COMS21202 Symbols, Patterns and Signals

Problem sheet: Classification and Clustering

1. (**Naive Bayes, revision of COMS10003**) Suppose a naive Bayesian spam filter uses a vocabulary consisting of the words 'Viagra', 'CONFIDENTIAL', 'COMS21202' and 'Gaussian', and has estimated the class-conditional likelihoods of these words occurring in spam and non-spam emails as in Table 1.

Table 1: Class-conditional likelihoods for words in the vocabulary.

word	$P(word\|spam)$	$P(word\|\neg spam)$
Viagra	0.20	0.01
CONFIDENTIAL	0.30	0.05
COMS21202	0.02	0.20
Gaussian	0.05	0.10

Consider three test emails as follows: A contains the word 'Viagra' but none of the others; B contains the word 'CONFIDENTIAL' but none of the others; C contains the words 'COMS21202' and 'Gaussian' but none of the others.

(a) Determine the most likely class of each of these emails by calculating the likelihood ratios $\frac{P(email|spam)}{P(email|\neg spam)}$.

(b) Now assume that typically 10% of your emails are spam. Using MAP estimation, investigate how this affects your predictions.

Answer:

(a) We use $P(word|spam)$ for words that occur in a spam email, and $P(\neg word|spam) = 1 - P(word|spam)$ for words that don't occur in a spam email (similarly for non-spam). We make the naive-Bayesian assumption that the occurrence or absence of words is independent within each class, and so we can decompose the likelihood ratio as follows:

$$\begin{aligned}\frac{P(A|spam)}{P(A|\neg spam)} &= \frac{P(Viagra|spam)P(\neg CONFIDENTIAL|spam)P(\neg COMS21202|spam)P(\neg Gaussian|spam)}{P(Viagra|\neg spam)P(\neg CONFIDENTIAL|\neg spam)P(\neg COMS21202|\neg spam)P(\neg Gaussian|\neg spam)} \\ &= \frac{0.20}{0.01} \times \frac{1-0.30}{1-0.05} \times \frac{1-0.02}{1-0.20} \times \frac{1-0.05}{1-0.10} \\ &= 20 \times 14/19 \times 49/40 \times 19/18 = 19.06\end{aligned}$$

This is (much) larger than 1, so the most likely class of A using ML estimation is spam. You can see that this is mostly due to the presence of the word 'Viagra'; the absence of the word 'CONFIDENTIAL' points somewhat in the direction of non-spam, and the absence of the other two words points somewhat in the direction of spam.

$$\begin{aligned}\frac{P(B|spam)}{P(B|\neg spam)} &= \frac{P(\neg Viagra|spam)P(CONFIDENTIAL|spam)P(\neg COMS21202|spam)P(\neg Gaussian|spam)}{P(\neg Viagra|\neg spam)P(CONFIDENTIAL|\neg spam)P(\neg COMS21202|\neg spam)P(\neg Gaussian|\neg spam)} \\ &= \frac{1-0.20}{1-0.01} \times \frac{0.30}{0.05} \times \frac{1-0.02}{1-0.20} \times \frac{1-0.05}{1-0.10} \\ &= 80/99 \times 6 \times 49/40 \times 19/18 = 6.27\end{aligned}$$

So the most likely class of B using ML estimation is again spam, mostly due to the presence of the word 'CONFIDENTIAL'.

$$\begin{aligned}\frac{P(C|spam)}{P(C|\neg spam)} &= \frac{P(\neg Viagra|spam)P(\neg CONFIDENTIAL|spam)P(COMS21202|spam)P(Gaussian|spam)}{P(\neg Viagra|\neg spam)P(\neg CONFIDENTIAL|\neg spam)P(COMS21202|\neg spam)P(Gaussian|\neg spam)} \\ &= \frac{1-0.20}{1-0.01}\times\frac{1-0.30}{1-0.05}\times\frac{0.02}{0.20}\times\frac{0.05}{0.10} \\ &= 80/99\times 14/19\times 1/10\times 1/2 = 0.030\end{aligned}$$

So the most likely class of C using ML estimation is non-spam – all factors point in that direction, but the presence of 'COMS21202' and 'Gaussian' more so than the absence of the other two.

