NOTES FOR MACHINE LEARNING

AARON NOTES SERIES

Aaron Xia
Department of Electronic Engineering
Tsinghua University
HaiDian District, Peking
muranqz@gmail.com

April 24, 2019

1 Linear Models for Regression

1.1 Maximum Likelihood and Least Squares

We assume that the target variable t is given by a deterministic function $y(\mathbf{x}, \mathbf{w})$ with additive Gaussian noise so that

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{1.1}$$

where ϵ is a zero mean Gaussian random variable with precision (inverse variance) β. Thus we can write

$$p(t|\mathbf{x}, \mathbf{w}, \beta) \sim \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \tag{1.2}$$

Remark 1.1 *In supervised learning problems such as regression (and classification), we are not seeking to model the distribution of the input variables. Thus x will always appear in the set of conditioning variables, and so from now on we will drop the explicit x from expressions such as $p(t|x, w, \beta)$ in order to keep the notation uncluttered.*

Making the assumption that these data points are drawn independently from the distribution (1.2), we obtain the following expression for the likelihood function, which is a function of the adjustable parameters **w** and β and taking the logarithm of the likelihood function, we have

$$\ln p(\mathbf{t}|\mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \tag{1.3}$$

Next, let's take some insight into the role of the bias parameter w_0. Since the error function is given by

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}_n))^2 \tag{1.4}$$

Setting the partial derivative with respect to w_0 to zero, and solving for w_0, we obtain

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \bar{\phi}_j \tag{1.5}$$

where we have defined

$$\bar{t} = \frac{1}{N} \sum_{i=1}^{N} t_n, \quad \bar{\phi}_j = \frac{1}{N} \sum_{n=1}^{N} \phi_j(\mathbf{x}_n). \tag{1.6}$$

Thus the bias w_0 compensates for the difference between the averages (over the training set) of the target values and the weighted sum of the averages of the basis function values.

If we maximize the log likelihood funciton (1.3) with respect to the noise precision parameter β, giving

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{ML}^T\phi(\mathbf{x}_n)\}^2 \tag{1.7}$$

The inverse of the noise precision is given by the residual variance of the target values around the regression function.

Adding a regularization term to an error function is useful to control over-fitting, so that the total error function to be minimized takes the form

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \tag{1.8}$$

where λ is the regularization coefficient that controls the relative importance of the data-dependent error $E_D(\mathbf{w})$ and the regularization term $E_W(\mathbf{w})$.

In general, the regularized error takes the form

$$\frac{1}{2}\sum_{i=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{1.9}$$

Remark 1.2 *minimizing (1.9) is equivalent to minimizing the unregularized sum-of-squares error (1.4) subject to the constraint*

$$\sum_{j=1}^{M}|w_j|^q \leq \eta \tag{1.10}$$

for an appropriate value of the parameter η, where the two approaches can be related using Lagrange multipliers.

1.2 The Bias Variance Decomposition

when we discussed decision theory for regression problems, we considered various loss functions each of which leads to a corresponding optimal prediction once we are given the conditional distribution $p(t|\mathbf{x})$. A popular choice is the squared loss function,for which the optimal prediction is given by the conditional expectation, which we denote by $h(\mathbf{x})$ and which is given by

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt \tag{1.11}$$

We obtain the following decomposition of the expected squared loss

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise} \tag{1.12}$$

where

$$(\text{bias})^2 = \int\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})d\mathbf{x} \tag{1.13a}$$

$$\text{variance} = \int\mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2]p(\mathbf{x})d\mathbf{x} \tag{1.13b}$$

$$\text{noise} = \int\{h(\mathbf{x}) - t\}p(\mathbf{x},t)d\mathbf{x}dt \tag{1.13c}$$

1.3 Bayesian