Polynomial Curve Fitting

We start by solving a simple problem in ML. Suppose we have a *training set* of N values, that is, a vector $\boldsymbol{x} = \{x_1, ..., x_N\}$ corresponding to $\boldsymbol{t} = \{t_1, ..., t_N\}$. Now given a new value x, we want to predict its corresponding value t.

Given a set of coefficients $\boldsymbol{w} = \{w_0, ..., w_M\}$ for some $M \in \mathbb{N}$, we have a polynomial of degree M expressed as $y(x, \boldsymbol{w}) = \sum_{i=0}^{M} w_i x^i$. Now we want to approximate the relation between $\boldsymbol{x}$ and $\boldsymbol{t}$ using a polynomial. We want to minimise an *error function* that tells us how good the approximation is given coefficients $\boldsymbol{w}$. The error function evaluates to 0 if and only if it passes through every $t \in \boldsymbol{t}$. Below is a simple way to do it.

Definition. The *sum of squares* error function E is given by

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, \boldsymbol{w}) - t_n)^2.$$

Deriving this gives us a linear map with a trivial kernel, so there is a unique $\boldsymbol{w}^*$ such that $E(\boldsymbol{w}^*)$ is minimal. We still need to choose M, if it's too large then we have the problem of over-fitting, if too little, we don't have enough flexibility to fit accurately to the training set.

We could rigorously test whether over-fitting is a problem by using our function $y(x, \boldsymbol{w}^*)$ against a set with way more datapoints. Hence, we require the following definition that generalises N.

Definition. The *root-means-square* error function is defined as follows.

$$E_{\text{RMS}}(\boldsymbol{w}) = \sqrt{\frac{1}{N} E(\boldsymbol{w})}.$$

Curve Fitting Revisited

(Please read probability theory an fundamentals of Gaussian distribution first.)

We now assume the value we are trying to predict, t, has a Gaussian distribution with $\mu = y(x, \boldsymbol{w})$, and that

$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(y(x, \boldsymbol{w}), \beta^{-1}).$$

Recall that $\beta = \frac{1}{\sigma^2}$. We assume the datapoints are independent, and we maximise the likelihood function below

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}).$$

This can be done by maximising its log.

$$\ln p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - y(x_n, \boldsymbol{w})\right)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi$$

If we simplify the process of maximising $\boldsymbol{w}$ by deleting the addition of the constant terms, then let $\beta = 1$, and minimise instead the negative log of $p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta)$, it follows now that this maximisation process is exactly the minimisation of the sum-of-squares error function. Also note that

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\left(y(x_n, \boldsymbol{w}_{\text{ML}}) - t_n\right)^2.$$

We may now predict a distribution given a new (x, t). This is called a *predictive distribution.*

$$p(t|x, \boldsymbol{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \boldsymbol{w}_{\text{ML}}), \beta_{\text{ML}}^{-1}).$$

Now we wish to introduce a more Bayesian approach. Given a precision α, we define a prior.

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}}\exp\left(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right).$$

A parameter like α is called a *hyper-parameter*. The motivation is to define a probability distribution over $\mathbb{R}^{M+1}$ around the origin with a precision α. Then, by Bayes' theorem,

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha).$$

We then use *maximum posterior* or *MAP* by minimising the negative log. We find that the maximum of the posterior is given by the minimum of the following

$$\frac{\beta}{2}\sum_{n=1}^{N}\left(y(x_n, \boldsymbol{w})\right)^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}.$$

So maximising the posterior is to minimise the sum of squares error function with regularisation parameter $\lambda = \alpha/\beta$.

For a full Bayesian approach, we repeatedly apply the sum and product rules. Ultimately, we wish to find $p(t|x, \boldsymbol{x}, \boldsymbol{t})$. Hence, we write it in the following form.

$$p(t|x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t|x, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}) \; d\boldsymbol{w}.$$

Recall that $p(t|x, \boldsymbol{w}) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1})$.

We will see that

$$p(t|x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t|m(x), s^2(x)).$$

where the mean and variance are given by

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$
$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S} \boldsymbol{\phi}(x)$$

where

$$\boldsymbol{S}^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^T$$
$$\boldsymbol{\phi}(x) = (x, x^2, x^3, ..., x^M)^T.$$