(b) We now need to take the prior probability of spam into account, and consider the posterior odds

$$\frac{P(spam|email)}{P(\neg spam|email)} = \frac{P(email|spam)P(spam)}{P(email|\neg spam)P(\neg spam)}$$

Equivalently, we can set a different threshold on the likelihood ratio, since $\frac{P(spam|email)}{P(\neg spam|email)} = 1$ *is equivalent to* $\frac{P(email|spam)}{P(email|\neg spam)} = \frac{P(\neg spam)}{P(spam)}$. *With the given prior distribution this threshold is* $0.90/0.10 = 9$, *which only affects the classification of B which is now classified as non-spam: the presence of the word 'CONFIDENTIAL' is not enough to pull us away from the strong prior.*

2. (**Mahalanobis distance**) Given a covariance matrix Σ, the Mahalanobis distance between vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\text{Dis}_M(\mathbf{x},\mathbf{y}|\Sigma) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mathbf{y})}$$

(a) Show that in the 2-D case, points with the same Mahalanobis distance D to a fixed $\mathbf{y} = \mu = (\mu_1\ \mu_2)^{\mathrm{T}}$ describe an ellipse. Use $\Sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with inverse $\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

(Hint: the general form of an ellipse is $Ax_1^2 + Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F = 0$, but in our case it is easier to use $(x_1-\mu_1)$ and $(x_2-\mu_2)$ as variables.)

(b) Derive the ellipse equation for $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

(c) Derive the ellipse equation for $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$.

(d) Derive the ellipse equation for $\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$.

Answer:

(a)

$$\begin{aligned}\sqrt{(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)} &= D \\ (\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu) &= D^2 \\ \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} &= |\Sigma|D^2 \\ d(x_1-\mu_1)^2 - (b+c)(x_1-\mu_1)(x_2-\mu_2) + a(x_2-\mu_2)^2 &= |\Sigma|D^2\end{aligned}$$

(b) $(x_1-\mu_1)^2 + (x_2-\mu_2)^2 = D^2$

(c) $2(x_1-\mu_1)^2-2(x_1-\mu_1)(x_2-\mu_2)+(x_2-\mu_2)^2=D^2$

(d) $(x_1-\mu_1)^2-2r(x_1-\mu_1)(x_2-\mu_2)+(x_2-\mu_2)^2=(1-r)^2D^2$

3. (**Maximum-likelihood decision boundaries using Mahalanobis distance**) The multivariate normal distribution can be expressed in terms of Mahalanobis distance as follows:

$$P(\mathbf{x}|\mu,\Sigma)=\frac{1}{E_d}\exp\left(-\frac{1}{2}\left(\mathrm{Dis}_M(\mathbf{x},\mu|\Sigma)\right)^2\right),\quad E_d=(2\pi)^{d/2}\sqrt{|\Sigma|}$$

Suppose we have two sets of bivariate normally distributed data points whose covariance matrices have the same determinant, then $P(\mathbf{x}|\mu_1,\Sigma_1)=P(\mathbf{x}|\mu_2,\Sigma_2)$ if and only if $\mathrm{Dis}_M(\mathbf{x},\mu_1|\Sigma_1)=\mathrm{Dis}_M(\mathbf{x},\mu_2|\Sigma_2)$. In other words, the decision boundary is formed by the set of points that have the same Mahalanobis distance to both means.

Use the results of the previous question to derive equations for the decision boundary in each of the cases below, using means $\mu_1=(1,1)$ and $\mu_2=(-1,-1)$.

(Sanity check: the first two should give a straight line through the mid-point between the two means, which is $(0,0)$.)

