Computational Statistics
Unsupervised Learning

Luca Bortolussi

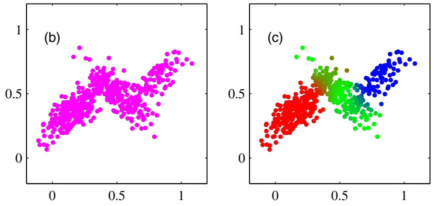
Department of Mathematics and Geosciences
University of Trieste

Office 238, third floor, H2bis
`luca@dmi.units.it`

Trieste, Winter Semester 2015/2016

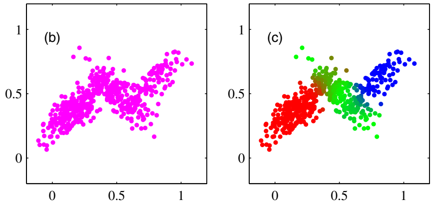
UNSUPERVISED LEARNING - OVERVIEW

Unsupervised learning: No labels are given to the learning algorithm (input only), leaving it on its own to find structure in its input.



Unsupervised Learning - overview

Unsupervised learning: No labels are given to the learning algorithm (input only), leaving it on its own to find structure in its input.



- Clustering: discover groups of similar examples within the data.
- Density estimation: determine the distribution of data within the input space.
- Dimensionality reduction: project the data from a high-dimensional space to a lower dimension space. Often down to two or three dimensions for the purpose of visualization.

Outline

1. Density Estimation
2. Clustering
3. Expectation Maximisation
4. Dimensionality Reduction

Density estimation

$X_i \in \mathbb{R}^d$ $p(x)$?

Given input data $\mathbf{x_1}, \ldots, \mathbf{x_N}$, sampled by an unknown distribution $p(X)$, estimate p.

- One way to solve this problem is to fix a parametric family of distributions $p(X|\theta)$ and then estimate parameters θ according to ML, MAP, or with a fully Bayesian treatment. The drawback is that a bad choice of the family of distributions can result in a poor fit of data.
- Non-parametric methods try to construct an estimate from data only, avoiding the pitfalls involved in choosing the correct family of models.

$\max_\theta \log \prod_i p(x_i|\theta)$

$\max_\theta \log \prod_i p(x_i|\theta) + \log p(\theta)$

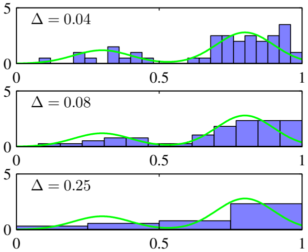
HISTOGRAM DENSITY

- (1D) Partition input space in bins (intervals) $B_1, \dots, B_k$,s each of size Δ_i, and count how many input points n_j fall inside each bin j. Define the density $p(x)$ as p_i if $x \in B_i$, where

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

- The resulting density is discontinuous, and the quality of the fit depends on the bin size.
- Curse of dimensionality: the number of bins grows exponentially with the dimension d of $\mathbf{x}$.

An illustration of the histogram approach to density estimation, in which a data set of 50 data points is generated from the distribution shown by the green curve. Histogram density estimates, based on (2.241), with a common bin width Δ are shown for various values of Δ.



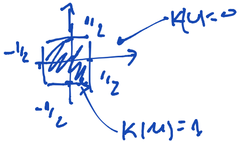
DATA-BASED ESTIMATOR

- Histogram estimation at a point x uses information only from few data points close to x, those lying in the same bin. But bins are rigid and result in discontinuous densities.
- We can do better “placing a (hard/ soft) box” in each point x.
- Consider now a little box B containing point $\mathbf{x}$, with volume V, and let P be the probability that a sampled point is in B, i.e. $P = \int_B p(\mathbf{x}) d\mathbf{x}$. The probability P can be estimated as $P = K/N$, for sufficiently large K and N (law of large numbers for Binomial), where K is the number of points falling into B. Furthermore, if B is sufficiently small, we can approximate P as $p(\mathbf{x})V$. It then follows that

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

for $\mathbf{x} \in B$.
- We can now either fix K and estimate V from data (K-nearest-neighbour) or fix V and estimate K from data (kernel-based or Parzen estimators)

PARZEN ESTIMATOR



- Consider the function (Parzen window)

$$k(\mathbf{u}) = \begin{cases} 1, & \|\mathbf{u}\|_\infty \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

- Then a data point $\mathbf{x}_n$ is inside the cube of edge length h centred in $\mathbf{x}$ if and only if

$$k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) = 1,$$

so that the number of data points in the cube is

$$K = \sum_n k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

- Then the estimate for the density p (in d dimensions) becomes:

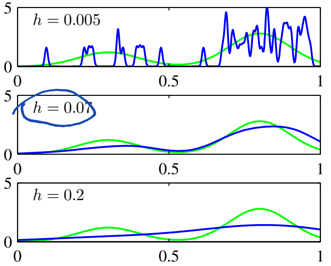
$$p(\mathbf{x}) = \frac{1}{Nh^d} \sum_n k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right).$$

Parzen Estimator

- The Parzen window is still discontinuous. An alternative approach is to use a smooth function, i.e. a kernel satisfying $k(\mathbf{x}) \geq 0$ and $\int k(\mathbf{x})d\mathbf{x} = 1$.
- a common choice is the Gaussian kernel, giving the estimate:

$$p(\mathbf{x}) = \frac{1}{N}\sum_n \frac{1}{(2\pi h^2)^{1/2}} \exp\left(\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{h^2}\right)$$

Illustration of the kernel density model (2.250) applied to the same data set used to demonstrate the histogram approach in Figure 2.24. We see that h acts as a smoothing parameter and that if it is set too small (top panel), the result is a very noisy density model, whereas if it is set too large (bottom panel), then the bimodal nature of the underlying distribution from which the data is generated (shown by the green curve) is washed out. The best density model is obtained for some intermediate value of h (middle panel).



K-NEAREST NEIGHBOUR ESTIMATOR

- It may be more convenient to have h depending on the local density of observations, to avoid over or under-smoothing.
- K-nearest neighbour solves this problem by setting the radius of the sphere/ box for Parzen estimation such that it exactly contains K points, i.e. equal to the distance of the K-th closest point to $\mathbf{x}$. Then $p(\mathbf{x})$ is estimated as $K/V(\mathbf{x})N$, where $V(\mathbf{x})$ is the volume of the sphere/box.
- K-NN can be used also for classification, by assigning to class C_k class-conditional probability in $\mathbf{x}$ equal to K_k/K, where K_k is the number of points of class K.

Illustration of K-nearest-neighbour density estimation using the same data set as in Figures 2.25 and 2.24. We see that the parameter K governs the degree of smoothing, so that a small value of K leads to a very noisy density model (top panel), whereas a large value (bottom panel) smoothes out the bimodal nature of the true distribution (shown by the green curve) from which the data set was generated.