Lecture Outline

Reading: Chapter 2

Section 2.5.1 will be left to the student to read.

This lecture covers

- Nonparametric methods
- Histogram method
- K-Nearest Neighbors
- Nonparametric classification

Density Estimation: Parametric vs. Non-Parametric

Parametric approach:

- Density must be chosen (good model or not?)
- (Usually a) Small number of parameters to be estimated and thus efficient in terms of storage and computation e.g. find $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ for $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or λ for $\text{Bern}(x|\lambda)$ via ML or MAP.
- Once parameters are estimated, dataset can be discarded

Non-Parametric approach:

- Few assumptions about the form of the data
- Require dataset to be stored (not histograms)
- Curse of dimensionality implies lots of data required in order to get a good estimation of the pdf

2.5 Nonparametric Methods

We have focussed on the use of probability distributions having specific functional forms governed by a small number of parameters whose values are to be determined from a data set. This is called the *parametric* approach to density modeling. An important limitation of this approach is that the chosen density might be a poor model of the distribution that generates the data, which will result in poor predictive performance. For instance, if the process that generates the data is multimodal, then this aspect of the distribution can never be captured by a Gaussian, which is necessarily unimodal.

We consider some *nonparametric* approaches to density estimation that make few assumptions about the form of the distribution.

Histogram Method

Standard histograms simply partition x into distinct bins of width Δ_i and then count the number n_i of observations of x falling in bin i. In order to turn this count into a normalized pdf, we simply divide by the

total number N of observations and by the width Δ_i of the bins to obtain probability values for each bin given by

$$p_i = \frac{n_i}{N\Delta_i} \tag{1}$$

for which it is easily seen that $\int p(x)\mathrm{d}x = 1$. This gives a model for the density $p(x)$ that is constant over the width of each bin, and often the bins are chosen to have the same width $\Delta_i = \Delta$.

In Figure 1, we show an example of histogram density estimation. We see that when Δ is very small (top figure), the resulting density model is very spiky, with a lot of structure that is not present in the underlying distribution that generated the dataset. Conversely, if Δ is too large (bottom figure) then the result is a model that is too smooth and the consequently fails to capture the bimodal property of the green curve. The best results are obtained for some intermediate value of Δ (middle figure).

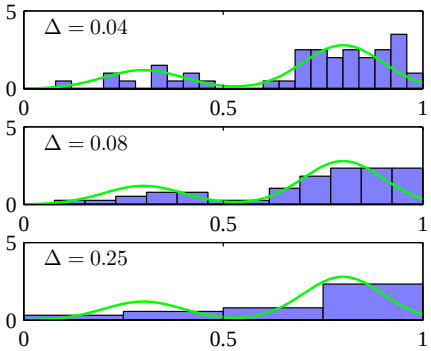

Figure 1: Textbook figure 2.24

Once the histogram has been computed, the dataset itself can be discarded, which can be advantageous if the dataset is large (some PRML methods require us to keep all the data). In practice, the histogram technique can be useful for obtaining a quick visualization of 1D or 2D data but is unsuited to most density estimation applications mainly because it does not scale well with dimensionality. If we divide each variable (element in data vector) in a D-dimensional space into M bins, the total number of bins will be M^D. This exponential scaling with D is an example of the curse of dimensionality.

2.5.2 K-Nearest Neighbors Method

Suppose we have a data set of N observations. We consider a small sphere centered on the data point $\mathbf{x}$ at which we wish to estimate the density $p(\mathbf{x})$, and we allow the radius of the sphere to grow until it contains precisely K data points.

Figure 2: K-Nearest Neighbors Method

The estimate of the density $p(\mathbf{x})$ is then given by

$$p(\mathbf{x}) \simeq \frac{K}{NV}. \tag{2}$$

with V set to the volume of the resulting sphere. This technique is known as K nearest neighbors.

Note: The model produced by K nearest neighbors is not a true density model because $\int p(\mathbf{x})d\mathbf{x} \to \infty$.

Fig. 3 shows density estimate for various choices of K. The parameter K governs the degree of smoothing: for $K = 1$ the density estimator is "spiky" because V can be small whereas for $K = 30$ the estimate is too smooth and the bimodal nature of the true distribution is not properly modeled.