(a) $\Sigma_1=\Sigma_2=\Sigma=\begin{pmatrix}1&0\\0&1\end{pmatrix}$ – i.e., the real-valued equivalent of the naive Bayes assumption.

(b) $\Sigma_1=\Sigma_2=\Sigma=\begin{pmatrix}1&1\\1&2\end{pmatrix}$.

(c) $\Sigma_1=\begin{pmatrix}1&r\\r&1\end{pmatrix}$ and $\Sigma_2=\begin{pmatrix}1&-r\\-r&1\end{pmatrix}$.

Answer:

(a)

$$\begin{aligned}(x_1-1)^2+(x_2-1)^2&=(x_1+1)^2+(x_2+1)^2\\x_1^2-2x_1+1+x_2^2-2x_2+1&=x_1^2+2x_1+1+x_2^2+2x_2+1\\x_1+x_2&=0\end{aligned}$$

This is a line through the origin with slope -1: the perpendicular bisector of the line connecting the two means.

(b)

$$\begin{aligned}2(x_1-1)^2-2(x_1-1)(x_2-1)+(x_2-1)^2&=2(x_1+1)^2-2(x_1+1)(x_2+1)+(x_2+1)^2\\2x_1^2-4x_1+2-2x_1x_2+2x_1+2x_2-2+x_2^2-2x_2+1&=2x_1^2+4x_1+2-2x_1x_2-2x_1-2x_2-2+x_2^2+2x_2+1\\x_1&=0\end{aligned}$$

This is a vertical line through the origin.

(c)

$$\begin{aligned}(x_1-1)^2-2r(x_1-1)(x_2-1)+(x_2-1)^2&=(x_1+1)^2+2r(x_1+1)(x_2+1)+(x_2+1)^2\\x_1^2-2x_1+1-2rx_1x_2+2rx_1+2rx_2-2r+x_2^2-2x_2+1&=x_1^2+2x_1+1+2rx_1x_2+2rx_1+2rx_2+2r+x_2^2+2x_2+1\\-2x_1-2rx_1x_2-2r-2x_2&=2x_1+2rx_1x_2+2r+2x_2\\x_1+rx_1x_2+r+x_2&=0\end{aligned}$$

This is a hyperbole. Notice that if $r=0$ we get (a) as a special case.

4. (**Decision trees**) Suppose we have a training set of 32 spam emails and 32 non-spam emails, and the numbers of emails containing particular words are as in Table 2. We want to build a decision tree using these words as boolean features: if the word occurs in an email the feature is true, else it is false. Which feature results in the best split, as measured by information gain?

Table 2: Numbers of spam and non-spam emails containing particular words.

word	spam	non-spam
Viagra	15	1
CONFIDENTIAL	28	4
COMS21202	1	15
Gaussian	4	12

Answer:
Information gain is calculated in this case as the entropy of the training set minus the weighted average entropy after splitting on the feature. The training set has entropy $-(32/64)\log_2(32/64) - (32/64)\log_2(32/64) = -\log_2(1/2) = 1$ *(i.e. a uniform distribution: no calculation necessary!)*

