Machine Learning

Lecture. 13.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Cluster Analysis

UNIVERSITY
of
GLASGOW

- Data Segmentation

Cluster Analysis

UNIVERSITY
of
GLASGOW

- Data Segmentation
- K-Means Clustering Algorithm

Cluster Analysis

- Data Segmentation
- K-Means Clustering Algorithm
- Kernel Based K-Means Clustering Algorithm

Cluster Analysis

UNIVERSITY of GLASGOW

- Data Segmentation
- K-Means Clustering Algorithm
- Kernel Based K-Means Clustering Algorithm
- Relation with EM Algorithm

Cluster Analysis

- Data Segmentation
- K-Means Clustering Algorithm
- Kernel Based K-Means Clustering Algorithm
- Relation with EM Algorithm
- Image Segmentation Examples

Cluster Analysis

UNIVERSITY
of
GLASGOW

- What does this scatter plot tell you?

Cluster Analysis

- What does this scatter plot tell you?

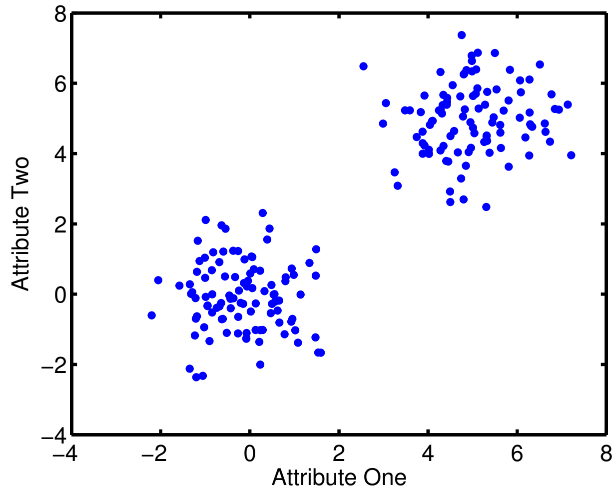

Figure 1: A sample of 200 examples of objects described by two attributes. Each dot represents a sample as defined by attribute 1 & 2, it should be obvious that there appears to be two groupings of objects which each share and internal cohesiveness and are somewhat separated from each of the other groups.

Cluster Analysis

UNIVERSITY
of
GLASGOW

- Cluster analysis aims to identify coherent structures in data

Cluster Analysis

UNIVERSITY of GLASGOW

- Cluster analysis aims to identify coherent structures in data
- How is coherence of groupings to be measured?

Cluster Analysis

- Cluster analysis aims to identify coherent structures in data
- How is coherence of groupings to be measured?
- How are coherent groupings to be identified?

Cluster Analysis

UNIVERSITY *of* GLASGOW

- Cluster analysis aims to identify coherent structures in data
- How is coherence of groupings to be measured?
- How are coherent groupings to be identified?
- Simple algorithm - K-Means clustering

Cluster Analysis

UNIVERSITY of GLASGOW

- Cluster analysis aims to identify coherent structures in data
- How is coherence of groupings to be measured?
- How are coherent groupings to be identified?
- Simple algorithm - K-Means clustering
- Direct connection with EM algorithm

K-Means Algorithm

UNIVERSITY of GLASGOW

- Data points $\mathbf{x}_n \in \mathbb{R}^D$

K-Means Algorithm

UNIVERSITY *of* GLASGOW

- Data points $\mathbf{x}_n \in \mathbb{R}^D$
- Assume at most K possible groupings or clusters

K-Means Algorithm

UNIVERSITY *of* GLASGOW

- Data points $\mathbf{x}_n \in \mathbb{R}^D$
- Assume at most K possible groupings or clusters
- Binary indicator variables associated with each data point and cluster $z_{kn} \in \{0, 1\}$

K-Means Algorithm

UNIVERSITY of GLASGOW

- Data points $\mathbf{x}_n \in \mathbb{R}^D$
- Assume at most K possible groupings or clusters
- Binary indicator variables associated with each data point and cluster $z_{kn} \in \{0, 1\}$
- Similarities with density estimation

K-Means Algorithm

UNIVERSITY of GLASGOW

- Data points $\mathbf{x}_n \in \mathbb{R}^D$
- Assume at most K possible groupings or clusters
- Binary indicator variables associated with each data point and cluster $z_{kn} \in \{0, 1\}$
- Similarities with density estimation
- Less complex as no function is required

