Problem 1:

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	SQRT(25)	SQRT(36)	SQRT(13)	SQRT(50)	SQRT(52)	SQRT(65)	SQRT(5)
A2		0	SQRT(37)	SQRT(18)	SQRT(25)	SQRT(17)	SQRT(10)	SQRT(20)
A3			0	SQRT(25)	SQRT(2)	SQRT(2)	SQRT(53)	SQRT(41)
A4				0	SQRT(13)	SQRT(17)	SQRT(52)	SQRT(2)
A5					0	SQRT(2)	SQRT(45)	SQRT(25)
A6						0	SQRT(29)	SQRT(29)
A7							0	SQRT(58)
A8								0

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:

a) The new clusters (i.e. the examples belonging to each cluster)

b) The centers of the new clusters

c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.

Problem 2:

Assume we are given n points $\{x^1, \ldots, x^n\} \in X^n$.

Let $k : X \times X \to R$ be a kernel, with the corresponding feature map $\Phi : X \to H$. We denote by K the $n \times n$ matrix such that $K_{ij} = k(x^i, x^j)$. We denote by Σ the sample covariance matrix of the images $\{\Phi(x^1), \ldots, \Phi(x^n)\}$ of our data. Let λ and V be an eigenvalue and corresponding eigenvector of Σ.

(a) Write $k(x, x')$ as a function of $\Phi(x)$ and $\Phi(x')$.
(b) Show that V is a linear combination of $\{\Phi(x^1), \ldots, \Phi(x^n)\}$.
(c) We can now write $V = \sum_{i=1}^{n} \alpha_i \Phi(x^i)$. Let us denote by α the n-dimensional vector $(\alpha_1, \ldots, \alpha_n)$. Show that λ,V are solution to $n\lambda K\alpha = K^2\alpha$.
(d) Show how PCA in the feature space can be conducted directly in the original space X, using only k and never computing $\Phi(x)$ for any x.
(e) What is the advantage of never having to compute $\Phi(x)$ explicitly?
(f) Can you get an explicit expression of the principal components without using Φ?

Problem 3:

Consider a random variable x that is categorical with M possible values $1,\cdots,M$. Suppose x is represented as a vector such that $x(i) = 1$ if x takes the ith value, and $\sum_i x(i) = 1$. The distribution of x is described by a mixture of K discrete Multinomial distributions such that

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k)$$

And

$$p(x|\mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)}$$

where π_k denotes the mixing coefficient for the kth component (aka the prior probability that the hidden variable z = k), and μ_k specifies the parameters of the kth component. Specifically, $\mu_k(j)$ represents the probabilities $p(x(j) = 1|z = k)$, and $\sum_j \mu_k(j) = 1$. Given an observed data set $\{x_i\}, i = 1,\cdots,N$, please derive the E and M step equations of the EM algorithm for optimizing the mixing coefficient and the component parameters $\mu_k(j)$ for this distribution.