Gaussian Process Regression

S.R. Khare, Poorvi Agrawal

11 August 2015

Regularization

Regularization is a technique to prevent overfitting problem by reducing model complexity. In this technique, we add an extra term to the error function which controls coefficients from attaining higher values.

$$E_D(w) + \lambda E_W(w)$$

where $\lambda E_W(w)$ is regularization term equals to $\frac{\lambda w^T w}{2}$. When λ tends to 0, we get overfitting model whereas λ tends to ∞ implies underfitting model.

One way to calculate λ is to use trial and error method. Dividing training set into training and cross validation set and try different values of λ. Compare different models based on different λ values with cross validation set and find the best value for λ. But this could be computationaly expensive. So there is an approach which can overcome the problem of selecting λ value called Bayesian approach.

Bayesian Approach

Bayesian Inference is the method in which bayes theorem is used to predict new data output (posterior predictive distribution) based on the past data (prior). It avoids overfitting problem and also lead to automatic methods for finding out model complexity. Consider a guassian prior distribution where α is distribution precision parameter and β is noise precision parameter.

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$$

Using Bayes theorem, the posterior distribution will be proportional to the product of the likelihood function and prior distribution

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

To get the best fit, we need to find the value of $\mathbf{w}$ which maximizes posterior distribution implies maximization of log of distribution.

$$\ln p(\mathbf{w}, \mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(x_n))^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const.$$

This is equivalent to the minimization of the error function with the additional quadratic regularization term where $\lambda = \frac{\alpha}{\beta}$. Practically we need to make predictions on new data. So the predictive distribution can be written as:

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)d(\mathbf{w})$$

On solving this we get the desired distribution One advantage of using Bayesian Inference is: it uses prior knowledge that results into less extreme conclusion. On the contrary if the prior analysis is done based on the poor choices then we may get bad results.