Cluster Quality

UNIVERSITY
of
GLASGOW

- Measure of internal cohesiveness of the points allocated

Cluster Quality

UNIVERSITY of GLASGOW

- Measure of internal cohesiveness of the points allocated
- How close points are to the cluster average

Cluster Quality

- Measure of internal cohesiveness of the points allocated
- How close points are to the cluster average
- Define a measure of cluster compactness as the total distance from the cluster mean in other words

$$\sum_{\mathbf{x}_n \in \mathcal{C}_k} ||\mathbf{x}_n - \mathbf{m}_k||^2 = \sum_{n=1}^{N} z_{kn} ||\mathbf{x}_n - \mathbf{m}_k||^2$$

where the cluster mean is defined as

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x}_n \in \mathcal{C}_k} \mathbf{x}_n$$

and $N_k = \sum_{n=1}^{N} z_{kn}$ is the total number of points allocated to cluster K

Cluster Quality

UNIVERSITY *of* GLASGOW

- The total goodness of the clustering will then be based on the sum of the cluster compactness measures for each of the K clusters. Using the indicator variables z_{kn} then we can define the overall cluster goodness as

$$\mathcal{E}_K = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{kn} ||\mathbf{x}_n - \mathbf{m}_k||^2$$

  So we have our overall measure of cluster quality the next step is to devise an algorithm which will allow us to optimise this.

Criterion Optimisation

- Two sets of parameters - the cluster mean values $\mathbf{m}_k$ and the cluster allocation indicator variables z_{kn}

Criterion Optimisation

UNIVERSITY of GLASGOW

- Two sets of parameters - the cluster mean values $\mathbf{m}_k$ and the cluster allocation indicator variables z_{kn}
- Optimise our criterion over each set of variables by holding one set fixed - similar to EM

Criterion Optimisation

- Two sets of parameters - the cluster mean values $\mathbf{m}_k$ and the cluster allocation indicator variables z_{kn}
- Optimise our criterion over each set of variables by holding one set fixed - similar to EM
- Given current z_{kn} optimal value of mean vectors $\mathbf{m}_k$ simply the estimates based on data points allocated to each cluster

Criterion Optimisation

UNIVERSITY of GLASGOW

- Two sets of parameters - the cluster mean values $\mathbf{m}_k$ and the cluster allocation indicator variables z_{kn}
- Optimise our criterion over each set of variables by holding one set fixed - similar to EM
- Given current z_{kn} optimal value of mean vectors $\mathbf{m}_k$ simply the estimates based on data points allocated to each cluster
- Therefore given each z_{kn} we obtain our K-means by

$$\mathbf{m}_k = \frac{\sum_{n=1}^{N} z_{kn}\mathbf{x}_{kn}}{\sum_{n'=1}^{N} z_{kn'}}$$

Criterion Optimisation

UNIVERSITY *of* GLASGOW

- Now given each of our new $\mathbf{m}_k$ we need to update the values of our indicator values z_{kn}.

Criterion Optimisation

UNIVERSITY of GLASGOW

- Now given each of our new $\mathbf{m}_k$ we need to update the values of our indicator values z_{kn}.
- From the expression for $\mathcal{E}_K$ we can see that each $\mathbf{x}_n$ should be assigned to the cluster k for which it has the shortest distance to the cluster centre

Criterion Optimisation

UNIVERSITY of GLASGOW

- Now given each of our new $\mathbf{m}_k$ we need to update the values of our indicator values z_{kn}.
- From the expression for $\mathcal{E}_K$ we can see that each $\mathbf{x}_n$ should be assigned to the cluster k for which it has the shortest distance to the cluster centre
- That is $||\mathbf{x}_n - \mathbf{m}_k||^2$ is the smallest for all values of $k = 1 \cdots K$

Criterion Optimisation

- Now given each of our new $\mathbf{m}_k$ we need to update the values of our indicator values z_{kn}.
- From the expression for $\mathcal{E}_K$ we can see that each $\mathbf{x}_n$ should be assigned to the cluster k for which it has the shortest distance to the cluster centre
- That is $||\mathbf{x}_n - \mathbf{m}_k||^2$ is the smallest for all values of $k = 1 \cdots K$
- So $z_{kn} = 1$ for k which yields the minimum of $||\mathbf{x}_n - \mathbf{m}_k||^2$

Criterion Optimisation

UNIVERSITY of GLASGOW

- Once these values have been redefined then we can go back and revise our estimates of each $\mathbf{m}_k$ and continue this iteration until $\mathcal{E}_K$ converges to some steady value.

