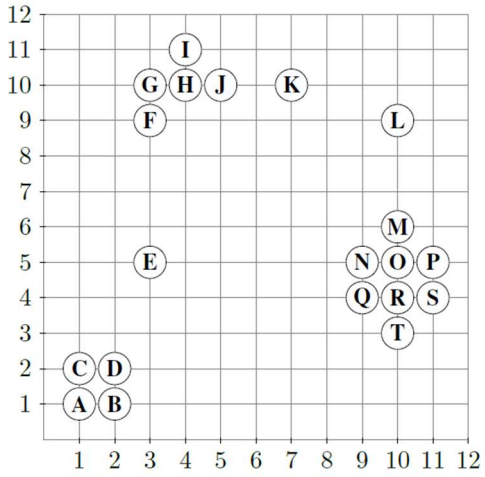
University of Southern Denmark
IMADA
Ricardo Campello

DM583: Data Mining

Exercise 5: Density Estimates and Probabilistic Clustering

Exercise 5-1 Kernel Density Estimates

a) Consider the following dataset in 2 dimensions:

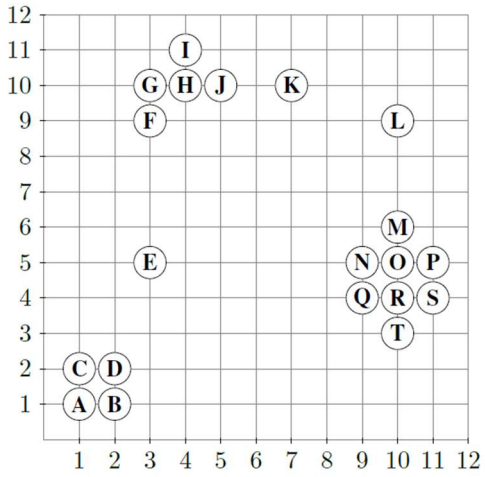


Estimate the density at the locations coinciding with data points H, K, E, as well as at location (3,7), using the **Discrete Kernel** (so-called square/cubic/hypercubic kernel) with window width equal to: (i) $h = 2$; (ii) $h = 4$

b) Now, consider the unidimensional dataset corresponding only to the horizontal coordinate (x-axis) in the figure in item (a). Estimate the density at location 6 using this 1D dataset and a **Gaussian Kernel** with width $h = 1$

Exercise 5-2 KNN Density Estimates

a) Consider the following dataset in 2 dimensions:



Estimate the density at the locations coinciding with data points H, K and E, as well as at location (3,7), using the **K-Nearest Neighbour (KNN) Estimator** with *Euclidean Distance* and $k = 2$. **Note 1** (density at data points): when we are estimating density at a location that coincides with one or more data points, these should be counted as part of the KNN neighbourhood of the location of interest. **Note 2** (ties): when there is more than one data point tied as kth nearest neighbour at the same distance from the query location, these should all be counted as part of the KNN neighbourhood of the location of interest, which in practice corresponds to adjusting the value of k used in the computation of the density estimate for that particular location.

b) Repeat item (a), but now with $k = 3$.

Exercise 5-3 EM-GMM Clustering

a) Suppose you are running EM-GMM to perform probabilistic clustering of a 2-dimensional dataset into $k = 2$ clusters and, during execution of the algorithm, at a given iteration, the estimated multivariate normal clusters, C_1 and C_2, are as follows:

$$Cov(C_1) = \begin{bmatrix} 0.125 & 0 \\ 0 & 0.125 \end{bmatrix}$$

$$Cov(C_2) = \begin{bmatrix} 0.8 & 0.27 \\ 0.27 & 0.11 \end{bmatrix}$$

$$\mathbf{v}_{C_1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\mathbf{v}_{C_2} = \begin{bmatrix} 4.5 \\ 2.5 \end{bmatrix}$$

where Cov(.) and $v_{(.)}$ stand for a cluster's estimated covariance matrix and multivariate mean (centroid), respectively. At the iteration in question, the priors $\pi_{(.)}$ of each model component are estimated as:

$$\pi_{C_1} = 0.4$$
$$\pi_{C_2} = 0.6$$

Now consider a hypothetical data point $\mathbf{x}$ that we have not observed. What is the probability that this point can be produced by each one of the model components (clusters), if we don't know its value?

b) Now, suppose we have observed a data point in our dataset as $\mathbf{x} = [2 \;\; 1.5]^T$. Assuming the EM-GMM estimates previously hypothesized in item (a), the Squared Mahalanobis distance from this observation to the clusters can be computed as $d_M(\mathbf{x}, \mathbf{v}_1)^2 = 10$ and $d_M(\mathbf{x}, \mathbf{v}_2)^2 = 9.10596$. What is the probability density of C_1 as evaluated at $\mathbf{x}$? Repeat for C_2.

c) What is the probability that the data point referred to in item (b), $\mathbf{x} = [2 \;\; 1.5]^T$, has been produced by each one of the model components (clusters), now that we have observed its value?

d) Now, consider a hypothetical dataset with 15 data points that has been clustered into $k = 3$ clusters using EM-GMM for a certain number of iterations. Suppose the estimated posterior probabilities μ_{ij} are as follows:

0.7	0.9	0.8	0.02	0.08	0.05	0.15	0.25	0.15	0.19	0.05	0.2	0.01	0.01	0.05
0.29	0.05	0.05	0.95	0.9	0.85	0.6	0.65	0.75	0.8	0.4	0.2	0.01	0.06	0.05
0.01	0.05	0.15	0.03	0.02	0.10	0.25	0.1	0.1	0.01	0.55	0.6	0.98	0.93	0.85

Answer the following questions: (i) If we want to obtain a hard (non-overlapping) partition of the dataset in a way that data objects are assigned to clusters according to their probabilities, what would be the result in this case? (ii) Interpret the 11^{th} and the 13^{th} columns of this matrix. (iii) If the algorithm were to be run for an extra iteration, what would be the updated priors associated with each cluster?