computation if the data set is large. This effect can be offset, at the expense of some additional one-off computation, by constructing tree-based search structures to allow (approximate) near neighbours to be found efficiently without doing an exhaustive search of the data set. Nevertheless, these nonparametric methods are still severely limited. On the other hand, we have seen that simple parametric models are very restricted in terms of the forms of distribution that they can represent. We therefore need to find density models that are very flexible and yet for which the complexity of the models can be controlled independently of the size of the training set, and this can be achieved using deep neural networks.

Exercises

3.1 ($\star$) Verify that the Bernoulli distribution (3.2) satisfies the following properties:

$$\sum_{x=0}^{1} p(x|\mu) = 1 \tag{3.191}$$

$$\mathbb{E}[x] = \mu \tag{3.192}$$

$$\mathrm{var}[x] = \mu(1-\mu). \tag{3.193}$$

Show that the entropy $\mathrm{H}[x]$ of a Bernoulli-distributed random binary variable x is given by

$$\mathrm{H}[x] = -\mu \ln \mu - (1-\mu)\ln(1-\mu). \tag{3.194}$$

3.2 ($\star\star$) The form of the Bernoulli distribution given by (3.2) is not symmetric between the two values of x. In some situations, it will be more convenient to use an equivalent formulation for which $x \in \{-1, 1\}$, in which case the distribution can be written

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2} \tag{3.195}$$

where $\mu \in [-1, 1]$. Show that the distribution (3.195) is normalized, and evaluate its mean, variance, and entropy.

3.3 ($\star\star$) In this exercise, we prove that the binomial distribution (3.9) is normalized. First, use the definition (3.10) of the number of combinations of m identical objects chosen from a total of N to show that

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}. \tag{3.196}$$

Use this result to prove by induction the following result:

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m, \tag{3.197}$$