

Nonlinear regression

Mark Andrews

September 13, 2019

Basis function regression

- ▶ Given a set of n observed predictor-outcome pairs $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$, a linear model with normally distributed errors is, for $i \in 1 \dots n$,

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2).$$

- ▶ How do we model a *nonlinear* relationship between x and y ?
- ▶ One general solution is to use a weighted sum of K *basis functions* $\phi_1(x_i), \phi_2(x_i) \dots \phi_K(x_i)$ and

$$y_i = \sum_{k=1}^K w_k \phi_k(x_i) + \epsilon_i$$

- ▶ Common choices of basis functions are *radial-basis functions* and *splines*.

Radial basis functions

- ▶ A radial, or squared exponential, basis function can have the following form:

$$\phi_k(x) = \exp\left(-\frac{|x - \mu_k|^2}{l^2}\right)$$

- ▶ Given a set of centers $\mu_1, \mu_2 \dots \mu_K$, we can then model a smooth function between x and y with additive noise as

$$y \sim \sum_{k=1}^K w_k \phi_k(x) + \epsilon.$$

- ▶ If $\mu_1, \mu_2 \dots \mu_K$ are known, then Bayesian inference in radial basis function regression involves inference of $w_1, w_2 \dots w_K$, the width parameter l^2 and the observation noise term σ^2 .

Gaussian Processes

- ▶ A Gaussian process can be seen as a generalization of a multivariate Gaussian to an infinite dimensional space.
- ▶ It can be seen as a distribution over smooth functions, i.e. a random function.
- ▶ For example, in a Gaussian process over a space X , the covariance in the distribution between any two points in that space is given by a positive definite *kernel* function.
- ▶ An example of a kernel could be

$$\kappa(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{l^2}\right)$$

- ▶ The joint probability distribution over any set of points $x = x_1, x_2 \dots x_n$ is then by a multivariate Gaussian with covariance matrix $K(x_1, x_2 \dots x_n)$ where

$$K_{ij} = \kappa(x_i, x_j).$$

Gaussian Processes regression

- ▶ If $f \sim \text{GP}(\mathbf{m}, \kappa)$ is a Gaussian process, i.e. a random distribution over smooth functions, we can model a smooth function between \mathbf{x} and y with additive noise as

$$y \sim f + \epsilon,$$

with the prior on the mean function \mathbf{m} as 0.

- ▶ Given a set of n observed predictor-outcome pairs $(\tilde{x}_1, y_1), (\tilde{x}_2, y_2) \dots (\tilde{x}_n, y_n)$, the posterior distribution over f is itself a Gaussian process

$$P(f|\tilde{\mathbf{x}}, \mathbf{y}) \sim \text{GP}(\tilde{\mathbf{m}}, \tilde{\kappa}),$$

- ▶ where

$$\tilde{\mathbf{m}} = \mathbf{K}(\tilde{\mathbf{x}}, \mathbf{x})(\mathbf{K}(\tilde{\mathbf{x}}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} \mathbf{y},$$

and

$$\tilde{\kappa} = \mathbf{K}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - \mathbf{K}(\tilde{\mathbf{x}}, \mathbf{x})(\mathbf{K}(\tilde{\mathbf{x}}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} \mathbf{K}(\mathbf{x}, \tilde{\mathbf{x}}).$$