19.6 BAYESIAN METHODS FOR LEARNING BELIEF NETWORKS

Part V made the case for the importance of probabilistic representations of uncertain knowledge, and presented belief networks as a general and useful performance element based on probability theory. In this section, we discuss the general problem of learning probabilistic knowledge, and the specific problem of learning belief networks. We will see that a Bayesian view of learning is extremely powerful, providing general solutions to the problems of noise, overfitting, and optimal prediction. We will also find striking parallels between belief networks and neural networks in their amenability to local, gradient-descent learning methods. Most of this section is fairly mathematical, although the general lessons can be understood without plunging into the details. It may be helpful at this point to review the material in Chapters 14 and 15.

Bayesian learning

BAYESIAN LEARNING

Bayesian learning views the problem of constructing hypotheses from data as a subproblem of the more fundamental problem of making predictions. The idea is to use hypotheses as intermediaries between data and predictions. First, the probability of each hypothesis is estimated, given the data. Predictions are then made from the hypotheses, using the posterior probabilities of the hypotheses to weight the predictions. As a simple example, consider the problem of predicting tomorrow's weather. Suppose the available experts are divided into two camps: some propose model A, and some propose model B. The Bayesian method, rather than choosing between A and B, gives some weight to each based on their likelihood. The likelihood will depend on how much the known data support each of the two models.

Suppose that we have data D and hypotheses $H_1, H_2, \ldots$, and that we are interested in making a prediction concerning an unknown quantity X. Furthermore, suppose that each H_i specifies a complete distribution for X. Then we have

$$\mathbf{P}(X|D) = \sum_i \mathbf{P}(X|D, H_i)\mathbf{P}(H_i|D) = \sum_i \mathbf{P}(X|H_i)\mathbf{P}(H_i|D)$$

This equation describes full Bayesian learning, and may require a calculation of $\mathbf{P}(H_i|D)$ for all H_i. In most cases, this is intractable; it can be shown, however, that there is no better way to make predictions.

MAXIMUM A POSTERIORI

The most common approximation is to use a *most probable* hypothesis, that is, an H_i that maximizes $\mathbf{P}(H_i|D)$. This often called a **maximum a posteriori** or MAP hypothesis H_{MAP}:

$$\mathbf{P}(X|D) \approx \mathbf{P}(X|H_{\mathrm{MAP}})$$

The problem is now to find H_{MAP}. By applying Bayes' rule, we can rewrite $\mathbf{P}(H_i|D)$ as follows:

$$\mathbf{P}(H_i|D) = \frac{\mathbf{P}(D|H_i)\mathbf{P}(H_i)}{\mathbf{P}(D)}$$

Notice that in comparing hypotheses, $P(D)$ remains fixed. Hence, to find H_{MAP}, we need only maximize the numerator of the fraction.

The first term, $\mathbf{P}(D|H_i)$, represents the probability that this particular data set would have been observed, given H_i as the underlying model of the world. The second term represents the