1999

p8q9

JGD

**Model Answer, Neural Computing, Question 2**

Neural computation deals with problems involving real-world data and must therefore address the issue of *uncertainty.* The uncertainty arises from a variety of sources including noise on the data, mislabelled data, the natural variability of data sources, and overlapping class distributions. Probability theory provides a consistent framework for the quantification of uncertainty, and is unique under a rather general set of axioms.

The goal in classifiction is to predict the class $\mathcal{C}_k$ of an input pattern, having observed a vector $\mathbf{x}$ of features extracted from that pattern. This can be achieved by estimating the conditional probabilities of each class given the input vector, i.e. $P(\mathcal{C}_k|\mathbf{x})$. The optimal decision rule, in the sense of minimising the average number of mis-classifications, is obtained by assigning each new $\mathbf{x}$ to the class having the largest posterior probability.

The likelihood function, for a particular probabilistic model and a particular observed data set, is defined as the probability of the data set given the model, viewed as a function of the adjustable parameters of the model. Maximum likelihood estimates the parameters to be those values for which the likelihood function is maximized. It therefore gives the parameter values for which the observed data set is the most probable.

Since the data points are assumed to be independent, the likelihood function is given by the product of the densities evaluated for each data point

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}) \\ &= \prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\{t_n - f(\mathbf{x}_n, \mathbf{w})\}^2}{2\sigma^2}\right).\end{aligned}$$

Following the standard convention, we can define an error function by the negative logarithm of the likelihood

$$E(\mathbf{w}) = -\ln \mathcal{L}(\mathbf{w}).$$

Since the negative logarithm is a monotonically decreasing function, maximization of $\mathcal{L}(\mathbf{w})$ is equivalent to minimization of $E(\mathbf{w})$. Hence we obtain

$$E(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}\{t_n - f(\mathbf{x}_n, \mathbf{w})\}^2 + \frac{N}{2}\ln(2\pi\sigma^2)$$

which, up to an additive constant independent of $\mathbf{w}$ and a multiplicative constant also independent of $\mathbf{w}$, is the sum-of-squares error function.