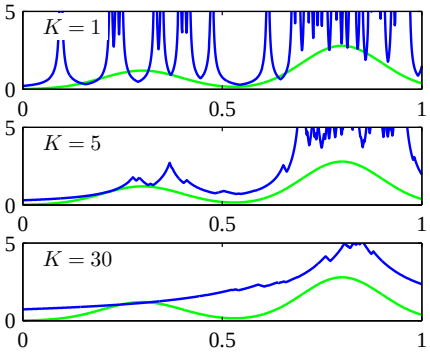

Figure 3: Textbook Figure 2.26

Application of K-Nearest Neighbors to Classification

Let us suppose that we have a data set comprising N_k points in class $\mathcal{C}_k$ with N total data points, i.e. $\sum_k N_k = N$. If we wish to classify a new point $\mathbf{x}$, we draw a sphere centered on $\mathbf{x}$ containing precisely K points irrespective of their class. Suppose this sphere has a volume V and contains K_k points from class $\mathcal{C}_k$. Then (2) provides an estimate of the density associated with each class, i.e. conditional pdf

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V} \tag{3}$$

and the class priors are given by

$$p(C_k) = \frac{N_k}{N}. \tag{4}$$

We use Bayes' theorem to obtain the posterior probability of class membership

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \\ &= \frac{\frac{K_k}{N_k V} \cdot \frac{N_k}{N}}{\frac{K}{NV}} \\ &= \frac{K_k}{K}. \end{aligned} \tag{5}$$

If we wish to minimize the probability of misclassification, this is done by assigning the test point $\mathbf{x}$ to the class having the largest posterior probability. Thus to classify a new point, we identify the K nearest points

(neighbors) from the training set and then assign the new point to the class having the largest number of representatives amongst this set. The particular case of $K = 1$ is called the *nearest neighbor* rule, because a test point is simply assigned to the same class as the nearest point from the training set. See Figs. 4 and 5. This classifier requires us to keep the training dataset.

We have seen that parametric models are restricted in terms of the forms of distributions that they can represent. Also, nonparametric methods are also severely limited. 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 we shall see in subsequent chapters how to do this.

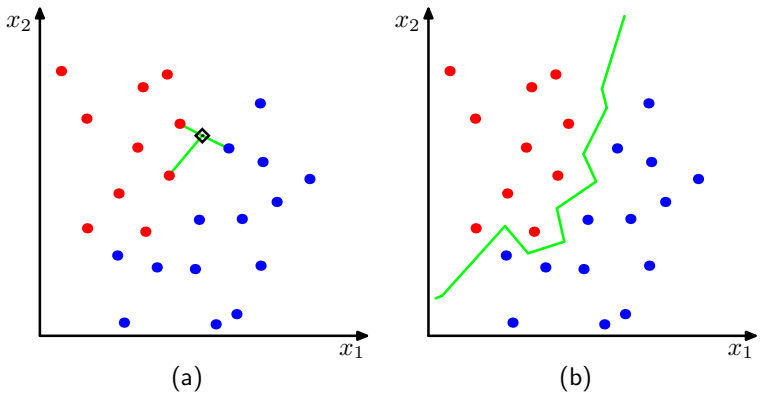

Figure 4:

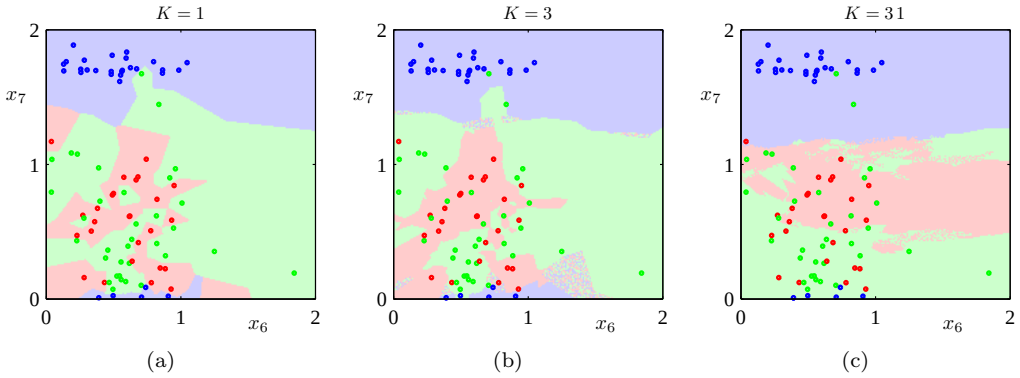

Figure 5: