between time elapsed, which we will denote as x, and winning distance, denoted by t. In other words the functional mapping

$$f(x; w_0, w_1) = w_1 x + w_0$$

defines our model. The slope w_1 and the intercept w_0 are the *free parameters* of our model which have to be assigned appropriate values in some way.

2.2 Loss Functions

We *identify the model parameters* by considering a *Loss Function* which defines a measurable indicator of the miss-match between our model output $f(x; w_0, w_1)$ and the actual target value t for *all* available input-output example pairs (x_n, t_n) where $n = 1, \cdots, N$ and in this case $N = 25$, the number of game results recorded. The sample average loss is given as

$$\frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(t_n, f(x_n; w_0, w_1))$$

Clearly we will select the values of the model parameters w_0 & w_1 to minimise the average loss incurred in modeling the targets by the linear function chosen. Now we have to consider a specific form of our loss function before we can go on and seek to minimise the average (over the examples available) model loss.

2.2.1 Mean-Squared-Error Loss

For a regression model such as this the *mean-squared error* (MSE)

$$\frac{1}{N} \sum_{n=1}^{N} |t_n - f(x_n; w_0, w_1)|^2 \tag{1}$$

is an appropriate loss function to use when identifying the model parameters. Indeed minimisation of the MSE is the basis of the venerable *Least-Squares* errors method of function approximation originally developed by Gauss & Legendre (1809) when predicting planetary motion. We will see in further lectures that the MSE actually has a probabilistic basis which we will employ throughout the module.

There are numerous other *loss-functions* which could be considered, for example the class of, what are know as *robust* losses, based on absolute deviations (the absolute value of the error). These loss-functions are particularly useful when there are many *outliers* in the data as only the sign of the deviation affects the optimal solution. However the main concepts we wish to introduce at this point can be covered with MSE.

2.2.2 Matrix Notation

In this course we will make liberal use of matrix notation as it will simplify much of the manipulations which will follow in the material to be covered. The student is **very strongly advised** to brush up on matrix manipulations and some useful resources have been made available on the class website (e.g. The Matrix Cookbook).

We can define the 2×1 dimensional column vector $\mathbf{w}$ and the $N \times 1$ dimensional column vector $\mathbf{t}$ such that

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \& \quad \mathbf{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

and finally the $N \times 2$ dimensional matrix $\mathbf{X}$ is defined as

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

Denoting the inner-product of two D-dimensional column vectors $\mathbf{a}$ & $\mathbf{b}$ as $\mathbf{a}^\mathsf{T}\mathbf{b} = \sum_{d=1}^{D} a_d b_d$, where a_d & b_d are the d-th elements of the respective vectors, then it is easy to show[3]the MSE (Equation 1) can be compactly written in matrix format as

$$MSE = \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w}) \tag{2}$$

2.3 Minimising Mean-Squared-Error

Now then we need to find the parameter set which will minimise MSE and the application of some straightforward calculus will enable us to do this. The minimum of MSE can be found by finding the stationary points with respect to the parameters w_0 & w_1, that is the point where the gradient of MSE is zeros.

[3]Convince yourself that this is the case by deriving the matrix expression for MSE.

$$\frac{\partial MSE}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial MSE}{\partial w_0} \\ \\ \frac{\partial MSE}{\partial w_1} \end{bmatrix} = \begin{bmatrix} -\frac{2}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1)) \\ \\ -\frac{2}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0} \tag{3}$$

As with the expression for MSE the corresponding gradient can be neatly written in matrix format[4]

$$\frac{\partial MSE}{\partial \mathbf{w}} = -\frac{2}{N}\mathbf{X}^{\mathsf{T}}(\mathbf{t} - \mathbf{X}\mathbf{w}) \tag{4}$$

Now the equations $\frac{\partial MSE}{\partial \mathbf{w}} = \mathbf{0}$ define a stationary point of the MSE and calling to mind our calculus for single parameter functions, if the second derivatives are strictly positive then this stationary point will be a (local) minimum. For multi-parameter functions then the matrix of all partial second-derivatives requires to be *positive-definite*[5] for the stationary point to be a (local) minimum.

$$\frac{\partial^2 MSE}{\partial \mathbf{w}\partial \mathbf{w}^{\mathsf{T}}} = \begin{bmatrix} \frac{\partial^2 MSE}{\partial w_0 \partial w_0} & \frac{\partial^2 MSE}{\partial w_0 \partial w_1} \\ \\ \frac{\partial^2 MSE}{\partial w_1 \partial w_0} & \frac{\partial^2 MSE}{\partial w_1 \partial w_1} \end{bmatrix} = \begin{bmatrix} 2 & \frac{2}{N}\sum_{n=1}^{N} x_n \\ \\ \frac{2}{N}\sum_{n=1}^{N} x_n & \frac{2}{N}\sum_{n=1}^{N} x_n^2 \end{bmatrix} \tag{5}$$

As will become usual in this course we can write the matrix of second-derivatives, also referred to as the *Hessian* matrix, succinctly as

$$\frac{\partial^2 MSE}{\partial \mathbf{w}\partial \mathbf{w}^{\mathsf{T}}} = \frac{2}{N}\mathbf{X}^{\mathsf{T}}\mathbf{X} \tag{6}$$

which follows from differentiation of Equation (4) with respect to $\mathbf{w}$.

[4] Have some more fun and convince yourself that this is the case.