Criterion Optimisation

UNIVERSITY of GLASGOW

- Once these values have been redefined then we can go back and revise our estimates of each $\mathbf{m}_k$ and continue this iteration until $\mathcal{E}_K$ converges to some steady value.
- This is very simple algorithm and is the K-Means Clustering algorithm for which a simple Matlab implementation is available for download form the class website.

Illustration

UNIVERSITY
of
GLASGOW

- Image of a 'wee dog' looking out to sea

Illustration

UNIVERSITY of GLASGOW

- Image of a 'wee dog' looking out to sea
- Image is a small 100 $\times$ 100 colour JPG thumbnail and we can represent each pixel in the image as a three-dimensional vector corresponding to the Red, Green & Blue channels of the JPEG image

Illustration

- Image of a 'wee dog' looking out to sea
- Image is a small 100 $\times$ 100 colour JPG thumbnail and we can represent each pixel in the image as a three-dimensional vector corresponding to the Red, Green & Blue channels of the JPEG image
- Segment the image into self consistent regions corresponding to the background or foreground (i.e. the dog) then we need to cluster the pixels together based on their Red, Green & Blue representations

Illustration

UNIVERSITY *of* GLASGOW

- Image of a 'wee dog' looking out to sea
- Image is a small 100 $\times$ 100 colour JPG thumbnail and we can represent each pixel in the image as a three-dimensional vector corresponding to the Red, Green & Blue channels of the JPEG image
- Segment the image into self consistent regions corresponding to the background or foreground (i.e. the dog) then we need to cluster the pixels together based on their Red, Green & Blue representations
- Employ K-Means to segment image

Illustration

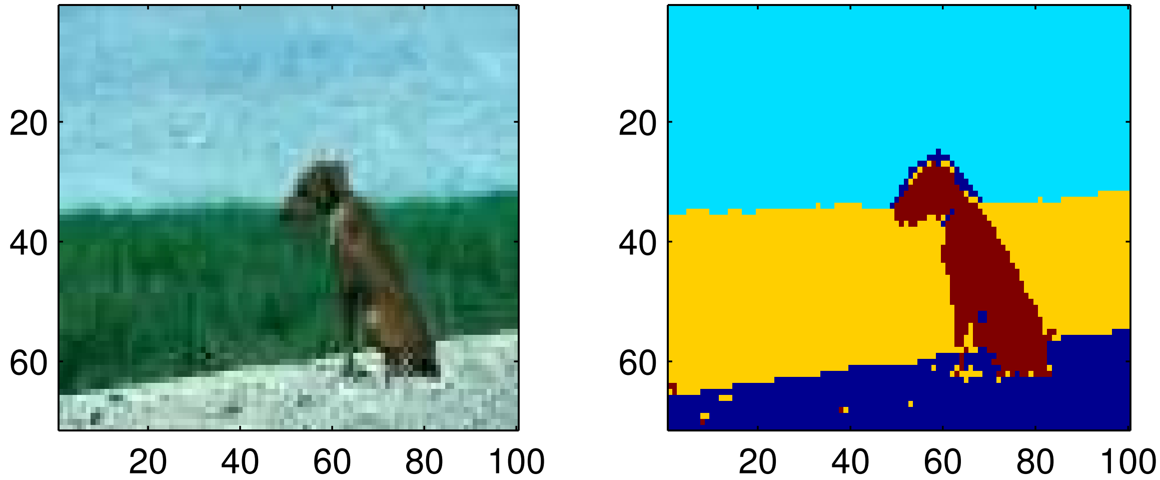

Figure 2: The image of a dog looking out to sea, the right hand image shows the areas of the original image which have been allocated to one of four possible clusters. We have managed to segment the image based on the regions corresponding to water, grass, road, dog

K-Means Issues

UNIVERSITY
of
GLASGOW

- The converged solution will vary with initial conditions

K-Means Issues

UNIVERSITY
of
GLASGOW

- The converged solution will vary with initial conditions
- The algorithm relies on a value of K being supplied by user

K-Means Issues

UNIVERSITY of GLASGOW

- The converged solution will vary with initial conditions
- The algorithm relies on a value of K being supplied by user
- As we shall see later K-Means relies on splitting feature space using linear hyper-planes

K-Means Issues

- The converged solution will vary with initial conditions
- The algorithm relies on a value of K being supplied by user
- As we shall see later K-Means relies on splitting feature space using linear hyper-planes
- Nonlinear feature dependencies exist then K-Means will fail.

