Representing Penalties on Basis Expansions as a Polynomial In the Parameters (Working Title)

February 25, 2013

1 Introduction To Basis Function Expansion Methods

One of the major fields in Statistics is that of fitting functions to data. A very large class of methods are known as *Basis Function Expansion Methods*, where we estimate the output as a weighted sum of basis functions. These models generally take a form as follows:

$$\mathbb{E}[y|t] = c_0\phi_0(t) + c_1\phi_1(t) + \cdots + c_N\phi_N(t)$$

Here we have our finite set of basis functions $\{\phi_k\}$ along with our parameters $\{c_k\}$, $\mathbb{E}[y|x]$ is the conditional expectation of our independent variable y, given our dependent data x. Roughly $\mathbb{E}[y|x]$ can be thought of as our best estimate of the value of y given x.

Basis Function Methods might seem a little abstract, but they are ubiquitous. For example, Simple Linear Regression is an example of a basis function method. Take two basis functions $\{1, t\}$, so that $\hat{y}(t)$, our estimate of y given the value t, can be written as a linear combination of the two:

$$\mathbb{E}[y|t] = \hat{y}(t) = c_0 + c_1 t$$

This is the form of a simple linear regression model. If our basis consists of just the single constant function $\phi(t) = 1$ then we get a model of the form:

$$\hat{y} = c_0$$

In this case we would generally use the mean or the median of the y values as our estimate of c_0. If we go in the other direction and add a quadratic function t^2 we get a quadratic regression model:

$$\hat{y}(t) = c_0 + c_1 t + c_2 t^2$$

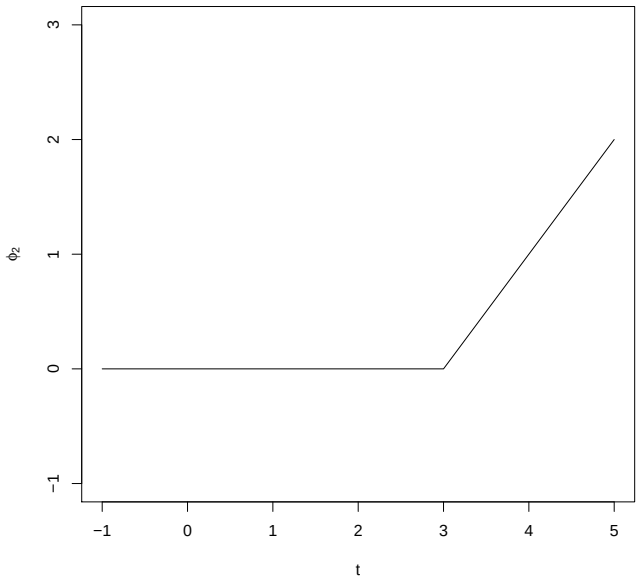

Figure 1: The Ramp Function ϕ_2

There are other choices of basis besides monomials we could use. We could have a basis of two functions $\{\phi_1, \phi_2\}$ Where the two functions are defined as follows:

$$\phi_1(t) = 1$$
$$\phi_2(t) = \begin{cases} 0 & \text{if } t < \tau \\ t - \tau & \text{if } t \geq \tau \end{cases}$$

The first function is a constant function, the second is a ramp function based at τ.

The above basis has two disadvantages though. Firstly the second basis function is only continuous, not differentiable. It would not be wise to use this basis if we wanted to estimate $\frac{dy}{dt}$. Secondly it is very arbitrary, it is not obvious why it would be useful

This underscores an important point. Real Analysis tells us that there are many sets of functions that are *dense* on some interval$[a, b]$, that is they can approximate any function on $[a, b]$ to arbitrary accuracy. We must be sensible in our choice of basis. If our choice of basis is good, then it will be able to fit the data with only a few terms, and we might be able to avoid estimating many coefficients. This is one of the reasons simple linear regression and quadratic

regression are useful; they can capture much variation in the observed data in spite of being very simple.

1.1 Fourier Basis Functions

If our data has a periodic component to it, such as the observed temperature over the course of the year, or a time series of annual sales, then it would be wise to use a basis consisting of periodic functions. This suggests that we should use a *Fourier Basis* consisting of the set of functions $\{cos(n\omega t), sin(n\omega t)\}$ where $0 \leq n \leq N$ for some N and ω is the frequency. The frequency depends on our time scale, they are related by the formula $\omega = \frac{2\pi}{T}$, where T is the period.

Furthermore Fourier basis functions have several other desirable properties. They are smooth, meaning that they can be used to estimate any derivivative of the data, at leastin theory. They are orthogonal, which can make certain problems more convienient and they are closed under differentiation, meaning that the derivative of a combination of Fourier basis functions, is itself a combination of Fourier basis functions. The latter two properties will prove very useful later, as we shall see.