- *16 emails in the training set contain the word 'Viagra', 15 of which are spam and 1 of which is non-spam. The entropy of those emails is* $-(15/16)\log_2(15/16) - (1/16)\log_2(1/16) = 0.337$.
  The remaining 48 emails in the training set do not contain the word 'Viagra', 17 of which are spam and 31 of which are non-spam. The entropy of those emails is $-(17/48)\log_2(17/48) - (31/48)\log_2(31/48) = 0.938$.
  The weighted average of these two entropies is $(16/64)0.337 + (48/64)0.938 = 0.788$. *The decrease in entropy before and after splitting is thus* $1 - 0.788 = 0.212$.
- *32 emails in the training set contain the word 'CONFIDENTIAL', 28 of which are spam and 4 of which are non-spam. The entropy of those emails is* $-(28/32)\log_2(28/32) - (4/32)\log_2(4/32) = 0.544$.
  The other 32 emails in the training set do not contain the word 'CONFIDENTIAL', 4 of which are spam and 28 of which are non-spam. This is the mirror image of emails containing the word 'CONFIDENTIAL', and so the entropy is again 0.544.
  The weighted average of these two entropies is of course 0.544, *and thus the decrease in entropy before and after splitting on this feature is* 0.456.
- *For emails containing the word 'COMS21202', a distribution of 1 spam and 15 non-spam works out the same as for 'Viagra' (why?), and thus the entropy is* 0.337. *The entropy of emails not containing 'COMS21202' again works out the same as in the 'Viagra' case (why?), i.e.,* 0.938. *We conclude that the decrease in entropy for 'COMS21202' is the same as for 'Viagra', i.e.,* 0.212.
- *Without giving the detailed calculation, the decrease in entropy for 'Gaussian' is* 0.062.

The best feature to split on is thus occurrence of the word 'CONFIDENTIAL', as this results in the biggest decrease in impurity on average. The worst feature is 'Gaussian', and the two remaining features are of equal quality.

5. (**Distance metrics**) Calculate the pairwise distances between the following points using the L_k norm with $k = 1$, $k = 2$ and $k = \infty$:

   (a) $x = 1$, $y = 2$ and $z = 4$.

   (b) $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$.

(c) $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$.

Answer:

(a) In the one-dimensional case $L_k(x,y) = \left(|x-y|^k\right)^{1/k} = |x-y|$*, independently of k. So we have* $L_k(x,y) = 1$*,* $L_k(x,z) = 3$ *and* $L_k(y,z) = 2$.

(b) i. $L_1(\mathbf{x},\mathbf{y}) = |1-2|+|2-3| = 2$*,* $L_1(\mathbf{x},\mathbf{z}) = |1-2|+|2-4| = 3$*, and* $L_1(\mathbf{y},\mathbf{z}) = |2-2|+|3-4| = 1$.

ii. $L_2(\mathbf{x},\mathbf{y}) = \sqrt{(1-2)^2+(2-3)^2} = \sqrt{2} = 1.41$*,* $L_2(\mathbf{x},\mathbf{z}) = \sqrt{(1-2)^2+(2-4)^2} = \sqrt{5} = 2.24$*, and* $L_2(\mathbf{y},\mathbf{z}) = \sqrt{(2-2)^2+(3-4)^2} = \sqrt{1} = 1$.

iii. $L_\infty(\mathbf{x},\mathbf{y}) = \max(|1-2|,|2-3|) = 1$*,* $L_\infty(\mathbf{x},\mathbf{z}) = \max(|1-2|,|2-4|) = 2$*, and* $L_\infty(\mathbf{y},\mathbf{z}) = \max(|2-2|,|3-4|) = 1$.

(c) i. $L_1(\mathbf{x},\mathbf{y}) = |1-2|+|2-3|+|3-4| = 3$*,* $L_1(\mathbf{x},\mathbf{z}) = |1-2|+|2-4|+|3-6| = 6$*, and* $L_1(\mathbf{y},\mathbf{z}) = |2-2|+|3-4|+|4-6| = 3$.

ii. $L_2(\mathbf{x},\mathbf{y}) = \sqrt{(1-2)^2+(2-3)^2+(3-4)^2} = \sqrt{3} = 1.73$*,* $L_2(\mathbf{x},\mathbf{z}) = \sqrt{(1-2)^2+(2-4)^2+(3-6)^2} = \sqrt{14} = 3.74$*, and* $L_2(\mathbf{y},\mathbf{z}) = \sqrt{(2-2)^2+(3-4)^2+(4-6)^2} = \sqrt{5} = 2.24$.

iii. $L_\infty(\mathbf{x},\mathbf{y}) = \max(|1-2|,|2-3|,|3-4|) = 1$*,* $L_\infty(\mathbf{x},\mathbf{z}) = \max(|1-2|,|2-4|,|3-6|) = 3$*, and* $L_\infty(\mathbf{y},\mathbf{z}) = \max(|2-2|,|3-4|,|4-6|) = 2$.