K-Means Issues

UNIVERSITY *of* GLASGOW

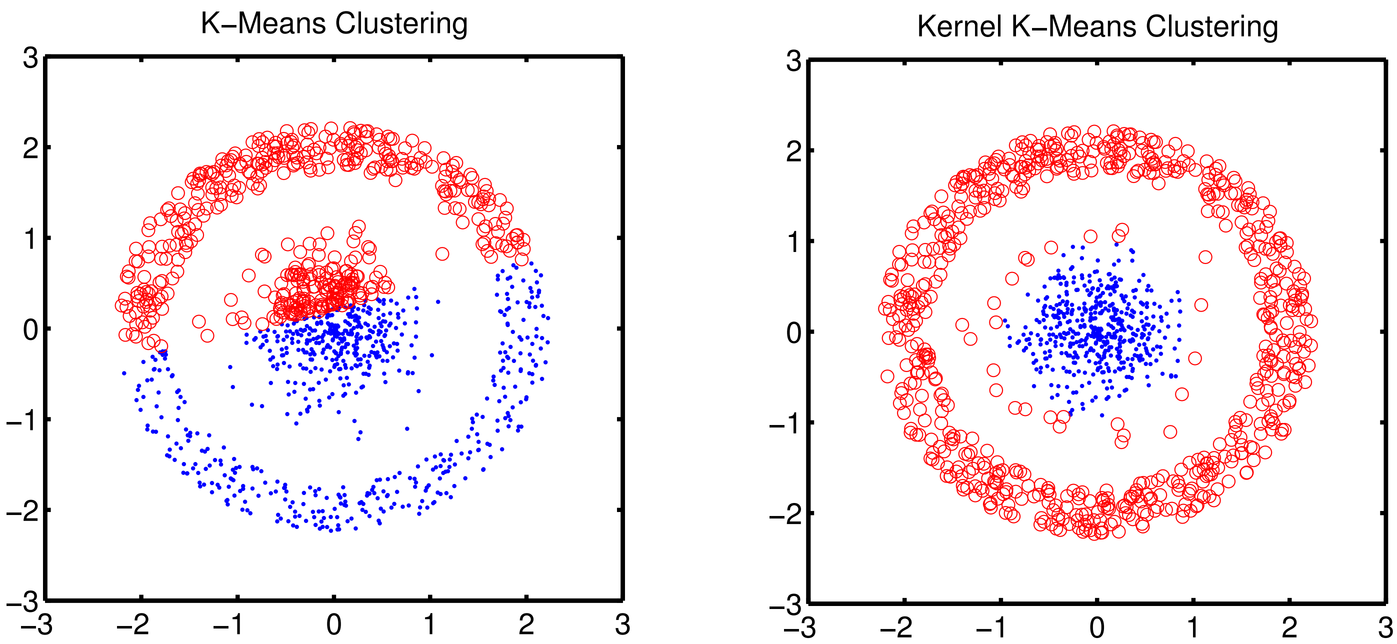

Figure 3: The data is generated such that two consistent clusters both share the same mean but are distributed as a Gaussian cloud and uniformly within a unit width annulus centered at the origin. The left hand plot shows the clustering using the standard K-Means algorithm. It fails to obtain a reasonable clustering. The right hand plot shows the clustering obtained by using Kernel K-Means clustering. A more sensible segmentation of the data is obtained.

