Linear models - splines

Data analytics

Jerzy Baranowski

Regression with design matrix

Or what we have established so far

- Returning to our model formulation

$$\text{outcome}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta \cdot \text{predictor}_i$$

- There is no problem with adding more predictors (or functions of predictors)

$$\text{outcome}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \mathbf{x}_i \cdot \beta$$

- Which can be represented in vector matrix notation

$$\mathbf{outcome} \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \mathbf{X} \cdot \beta$$

Design matrix

So what can we put into design matrix?

Complication in regression

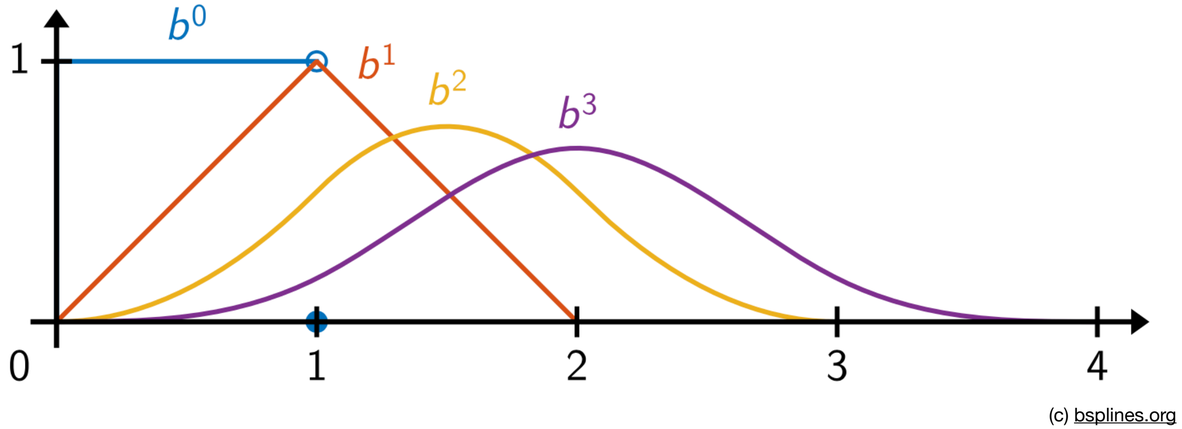
- First of all we can use multiple different predictors. We won’t talk about this in detail, but you will be given material to read.
- We have used polynomials already
- Another options are B-splines

B-spline approximation

Making function out of pieces

- Splines are functions on so called compact support
- That means that function outside certain interval (compact support) is equal to zero.
- B-splines (basis splines) are polynomials (or similar to them) on the support, while from order 1+ they are zero at the edges.

Cardinal B-splines



(c) bsplines.org

B-splines are defined by **knots**, that is point spanning on the support.

B-spline approximation

So what with knots?

- We locate knots in the are of interest and select the orders polynomial consisting of B-splines.
- Full basis of for example third order splines on considered interval consists of the functions that are nonzero at 4 knots and zero outside of them. If all such knots are different this function is differentiable. On the edges we need to repeat some knots (losing differentiability).

B-splines approximation

Construction of approximation

- Function of interest is approximated by a sum

$$f(x) \approx \beta_1 B_1(x) + \beta_2 B_2(x) + \beta_3 B_3(x) + \dots$$

- So generally it is very similar to regression

B-spline approximation

How to select knots

- Up to debate, if we have evenly distributed samples (for example time series samples) uniform distribution of knots might be a good idea.
- If samples are not evenly distributed, one rule of thumb is to locate knots at quantiles of the samples. This gives us more knots where there are more datapoints.

B-spline approximation

Why to do it?

- Splines can express local behavior better (localized effects might be expressed by few base functions)
- Splines result in sparse design matrices which can be exploited for large problems
- Splines can be considered a special case of Gaussian Processes.

Construction of design matrix

Go to the solution

- Situation is very similar to normal regression

$$\mu_i = \beta_1 B_1(x_i) + \beta_2 B_2(x_i) + \dots$$

- So that gives a design matrix:

$$\mathbf{X} = \begin{bmatrix} B_1(x_1) & B_2(x_1) & \cdots & \cdots \\ B_1(x_2) & B_2(x_2) & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_N) & B_2(x_N) & \cdots & \cdots \end{bmatrix}$$

Example

Cherry blossoming day

- Cherry trees blossom all over Japan in the spring each year, and the tradition of flower viewing (Hanami 花見) follows. The timing of the blossoms can vary a lot by year and century.
- We will consider dataset of 1000 years.

Photo by Bagus Pangestu