Reading Notes for ch3 Linear Models for Regression

Xiang Pan

September 11, 2021

1 Linear Basis Function Models

This section just reviews the basic linear regression setting.

2 Bias-Variance Decomposition

$$
\begin{aligned}
\text{expected loss } &= (\text{ bias })^2+\text{ variance }+\text{ noise} \\
(\text{ bias })^2 &= \int\left\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x} ; \mathcal{D})]-h(\mathbf{x})\right\}^2 p(\mathbf{x}) \mathrm{d} \mathbf{x} \\
\text{variance } &= \int \mathbb{E}_{\mathcal{D}}\left[\left\{y(\mathbf{x} ; \mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\mathbf{x} ; \mathcal{D})]\right\}^2\right] p(\mathbf{x}) \mathrm{d} \mathbf{x} \\
\text{noise } &= \int\{h(\mathbf{x})-t\}^2 p(\mathbf{x}, t) \mathrm{d} \mathbf{x} \mathrm{d} t
\end{aligned}
$$

Very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance.

3 Bayesian Liear Regression

3.1 Error Decoposition

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

$$
p(t \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)=\mathcal{N}\left(t \mid \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right)
$$

$$
\sigma_N^2(\mathbf{x}) = \text{noise on the data} + \text{uncertainty associated with the parameters w.}
$$

$$
\sigma_N^2(\mathbf{x})=\frac{1}{\beta}+\phi(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \phi(\mathbf{x})
$$

When dataset size get unlimited, the second term goes to zero.

3.2 Equivalent kernel(smoother matrix)

For predictive mean,

$$
y\left(\mathbf{x}, \mathbf{m}_N\right)=\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})=\beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}=\sum_{n=1}^N \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}\left(\mathbf{x}_n\right) t_n
$$

Forming a weighted combination of the target values in which data points close to x are given higher weight than points further removed from x.

$$
y\left(\mathbf{x}, \mathbf{m}_N\right)=\sum_{n=1}^N k\left(\mathbf{x}, \mathbf{x}_n\right) t_n
$$

$$
\begin{aligned}
\operatorname{cov}\left[y(\mathbf{x}), y\left(\mathbf{x}^{\prime}\right)\right] &= \operatorname{cov}\left[\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{w}, \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}\left(\mathbf{x}^{\prime}\right)\right] \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}\left(\mathbf{x}^{\prime}\right)=\beta^{-1} k\left(\mathbf{x}, \mathbf{x}^{\prime}\right)
\end{aligned}
$$

4 Bayesian Model Comparison

4.1 posterior distribution

$$\text{(model posterior distribution)} \propto \text{(model prior probability distribution)} * \text{(model evidence)}$$

$$p\left(\mathcal{M}_i \mid \mathcal{D}\right) \propto p\left(\mathcal{M}_i\right) p\left(\mathcal{D} \mid \mathcal{M}_i\right)$$

prior probability distribution $p(M_i)$: allows us to express a preference for different models
model evidence $p(D|M_i)$: preference shown by the data for different models
Bayes factor $p(D|M_i)/p(D|M_j)$: the ratio of model evidences for two models.

4.2 predictive distribution

$$p(t \mid \mathbf{x}, \mathcal{D})=\sum_{i=1}^{L} p\left(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D}\right) p\left(\mathcal{M}_i \mid \mathcal{D}\right)$$

4.3 model evidence

We can obtain a rough approximation to the model evidence if we assume that the posterior distribution over parameters is sharply peaked around its mode w_{MAP}.

$$\ln p(\mathcal{D}) \simeq \ln p\left(\mathcal{D} \mid w_{\mathrm{MAP}}\right)+\ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}\right)$$

5 Evidence Approximation

We set the hyperparameters to specific values determined by maximizing the marginal likelihood function obtained by first integrating over the parameters w.

$$p(t \mid \mathbf{t}) \simeq p(t \mid \mathbf{t}, \widehat{\alpha}, \widehat{\beta})=\int p(t \mid \mathbf{w}, \widehat{\beta}) p(\mathbf{w} \mid \mathbf{t}, \widehat{\alpha}, \widehat{\beta}) \mathrm{d} \mathbf{w}$$

update parameter iteraly.