6. (**Nearest-neighbour classification**) Assume $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ are three training instances labelled $+$, $+$ and $-$, respectively. Derive the k-nearest neighbour prediction for the test points $\mathbf{p} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\mathbf{q} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$, using Euclidean distance, for $k = 1$ and $k = 3$.

Answer:

- *($k = 1$) We have* $L_2(\mathbf{p},\mathbf{x}) = 1$*,* $L_2(\mathbf{p},\mathbf{y}) = \sqrt{5}$*, and* $L_2(\mathbf{p},\mathbf{z}) = \sqrt{10}$*. So* $\mathbf{x}$ *is the nearest neighbour of* $\mathbf{p}$*, and we predict* $+$.
  Similarly, $L_2(\mathbf{q},\mathbf{x}) = 2$*,* $L_2(\mathbf{q},\mathbf{y}) = \sqrt{2}$*, and* $L_2(\mathbf{q},\mathbf{z}) = 1$*. So* $\mathbf{z}$ *is the nearest neighbour of* $\mathbf{q}$*, and we predict* $-$.
- *($k = 3$) In this case k is equal to the size of the training set, and k-nearest neighbour will always predict the majority class in the training set (i.e., $+$).*

7. (***K*-means clustering**) Assume $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ form a cluster. Using Euclidean distance:
   (a) calculate the sum of the squared distances of cluster points to the mean (within-cluster scatter); and
   (b) calculate the pairwise squared distance between cluster points, summed over all pairs.

Answer:

(a) The cluster mean is $\mu = 1/3 \begin{bmatrix} 1+2+2 \\ 2+3+4 \end{bmatrix} = \begin{bmatrix} 5/3 \\ 3 \end{bmatrix}$*. We have* $L_2(\mathbf{x},\mu) = \sqrt{13/9}$*,* $L_2(\mathbf{y},\mu) = \sqrt{1/9}$*, and* $L_2(\mathbf{z},\mu) = \sqrt{10/9}$*, so the total squared distance to the mean is* $24/9 = 8/3$.

(b) The Euclidean distances between cluster points have already been calculated in Q5b as $\sqrt{2}$, $\sqrt{5}$, and 1, so the total squared distance is 8. Notice this is equal to the sum of squared distances to the mean times the number of data points, which is true in general.

8. (***K*-means clustering**) Let $\mathbf{X} = \begin{bmatrix} 1 & 2 & 2 & 2 & 3 \\ 2 & 3 & 4 & 1 & 2 \end{bmatrix}$ be a data matrix, with data points in columns. Calculate new centroids $\mu_i(1)$ after one iteration of K-means for each of the following initial centroids:

   (a) $\mu_1(0) = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$;

   (b) $\mu_1(0) = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$;

   (c) $\mu_1(0) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$.

   Also indicate in each of these cases whether K-means has converged. Which of these sets of centroids are optimal?

   Answer:

   (a) The first three points are closest to $\mu_1(0)$, and the last two points are closest to $\mu_2(0)$ (this can be seen by plotting the points in 2-D space, or you can calculate the relevant Euclidean distances). This gives $\mu_1(1) = \begin{bmatrix} 5/3 \\ 3 \end{bmatrix}$ and $\mu_2(1) = \begin{bmatrix} 5/2 \\ 3/2 \end{bmatrix}$. You will then find that again the first three points are closest to $\mu_1(1)$ and the last two points are closest to $\mu_2(1)$, so this is a convergence point. We now calculate the sum of squared distances to the centroid for each cluster, this is $8/3$ for the first cluster (see Q7) and 1 for the second cluster.