Kernel K-Means

UNIVERSITY of GLASGOW

- The clustering criterion upon which the K-Means algorithm is based is can be written as follows

$$
\begin{aligned}
\mathcal{E}_K &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}||\mathbf{x}_n - \mathbf{m}_k||^2 \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}(\mathbf{x}_n - \mathbf{m}_k)^{\mathsf{T}}(\mathbf{x}_n - \mathbf{m}_k) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\left(\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - 2\mathbf{m}_k^{\mathsf{T}}\mathbf{x}_n + \mathbf{m}_k^{\mathsf{T}}\mathbf{m}_k\right)
\end{aligned}
$$

Kernel K-Means

- Note that

$$\mathbf{m}_k^{\mathsf{T}}\mathbf{x}_n = \frac{1}{N_k}\sum_{m=1}^{N} z_{km}\mathbf{x}_m^{\mathsf{T}}\mathbf{x}_n$$

Kernel K-Means

UNIVERSITY of GLASGOW

- Note that

$$\mathbf{m}_k^\mathsf{T}\mathbf{x}_n = \frac{1}{N_k}\sum_{m=1}^{N} z_{km}\mathbf{x}_m^\mathsf{T}\mathbf{x}_n$$

- and

$$\begin{aligned}\mathbf{m}_k^\mathsf{T}\mathbf{m}_k &= \left(\frac{1}{N_k}\sum_{p=1}^{N} z_{kp}\mathbf{x}_p\right)^2 \\ &= \frac{1}{N_k^2}\sum_{p=1}^{N}\sum_{l=1}^{N} z_{kp}z_{kl}\mathbf{x}_p^\mathsf{T}\mathbf{x}_l\end{aligned}$$

Kernel K-Means

UNIVERSITY of GLASGOW

$$
\begin{aligned}
\mathcal{E}_K^{\phi} &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}||\phi(\mathbf{x}_n) - \mathbf{m}_k^{\phi}||^2 \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\left(\begin{array}{c} \phi(\mathbf{x}_n)^{\mathsf{T}}\phi(\mathbf{x}_n) - \\ \frac{2}{N_k}\sum_{m=1}^{N} z_{km}\phi(\mathbf{x}_m)^{\mathsf{T}}\phi(\mathbf{x}_n) + \\ \frac{1}{N_k^2}\sum_{p=1}^{N}\sum_{l=1}^{N} z_{kp}z_{kl}\phi(\mathbf{x}_p)^{\mathsf{T}}\phi(\mathbf{x}_l) \end{array}\right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\left(\begin{array}{c} K(\mathbf{x}_n, \mathbf{x}_n) - \\ \frac{2}{N_k}\sum_{m=1}^{N} z_{km}K(\mathbf{x}_m, \mathbf{x}_n) \\ \frac{1}{N_k^2}\sum_{p=1}^{N}\sum_{l=1}^{N} z_{kp}z_{kl}K(\mathbf{x}_p, \mathbf{x}_l) \end{array}\right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\delta_{kn}
\end{aligned}
$$

Kernel K-Means

UNIVERSITY
of
GLASGOW

- The first point to notice is that the clustering criterion can be written solely in terms of the kernel functions computed at each of the data point pairs

Kernel K-Means

UNIVERSITY
of
GLASGOW

- The first point to notice is that the clustering criterion can be written solely in terms of the kernel functions computed at each of the data point pairs
- An algorithm can be developed which only requires the step of updating the indicator variables z_{kn} as no explicit updating of cluster mean values of required

Kernel K-Means

UNIVERSITY
of
GLASGOW

- The first point to notice is that the clustering criterion can be written solely in terms of the kernel functions computed at each of the data point pairs
- An algorithm can be developed which only requires the step of updating the indicator variables z_{kn} as no explicit updating of cluster mean values of required
- An implementation of Kernel K-means clustering is available at the course website.

Kernel K-Means

UNIVERSITY of GLASGOW

- The first point to notice is that the clustering criterion can be written solely in terms of the kernel functions computed at each of the data point pairs
- An algorithm can be developed which only requires the step of updating the indicator variables z_{kn} as no explicit updating of cluster mean values of required
- An implementation of Kernel K-means clustering is available at the course website.
- Now we have a kernel-based clustering method which will allow us to segment our data in a nonlinear manner - (hoop & blob)

Kernel K-Means

- We have generalised our K-means algorithm to a more flexible representation which takes account of nonlinear relationships

Kernel K-Means

UNIVERSITY
of
GLASGOW

- We have generalised our K-means algorithm to a more flexible representation which takes account of nonlinear relationships
- There is, of course, a small price to pay for this flexibility

Kernel K-Means

- We have generalised our K-means algorithm to a more flexible representation which takes account of nonlinear relationships
- There is, of course, a small price to pay for this flexibility
- The kernel function used may well have a parameter (or a number of parameters) of its own - which will need to be chosen in some way - so we have added an additional layer of parameters into our representation

