1 Intorduction Problem Setting

The introduction use polynomial curve fitting as the example.

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

we can use the general defined loss function to describe the fitting performance and the normilized RMS loss function(in order to make compare different dataset size).

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

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$

We can use the penalty term to control the model complexity in order to match the problem complexity. (As mentioned in class, the DL model can be overparameterized model, which does not follow the traditional model complexity control theory)

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

2 Probability Theory

This section describe the general probability defintion in descrete conditoin and continuous condition. With Bayes' theorem and the distribution, we can describe the learning processure in probability perspective. [1]

3 The curve fitting revisted in probabilistic perspective

3.1 Piror

Given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$

$$p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}\left(t \mid y(x, \mathbf{w}), \beta^{-1}\right)$$

, where precision parameter β corresponding to the inverse variance of the distribution.

[1] limited to the page limit, you can get the full version note at https://github.com/Xiang-Pan/NYU_Baysian_Machine_Learning

If the data point is i.i.d.,

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid y\left(x_n, \mathbf{w}\right), \beta^{-1}\right),$$

thus we can have the log likelihood,

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

.

We can remove the term that do not depend on w, and scale the whole equation, then we can minimize the negative log likelihood,

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

.

For inference, we can have,

$$p\left(t \mid x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}\right) = \mathcal{N}\left(t \mid y\left(x, \mathbf{w}_{\mathrm{ML}}\right), \beta_{\mathrm{ML}}^{-1}\right)$$

4 Model Selection

cross-validation drawbacks:

- training is itself computationally expensive
- multiple complexity parameters for a single model

5 The Curse of Dimensionality

We have some observation and assumption for the high-dimensional data, which helps us overcome the curse of dimensionality.

6 Decision Theory

Bayes' theorem,

$$p\left(\mathcal{C}_k \mid \mathbf{x}\right) = \frac{p\left(\mathbf{x} \mid \mathcal{C}_k\right) p\left(\mathcal{C}_k\right)}{p(\mathbf{x})}$$

Minimizing the misclassification rate,

$$\begin{aligned} p(\text{ mistake }) &= p\left(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2\right) + p\left(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\ &= \int_{\mathcal{R}_1} p\left(\mathbf{x}, \mathcal{C}_2\right) \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p\left(\mathbf{x}, \mathcal{C}_1\right) \mathrm{d}\mathbf{x} \end{aligned}$$

We can maximize the probability of being correct

$$\begin{aligned} p(\text{ correct }) &= \sum_{k=1}^{K} p\left(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k\right) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p\left(\mathbf{x}, \mathcal{C}_k\right) \mathrm{d}\mathbf{x} \end{aligned}$$

(PS, these two inference methods can be the same at the inference time. However, they MAY be different on the training procedure)

7 Inference and decision

Inference stage in which we use training data to learn a model for $p(C_k|x)$.

Decision stage in which we use these posterior probabilities to make optimal class assignments.

Three ways,

- Bayes' theorem
- Directly assign class for each x
- Discriminant function

8 Information Theory

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x} \mid \mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y} \mid \mathbf{x}]$$

Thus we can view the mutual information as the reduction in the uncertainty about x by virtue of being told the value of y (or vice versa). From a Bayesian perspective, we can view p(x) as the prior distribution for x and p(x| y) as the posterior distribution after we have observed new data y. The mutual information therefore represents the reduction in uncertainty about x as a consequence of the new observation y.