1.2 B Spline Basis Functions

Roughly speaking, *B Splines* are compactly supported polynomial functions, or more practically they are nonzero only inside of a given interval. More formally a B-Spline basis consists of a *degree* n, which determines the degree of the basis functions, and a set of *knots*, that is a set of K time points $\{t_0, \ldots, t_K\}$.

Since they are compact they generally can only indvidually capture local information about the data. One of the main advantages of B Splines is that they can represent any other spline of the same degree and smoothness with the same knots. This makes them useful for Statistics since they assume less about the form of the data, they can help us avoid bias.

2 Least Squares Fitting

Least Squares is one of the most well known statistical estimation techniques. If we have a set of n observations y_i, measured at times t_i, then the *Least Squares Criterion*, chooses the estimated function $\hat{y}$ from some set of functions S that minimises the sum of the squares of the error:

$$\hat{y(t)} = \underset{f \in S}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f(t_i))^2$$

We assume S is the span of some set of m basis functions i.e $s = \{\sum_{i=1}^{m} c_i \phi_i(t) | c_i \in \mathbb{R}\}$. This suggests that to find $\hat{y}$ we only need to estimate to coefficients $\hat{c}_i$. Then we have completely deterimined $\hat{y}$. We can then write the least squares criterion as:

$$(\hat{c}_1, \ldots, \hat{c}_m) = \underset{c_1 \ldots c_K}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} c_j \phi_j(t_i))^2$$

The above expression is cumbersome, we can use vector notation to simplify it. Firstly, we have $f(t) = \mathbf{c}'\phi(\mathbf{t})$, where $\mathbf{c} = (c_1, ..., c_m)$ and $\phi = (\phi_1(t), \ldots, \phi_2(t))'$. If we construct a matrix $\mathbf{\Phi}$, where the ith row of $\mathbf{\Phi}$ is $\phi(t_i)$, and let $\mathbf{y} = (y_1, \ldots, y_n)$ we can write the least squares problem as:

$$\hat{\mathbf{c}} = \underset{\mathbf{c} \in \mathbb{R}^m}{\operatorname{argmin}} (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c})$$

The expression on the right hand side is an example of a *Quadratic Form*, they are the generalisation of quadratic functions to finite dimensional vector spaces. We can have quadratic forms on infinite dimensional vectors spaces too, but that is not relevant here.

3 Roughness Penalties

It is well known that the Least Squares gives us the *Best Linear Unbiased Estimator* for $\mathbf{y}$ the function we assume to be generating the data. Nonetheless it is often useful to employ a form of *regularisation* which constrains how much $\hat{y}$ is allowed to vary. Intuitively, this reduces the *variance* of $\hat{y}$ and helps guard against overfitting. In the context of basis function expansions, the most commonly used penalty is the curvature penalty:

$$PEN(\hat{y}) = \int_D [\hat{y}(t)'']^2 dt$$

Here D is our domain of interest and f'' stands for the second derivative of the function f. This criterion for fitting a curve is very appealing in many regards. It can be shown using the theory of Finite Elements that if we constrain $\hat{y}$ to pass through two points, then the $\hat{y}$ minimises the penalty is the best approximation to the straight line going through the points that we can form with this basis.

This penalty can also be represented as a polynomial. Let $\langle f, g \rangle = \int_D f(t)g(t)dt$. We can see that $\langle \cdot, \cdot \rangle$ defines an inner product on $\mathrm{L}^2(D)$, the set of square integrable functions on D. The penalty can be written in the form $PEN(f) = \langle f'', f'' \rangle$. Substituting in the expansion for f we get:

$$\langle \hat{f}'', \hat{f}'' \rangle = \langle \sum_{i=1}^{m} c_i \phi_i'', \sum_{i=1}^{m} c_i \phi_i'' \rangle \quad = \sum_{i=1}^{m} \sum_{j=1}^{m} \langle c_i \phi_i'', c_j \phi_j'' \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \langle \phi_i'', \phi_j'' \rangle$$

We can see that the terms $\langle \phi_i'', \phi_j'' \rangle$ depend only on our choice of basis and so are "fixed" for our purposes. This implies that the penalty can be represented as a penalty in the c_i. We can do better however and represent the penalty as

a quadratic form. If we define an $m \times m$ matrix $\mathbf{K}$ by $\mathbf{K}_i j = \langle \phi_i'', \phi_j'' \rangle$ it can be seen that $\langle f'', f'' \rangle = \mathbf{c}'\mathbf{K}\mathbf{c}$.

Note that this relies on the fact that the second derivative operator maps a linear combination of ϕ_i into a linear combination ϕ_i''. We can always represent an inner product on some finite vector space as $\mathbf{b}'\mathbf{A}\mathbf{b}$, where the terms depend on the problem at hand. We can simply construct the matrix $\mathbf{A}$ for the ϕ_i'' and have the $\mathbf{c}$ in place of the $\mathbf{b}$ because differentiation is linear.

4 Non Linear Penalties