1 Nonparametric regression

Suppose $Y \in \mathbb{R}$ and we are trying to nonparametrically fit the regression function

$$\mathbb{E}Y|X = f_*(X)$$

A common approach (particularly when p is small) is to specify

- A fixed basis, $(\phi_k)_{k=1}^{\infty}$
- A tuning parameter K

We follow this prescription:

1. Write[1]
$$f_*(X) = \sum_{k=1}^{\infty} \beta_k \phi_k(x)$$
where $\beta_k = \langle f_*, \phi_k \rangle$
2. Truncate this expansion[2] at K
$$f_*^K(X) = \sum_{k=1}^{K} \beta_k \phi_k(x)$$
3. Estimate β_k with least squares

The weaknesses of this approach are:

- The basis is fixed and independent of the data
- If p is large, then nonparametrics doesn't work well at all
- If the basis doesn't 'agree' with f_*, then K will have to be large to capture the structure $(f_* = \sum_{k=1}^{\infty} \langle f_*, \phi_k \rangle \phi_k)$
- What if parts of f_* have substantially different structure?

An alternative would be to have the data tell us what kind of basis to use

[1] Technically, f_* might not be in the span of the basis, in which case we have incurred an irreducible approximation error. Here, I'll just write f_* as the projection of f_* onto that span

[2] Often higher k are more rough $\Rightarrow$ this is a smoothness assumption

2 Neural networks

2.1 Definitions

$$L(\mu(X)) = \beta_0 + \sum_{k=1}^{K} \beta_k \sigma(\alpha_{k0} + \alpha_k^\top X)$$

The main components are

- The derived features $Z_k = \sigma(\alpha_{k0} + \alpha_k^\top X)$ and are called the hidden units
  - The function σ is called the activation function and is very often $\sigma(u) = (1 + e^{-u})^{-1}$ (This particular $\sigma(u)$ is known as the sigmoid function)
  - The parameters $\beta_0, \beta_k, \alpha_{k0}, \alpha_k$ are estimated from the data.
- The number of hidden units K is a tuning parameter

2.2 Observation 1: Feature map

We start with p covariates

We generate K features

$$\begin{aligned}\Phi(X) &= (1, x_1, x_2, \ldots, x_p, x_1^2, x_2^2, \ldots, x_p^2, x_1 x_2, \ldots, x_{p-1} x_p) \in \mathbb{R}^K \\ &= (\phi_1(X), \ldots, \phi_K(X))\end{aligned}$$

Before feature map:

$$L(\mu(X)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

After feature map:

$$L(\mu(X)) = \beta^\top \Phi(X) = \sum_{k=1}^{K} \beta_k \phi_k(X)$$

For neural networks write:

$$Z_k = \sigma\left(\alpha_{k0} + \sum_{j=1}^{p} \alpha_{kj} x_j\right) = \sigma\left(\alpha_{k0} + \alpha_k^\top X\right)$$

Then we have

$$\Phi(X) = (1, Z_1, \ldots, Z_K)^\top \in \mathbb{R}^{K+1}$$

and

$$\mu(X) = \beta^\top \Phi(X) = \beta_0 + \sum_{k=1}^{K} \beta_k \sigma\left(\alpha_{k0} + \sum_{j=1}^{p} \alpha_{kj} x_j\right)$$

2.3 Observation 2: Activation function

If $\sigma(u) = u$ is linear, then we recover classical methods

$$\begin{aligned}
L(\mu(X)) &= \beta_0 + \sum_{k=1}^{K} \beta_k \sigma(\alpha_{k0} + \alpha_k^\top X) \\
&= \beta_0 + \sum_{k=1}^{K} \beta_k (\alpha_{k0} + \alpha_k^\top X) \\
&= \beta_0 + \sum_{k=1}^{K} \beta_k \alpha_{k0} + \sum_{k=1}^{K} \beta_k \alpha_k^\top X \\
&= \gamma_0 + \gamma^\top X \\
&= \gamma_0 + \sum_{j=1}^{p} \gamma_j^\top x_j
\end{aligned}$$