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

2.1 Discrete Conditoin

$$\text{sum rule} \quad p(X, Y) = p(Y \mid X) p(X)$$

$$\text{product rule} \quad p(X) = \sum_{Y} p(X, Y)$$

2.2 Continuous Conditoin

In continus condition, we can use probability densities to define the probability for the variable x with continuous values.

$$p(x \in (a, b)) = \int_a^b p(x) \mathrm{d}x$$

2.3 Expection and Covariances

$$\mathbb{E}[f] = \sum_{x} p(x) f(x)$$

$$\begin{aligned}\operatorname{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

2.4 Bayesian Probability

Bayes theorem

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\times$ prior

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

2.5 Gaussian Distribution

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\operatorname{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 = \sigma^2$$

[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/blob/master/reading_notes/ch1/note1.pdf

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,

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

3.2 Prediction Distribution

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid x, \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{w}$$

4 Model Selection

cross-validation drawbacks:

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

We need measure of performance which depends only on the training data and which does not suffer from bias due to over-fitting.

$$\ln p\left(\mathcal{D} \mid \mathbf{w}_{\mathrm{ML}}\right) - M \quad \text{(M is the number of adjustable parameters)}$$

Such criteria do not take account of the uncertainty in the model parameters, however, and in practice they tend to favour overly simple models.

5 The Curse of Dimensionality

- Real data will often be confined to a region of the space having lower effective dimensionality, and in particular the directions over which important variations in the target variables occur may be so confined.

- Real data will typically exhibit some smoothness properties (at least locally) so that for the most part small changes in the input variables will produce small changes in the target variables, and so we can exploit local interpolation-like techniques to allow us to make predictions of the target variables for new values of the input variables.

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.