2 classes) or softmax ($K \geqslant 2$ classes) activation functions. These are particular cases of a more general result obtained by assuming that the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ are members of the subset of the exponential family of distributions given by

Section 3.4

$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s} h\left(\frac{1}{s}\mathbf{x}\right) g(\boldsymbol{\lambda}_k) \exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}. \tag{5.66}$$

Here the scaling parameter s is shared across all the classes.

For the two-class problem, we substitute this expression for the class-conditional densities into (5.41) and we see that the posterior class probability is again given by a logistic sigmoid acting on a linear function $a(\mathbf{x})$, which is given by

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2). \tag{5.67}$$

Similarly, for the K-class problem, we substitute the class-conditional density expression into (5.46) to give

$$a_k(\mathbf{x}) = \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_k) + \ln p(\mathcal{C}_k) \tag{5.68}$$

and so again is a linear function of $\mathbf{x}$.

5.4. Discriminative Classifiers

For the two-class classification problem, we have seen that the posterior probability of class $\mathcal{C}_1$ can be written as a logistic sigmoid acting on a linear function of $\mathbf{x}$, for a wide choice of class-conditional distributions $p(\mathbf{x}|\mathcal{C}_k)$ from the exponential family. Similarly, for the multi-class case, the posterior probability of class $\mathcal{C}_k$ is given by a softmax transformation of linear functions of $\mathbf{x}$. For specific choices of the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$, we have used maximum likelihood to determine the parameters of the densities as well as the class priors $p(\mathcal{C}_k)$ and then used Bayes' theorem to find the posterior class probabilities. This represents an example of *generative* modelling, because we could take such a model and generate synthetic data by drawing values of $\mathbf{x}$ from the marginal distribution $p(\mathbf{x})$ or from any of the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$.

However, an alternative approach is to use the functional form of the generalized linear model explicitly and to determine its parameters directly by using maximum likelihood. In this direct approach, we maximize a likelihood function defined through the conditional distribution $p(\mathcal{C}_k|\mathbf{x})$, which represents a form of *discriminative* probabilistic modelling. One advantage of the discriminative approach is that there will typically be fewer learnable parameters to be determined, as we will see shortly. It may also lead to improved predictive performance, particularly when the assumed forms for the class-conditional densities represent a poor approximation to the true distributions.