INM431 Machine Learning

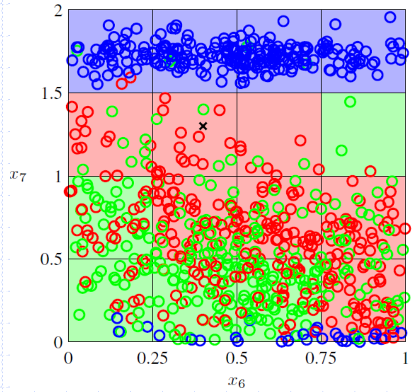
Artur S. d'Avila Garcez
a.garcez@city.ac.uk
http://www.staff.city.ac.uk/~aag/

Content

- The Curse of Dimensionality
- Kernel Density Estimation
- K-nearest neighbours

The Curse of Dimensionality

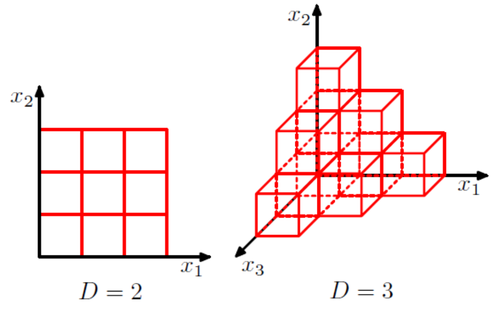
A simple approach:



The test point is predicted as being in the class having the largest number of training points in the cell (with ties broken at random)

The curse…

The number of cells grows exponentially with the number of dimensions (i.e. variables)



Requires exponentially large training data (big data?) to ensure that cells are not empty

The antidotes…

Data often confined to region of space with lower effective dimensionality (dimensionality reduction)

Smoothness: normally, small changes in the input produce small changes in the target variable (thus allowing prediction)

Big data? Quality data with labels still difficult to get… semi-supervised learning!

Nonparametric Methods (1)

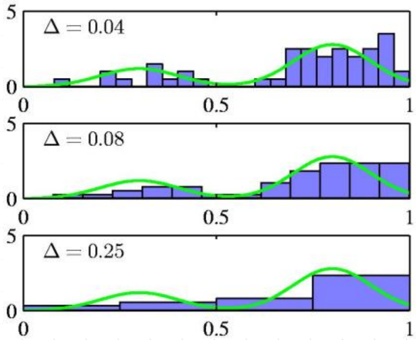
- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

Nonparametric Methods (2)

Histogram methods partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin

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

Often, the same width is used for all bins, $\Delta_i = \Delta$



In a D-dimensional space, using M bins in each dimension will require M^D bins!

Nonparametric Methods (3)

Assume observations drawn from a density p(x) and consider a small region R containing x such that:

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

The probability that K out of N observations lie inside R follows a binomial distribution. For large N:

$$K \simeq NP.$$

If the volume V of R is sufficiently small, p(x) is approximately constant over R and:

$$P \simeq p(\mathbf{x})V$$

Thus:

$$\boxed{p(\mathbf{x}) = \frac{K}{NV}} \quad \text{(Eq.1)}$$

Nonparametric Methods (4)

Kernel Density Estimation: fix V, estimate K from the data.

Let R be a hypercube of side h centred on x and define the kernel function (Parzen window):

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

i.e. k = 1 iff x_n is inside the cube (for each dimension)

It follows that the total number K of points inside the cube… substituting on Eq.1, one gets p(x)

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \qquad\qquad p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$

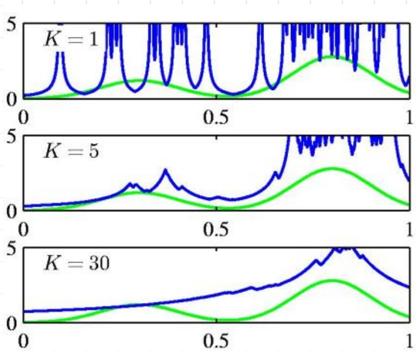
Nonparametric Methods (5)

Nearest Neighbour Density Estimation: Alternatively, fix K, and estimate V from the data.

Consider a hypersphere centred on x and let it grow to a volume V* that includes K of the given N data points.

Then:

$$p(\mathbf{x}) \simeq \frac{K}{NV^{\star}}.$$



K acts as a smoother.

K-Nearest-Neighbours for Classification (1)

Given a data set with N_k data points from class C_k and $\sum_k N_k = N$, we have

$$p(\mathbf{x}) = \frac{K}{NV}$$

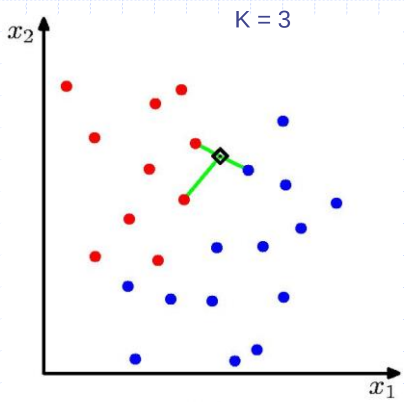
and correspondingly

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

Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

K-Nearest-Neighbours for Classification (2)



Extension: use 1-NN classifier on centroids obtained by K-means to classify new data into clusters