Bayesian polynomial regression

Kristian Wichmann

October 25, 2017

1 The setup

Given a training set $\mathbf{x}, \mathbf{t}$ of N points where the random t's are thought to depend of the non-random x's, we wish to find a set of weights $\mathbf{w}$ that fits the model where the t's are normally distributed with precision β and mean:

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

More generally, x^j could be substituted for another set of basis functions $f_j(x)$.

2 Maximum likelihood estimation

The usual frequentist approach would be a maximum likelihood estimation of the parameters β and $\mathbf{w}$. The likelihood function for the model is:

$$L(\beta, \mathbf{w}|\mathbf{x}, \mathbf{t}) = \prod_{n=1}^{N} \mathcal{N}(y(x_n, \mathbf{w}), \beta^{-1}) = \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{1}{2}\beta(y(x_n, \mathbf{w}) - t_n)^2\right] \tag{2.1}$$

To turn the product into the sum, find the log-likelihood[1]

$$\ell(\beta, \mathbf{w}|\mathbf{x}, \mathbf{t}) = -\frac{N}{2}(\ln\beta - \ln(2\pi)) + \frac{\beta}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 \tag{2.2}$$

Now differentiate with respect to weights and precision to find the minimum:

$$\frac{\partial \ell}{\partial w_j} = \frac{\beta}{2}\sum_{n=1}^{N} 2(y(x_n, \mathbf{w}) - t_n)\frac{\partial y_n}{\partial w_j} = \beta \sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n) x_n^j \tag{2.3}$$

[1] Here including a minus as well. Conventions might differ.

Setting equal to zero we get the usual least squares equations:

$$\sum_{n=1}^{N}(y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n)x_n^j = 0 \tag{2.4}$$

For precision:

$$\frac{\partial \ell}{\partial \beta} = -\frac{N}{2\beta} + \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 \tag{2.5}$$

Setting this equal to zero:

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 \tag{2.6}$$

3 Bayesian treatment

In the Bayesian formulation, we need to specify a *prior distribution*. Initially, without any specific information, we have no idea whether a weight should be positive or negative, so let's choose a prior where each is centered at zero. The normal distribution is an obvious pick. Assuming independence of weights and them all having the same precision α, this means our prior is:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_{M+1}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^t\mathbf{w}\right\} \tag{3.1}$$

We can now use Bayes' theorem to obtain the posterior distribution as proportional to the likelihood function (2.1) times the prior (3.1):

$$p(\mathbf{w}|\alpha, \beta, \mathbf{x}, \mathbf{t}) \propto L(\beta, \mathbf{w}|\mathbf{x}, \mathbf{t})p(\mathbf{w}|\alpha) \tag{3.2}$$

Ignoring factors, this means:

$$p(\mathbf{w}|\alpha, \beta, \mathbf{x}, \mathbf{t}) \propto \prod_{n=1}^{N} \exp\left[-\frac{1}{2}\beta(y(x_n, \mathbf{w}) - t_n)^2\right] \cdot \exp\left\{-\frac{\alpha}{2}\mathbf{w}^t\mathbf{w}\right\} \tag{3.3}$$

3.1 Maximum posterity estimation

If we're just looking for a point estimate, we can now choose the parameters which maximizes equation 3.3, a process known as *maximum posterior estimation* or *MAP* for short. Since the exponential is strictly monotonous, we can maximize the exponent to get:

$$\frac{\beta}{2} sum_{i=1}^{N}(y(x_n, \mathbf{w}_{\mathrm{MAP}}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}_{\mathrm{MAP}}^t\mathbf{w}_{\mathrm{MAP}} = 0 \tag{3.4}$$