   (b) The first, fourth and fifth points are closest to $\mu_1(0)$, and the second and third points are closest to $\mu_2(0)$. This gives $\mu_1(1) = \begin{bmatrix} 2 \\ 5/3 \end{bmatrix}$ and $\mu_2(1) = \begin{bmatrix} 2 \\ 7/2 \end{bmatrix}$. This is also a convergence point. Average pairwise squared distance is $8/3$ for the three-point cluster and $1/2$ for the second cluster, so this is a more compact clustering.

   (c) Same solution as in Q8b.

9. (**Gaussian mixtures**) Using the same data as in the previous question, use Gaussian mixtures and one iteration of Expectation-Maximisation to calculate new centroids from $\mu_1(0) = \begin{bmatrix} 2 \\ 2.15 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 2 \\ 2.65 \end{bmatrix}$.

   Answer:

   The lectures covered the 1-D case; the 2-D case is a straightforward extension using Euclidean distance.

   Expectation: we need to calculate the expected cluster assignments z_{ij} given the initial centroids. For $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, we have $d_{11}(0) = L_2(\mathbf{x_1}, \mu_1(0)) = \sqrt{1.02}$ and $d_{12}(0) = L_2(\mathbf{x_1}, \mu_2(0)) = \sqrt{1.42}$. Because we assume the clusters to be Gaussian with identity covariance, the probability of point $\mathbf{x}_1$ to belong to cluster j is proportional to $e^{-(d_{1j}(0))^2/2}$, where $d_{1j}(0)$ denotes the Euclidean distance from the j-th cluster mean. Since $z_{11} + z_{12} = 1$ we can use the sum these exponential terms as a normaliser, giving

$$z_{11}(0) = \frac{e^{-(d_{11}(0))^2/2}}{e^{-(d_{11}(0))^2/2} + e^{-(d_{12}(0))^2/2}} = 0.55$$

   and $z_{12}(0) = 1 - z_{11}(0) = 0.45$.

The interpretation of this is that $\mathbf{x}_1$ is slightly more likely to belong to the first cluster, but since the centroids are so close there is little certainty regarding cluster membership.
Similar calculations for the other four points yield the following tables:

$(d_{ij})^2$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$j=1$	*1.02*	*0.72*	*3.42*	*1.32*	*1.02*
$j=2$	*1.42*	*0.12*	*1.82*	*2.72*	*1.42*

z_{ij}	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$j=1$	*0.55*	*0.43*	*0.31*	*0.67*	*0.55*
$j=2$	*0.45*	*0.57*	*0.69*	*0.33*	*0.45*

Maximisation: we need to calculate the maximum-likelihood estimate of the means given the cluster assignments just calculated. Since we have soft cluster membership, we need to take a weighted average over all points:

$$\mu_j(1) = \frac{\sum_{i=1}^{5} z_{ij}(0)\mathbf{x}_i}{\sum_{i=1}^{5} z_{ij}(0)}$$

This yields

$$\mu_1(1) = \frac{z_{11}(0)\mathbf{x}_1 + z_{21}(0)\mathbf{x}_2 + z_{31}(0)\mathbf{x}_3 + z_{41}(0)\mathbf{x}_4 + z_{51}(0)\mathbf{x}_5}{z_{11}(0) + z_{21}(0) + z_{31}(0) + z_{41}(0) + z_{51}(0)} = \begin{bmatrix} 2 \\ 2.15 \end{bmatrix}$$

With a similar calculation we obtain $\mu_2(1) = \begin{bmatrix} 2 \\ 2.65 \end{bmatrix}$. *Thus,* $\mu_j(1) = \mu_j(0)$, *from which we conclude that the EM algorithm has converged.*

It will be helpful to verify the solution by plotting the data points and the means; notice in particular the symmetry of the configuration, and also that we can be most certain about the cluster membership of $\mathbf{x}_3$, *and least certain about* $\mathbf{x}_1$ *and* $\mathbf{x}_5$.