Chapter 14. Combining Models

14.1. Bayesian Model Averaging

(pp. 654 - 655) Difference between ***Model Combination*** *methods* and ***Bayesian Model Averaging***

- In Model Combination, different data points withn the full set can potentially be generated from different values of the latent variable $\mathbf{z}$ and hence by different components/models

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N} [\sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n)]$$

  i.e., each individual data $\mathbf{x}_n$ has a corresponding latent variable $\mathbf{z}_n$
- In Bayesian Model Averaging,

$$p(\mathbf{X}) = \sum_{h=1}^{H} p(\mathbf{X}|h)p(h)$$

  Each model indexed by h contributes to the entire dataset $\mathbf{X}$ with prior probability $p(h)$. Of course as the size of the data set increases, this uncertainty reduces as the posterior probabilities $p(h|\mathbf{X})$ become more focussed on one of the models.

14.2. Committees

(pp. 656) Bootstrap aggregation or bagging

$$y_{COM}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} y_m(\mathbf{x})$$

Consider regression models with $h(\mathbf{x})$ being the true function, and the average SSE by each individual model is

$$E_{AV} = \frac{1}{M} \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]$$

The SSE of the committee model is given by

$$E_{COM} = \mathbb{E}_{\mathbf{x}}[\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\}^2]$$

Notice that $E_{COM} = \frac{1}{M}E_{AV}$ only when $\mathbb{E}_{\mathbf{x}}[\epsilon_m] = 0$ and $\mathbb{E}_{\mathbf{x}}[\epsilon_m \cdot \epsilon_l] \neq 0$, which is often not the case in practice. But we have $E_{COM} \leq E_{AV}$ regardless.

14.3. Boosting (AdaBoost)

- See previous notes for the AdaBoost algorithm step by step.
- (pp. 659) exponential error function

  $$E = \sum_{n=1}^{N}\exp\{-t_n f_m(\mathbf{x}_n)\}$$

  can be re-write by separating off the contribution from $y_m(\mathbf{x})$ as

  $$E = \sum_{n=1}^{N}\omega_n^{(m)}\exp\{-\frac{1}{2}\alpha_m y_m(\mathbf{x}_n)\}$$

  Notice how this sequential minimization leads to the simple AdaBoost scheme.
- (pp. 662) AdaBoost seeks the best approximation to the log odds ratio within the space of functions represente by the linear combination of base classfiers, subject to the contrained minimization resulting from the sequential optimization strategy.
- **Drawbacks**: less robust to outliers or misclassifeid data points compared with cross-entropy objective, cannot be interpreted as the log likelihood function of any well-defined probabilistic model.

14.4. Tree-based Models

14.5. Conditional Mixture Models