[5] A symmetric matrix $\mathbf{H}$ is positive-definite if for any appropriately dimensioned vector $\mathbf{a}$ then $\mathbf{a}^{\mathsf{T}}\mathbf{H}\mathbf{a} > 0$, in other words $\mathbf{H}$ has eigenvalues which are greater than zero and so $\mathsf{det}(\mathbf{H}) > 0$.

2.4 Least-Squares Solution

Now if $N >> 2$ then $\mathbf{X}^\mathsf{T}\mathbf{X}$ will be positive-definite and so we can say that the stationary point is the (global) minimum of MSE[6]. Therefore

$$\frac{\partial MSE}{\partial \mathbf{w}} = -\frac{2}{N}\mathbf{X}^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w}) = \mathbf{0} \Rightarrow \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} = \mathbf{X}^\mathsf{T}\mathbf{t} \tag{7}$$

As the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive-definite it can be inverted and so we obtain the estimate $\widehat{\mathbf{w}}$[7] for the set of parameters which minimise the MSE fit between our linear model and the target values.

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \tag{8}$$

The Least-Squares solution is

$$\widehat{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 276.78 \\ 0.748 \end{bmatrix}$$

2.5 Making Predictions

With this parameter estimate our predictions for the given target values $\widehat{\mathbf{t}}$ follow as

$$\widehat{\mathbf{t}} = \mathbf{X}\widehat{\mathbf{w}} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \tag{9}$$

Figure (2) shows the line of best fit in the least-squares sense to the data under consideration. To make *new* predictions, say for example the winning distance at the London 2012 Olympic Games denoted by $\widehat{t}_{2012}$, then

$$\widehat{t}_{2012} = \mathbf{x}_{2012}^\mathsf{T}\widehat{\mathbf{w}} = [1 \;\; 112]\widehat{\mathbf{w}} = [1 \;\; 112]\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \tag{10}$$

So based on our linear regression model we can expect a gold medal winning distance of $276.78 + 0.748 \times 112 = 360.5$ inches in London. The current Olympic record, set by Bob Beamon in 1968, stands at 350.39 inches and the current World Record was set in 1991 by Mike Powell with a distance of 352.36 inches. It would appear that our model is somewhat optimistic and we have a rather long wait for verification of our prediction.

[6]There will be many instances where the matrix is not positive-definite and so not uniquely invertible. In such a case solutions can still be obtained however they will not be unique and this throws up other problems such as how do we choose the most appropriate solution? we will consider these problems later in the course.

[7]The *hat* notation is used to denote an estimate for the parameters.

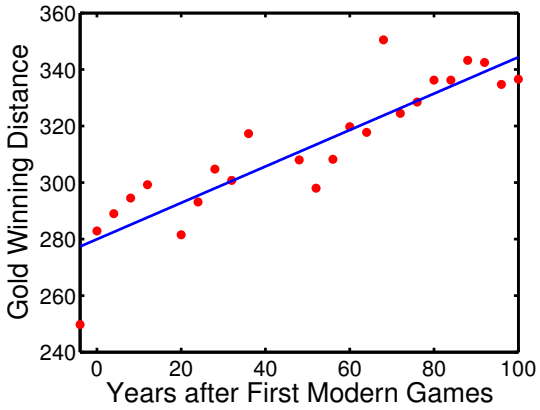

Figure 2: Least Squares fit of linear model to Olympic long Jump Gold data.

2.6 Nonlinear Response from a Linear Model

The least-squares solution provides the best fit in the sense of minimising the average squared missmatch between our linear model and the data. The model is linear in the sense that it is a linear function of the associated parameters but says nothing of any transformations of the input variables. For example we may believe that the underlying function mapping time-elapsed and distance achieved is a polynomial function (due to for example cyclic levels of performances caused by various world events e.g. wars) such that for example a cubic polynomial relation holds i.e.

$$f(x; \mathbf{w}) = w_3 x^3 + w_2 x^2 + w_1 x + w_0 \tag{11}$$

or more generally an arbitrary K'th order polynomial holds

$$f(x; \mathbf{w}) = \sum_{i=0}^{K} w_i x^i \tag{12}$$

It should be straightforward to see that by now defining the $N \times (K+1)$ dimensional matrix $\mathbf{X}$ such that

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^K \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^K \end{bmatrix} \tag{13}$$

then the best polynomial fit, in the least-squares sense, will still be obtained using Equation (8) where now $\widehat{\mathbf{w}}$ will be a $(K+1) \times 1$ column vector. Figure (3) shows the fit of a $K = 9$ - order polynomial model of the long-jump data, does this look to be a better fit to the underlying trend than the $K = 1$ - order model? Is it a better predictor than a strictly linear model?

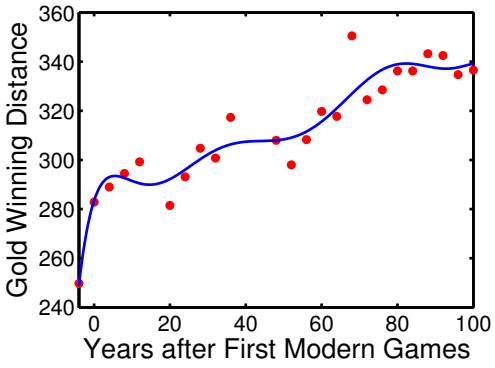

Figure 3: Least Squares fit of 9'th order polynomial model to Olympic long Jump Gold data.

To answer these questions we have to consider objective assessments of the goodness of our models in terms of how faithfully they represent the data generating process whatever that may be.

The laboratory exercise will examine these questions in some detail and a tutorial sheet is available to assist in the vector and matrix manipulations.