Kernel K-Means

UNIVERSITY
of
GLASGOW

- We have generalised our K-means algorithm to a more flexible representation which takes account of nonlinear relationships
- There is, of course, a small price to pay for this flexibility
- The kernel function used may well have a parameter (or a number of parameters) of its own - which will need to be chosen in some way - so we have added an additional layer of parameters into our representation
- This weeks laboratory session will explore these two forms of clustering in some detail

EM & K-Means Clustering

UNIVERSITY of GLASGOW

- In Week 6 we developed an EM algorithm for a Gaussian Mixture model. Lets have another look at this algorithm and make some simplifying assumptions

EM & K-Means Clustering

UNIVERSITY *of* GLASGOW

- In Week 6 we developed an EM algorithm for a Gaussian Mixture model. Lets have another look at this algorithm and make some simplifying assumptions
- Assume that the covariance for each mixture component is simply an identity matrix which is fixed

EM & K-Means Clustering

UNIVERSITY *of* GLASGOW

- In Week 6 we developed an EM algorithm for a Gaussian Mixture model. Lets have another look at this algorithm and make some simplifying assumptions
- Assume that the covariance for each mixture component is simply an identity matrix which is fixed
- There is no need to estimate the covariance matrices as these are set

EM & K-Means Clustering

UNIVERSITY *of* GLASGOW

- In Week 6 we developed an EM algorithm for a Gaussian Mixture model. Lets have another look at this algorithm and make some simplifying assumptions
- Assume that the covariance for each mixture component is simply an identity matrix which is fixed
- There is no need to estimate the covariance matrices as these are set
- Now lets think of the posterior probabilities of data points being allocated to a Gaussian component

EM & K-Means Clustering

UNIVERSITY
of
GLASGOW

- Now lets think of the posterior probabilities of data points being allocated to a Gaussian component

EM & K-Means Clustering

UNIVERSITY of GLASGOW

- Now lets think of the posterior probabilities of data points being allocated to a Gaussian component
- In this case (where each $\mathbf{\Sigma_k}$ is an identity then in the E-step each

$$E\{z_{kn}\} = P(k|\mathbf{x}_n) \propto \exp\left(-\frac{1}{2}||\mathbf{x}_n - \mathbf{m}_k||^2\right)$$

EM & K-Means Clustering

UNIVERSITY *of* GLASGOW

- Now lets think of the posterior probabilities of data points being allocated to a Gaussian component
- In this case (where each $\mathbf{\Sigma_k}$ is an identity then in the E-step each

$$E\{z_{kn}\} = P(k|\mathbf{x}_n) \propto \exp\left(-\frac{1}{2}||\mathbf{x}_n - \mathbf{m}_k||^2\right)$$

- The M-step boils down to

$$\mathbf{m}_k = \frac{\sum_{n=1}^{N} P(k|\mathbf{x}_n)\mathbf{x}_n}{\sum_{m=1}^{N} P(k|\mathbf{x}_n)}$$

EM & K-Means Clustering

UNIVERSITY of GLASGOW

- If we make a hard decision about the expected value of z_{kn} based on the maximum of posterior we should be able to see that the maximum posterior corresponds to the minimum of $||\mathbf{x}_n - \mathbf{m}_k||^2$ which is exactly what we are doing in K-means

EM & K-Means Clustering

UNIVERSITY
of
GLASGOW

- If we make a hard decision about the expected value of z_{kn} based on the maximum of posterior we should be able to see that the maximum posterior corresponds to the minimum of $||\mathbf{x}_n - \mathbf{m}_k||^2$ which is exactly what we are doing in K-means
- So if we choose z_{kn} based on the maximum posterior our M-step is precisely the cluster centre updates for K-means clustering

EM & K-Means Clustering

UNIVERSITY of GLASGOW

- If we make a hard decision about the expected value of z_{kn} based on the maximum of posterior we should be able to see that the maximum posterior corresponds to the minimum of $||\mathbf{x}_n - \mathbf{m}_k||^2$ which is exactly what we are doing in K-means
- So if we choose z_{kn} based on the maximum posterior our M-step is precisely the cluster centre updates for K-means clustering
- K-Means clustering can be obtained directly from the EM algorithm from a mixture of unit radius spherical Gaussians where at the E-step a hard decision about component membership is made