HW3 Written Questions

4.1 Problem 1

Read Section 9.1 in Bishop that discusses the K-means algorithm, and solve problem 9.1 which asks you to prove that it converges.
Consider the K-means algorithm discussed in Section 9.1. Show that as a consequence of there being a finite number of possible assignments for the set of discrete indicator variables r_{nk} and that for each such assignment there is a unique optimum for the $\{\mu_k\}$ the K-means algorithm must converge after a finite number of iterations.

Answer:
If there are N data points which are assigned to K clusters. There will be $\binom{N}{K}$ possible solutions, which is a finite number. Each iteration, we attempt to lower the J function (9.1) to update parameters $\{ r_{nk} \}$ and $\{ \mu_k \}$ for new clustering based on the previous clustering. At last, we finally will get a clustering which will keep the same parameters in the following iterations, which means the algorithm finally converge in a finite number of iterations.

4.2 Problem 2

Read the beginning of Section 9.2 which describes Gaussian mixture models, and solve Problem 9.3.
Consider a Gaussian mixture model in which the marginal distribution p(z) for the latent variable is given by (9.10) and the conditional distribution p(x|z) for the observed variable is given by (9.11). Show that the marginal distribution p(x) obtained by summing p(z)p(x|z) over all possible values of z is a Gaussian mixture of the form (9.7).

Answer:
Since z is binary 1-of-K coding variable where $z_k \in \{0, 1\}$, which mean from 1 to K, there is only one k makes z_k=1, otherwise, z=0.
Thus, marginal distribution of z (9.10) can also be written as:

$$p(z) = \pi_1^{z_1} \cdot \pi_2^{z_2} \cdot \ldots \cdot \pi_K^{z_K}$$

$1 \times 1 \times \ldots \times \pi_k \times 1 \times 1$,
Where only when z_k=1, the exponent of π_k will be one, otherwise the exponent will be 0.
Then, when comes p(x)

$$p(x) = \sum_z p(z)p(x|z) = \sum_z \prod_{k=1}^{K} (\pi_k^{z_k} N(x|\mu_k, \Sigma_k)^{z_k} = \sum_{k=1}^{K} (\pi_k^{z_k} N(x|\mu_k, \Sigma_k)$$

4.3 Problem 3

Go through Section 12.1.2 which describes the Minimum-error formulation of PCA and perform omitted computations. Specifically, do all the derivations necessary to show that

0. Before (12.9) $\alpha_{nj} = x^T{}_n u_j$

1. (12.12) $z_{nj} = x^T{}_n u_j$

2. (12.13) $b_j = x^T u_j$

3. In case of two-dimensional data space

$$Su_2 = \lambda_2 u_2$$

$$J = \lambda_2$$

Answer:

0. Before (12.9) $\alpha_{nj} = x_n^T u_j$.

· In order to rotate coordinate system to a new system defined by $\{u_i\}$, we need x^T

· Since $\{u_i\}$ is an orthonormal basis, by taking the dot product of x_n with $\{u_i\}$, we get:

$$u_i^T \cdot x_n = \alpha_1 u_1^T + \alpha_2 u_2^T + \cdots + \alpha_i u_i^T + \cdots + \alpha_D u_D^T.$$

$$= \alpha_1 \cdot 0 + \alpha_2 \cdot 0 + \cdots + \alpha_i \cdot 1 + \cdots \alpha_D \cdot 0$$

$$= \alpha_i$$

$$\alpha_i = (x_n^T u_i)^T$$

$$\alpha_i = x_n^T u_i \quad (12.9)$$

1. (12.12) $z_{nj} = x_n^T u_j$.

$\because \begin{cases} J = \frac{1}{N} \sum_{n=1}^{N} \| x_n - \tilde{x}_n \|^2. \\ \tilde{x}_n = \sum_{i=1}^{M} z_{ni} u_i + \sum_{i=M+1}^{D} b_i u_i \end{cases}$

to minimize J

$\rightarrow \partial J / \partial z_{nj} = 2/N \sum_{n=1}^{N} (x_n - \sum_{i=1}^{M} z_{ni} u_i - \sum_{i=M+1}^{D} b_i u_i) \cdot \sum_{i=1}^{M} u_i = 0.$

when $i = 1$ to M.

$$\sum_{n=1}^{N} (x_n - \sum_{i=1}^{M} z_{ni} u_i) = 0$$

$\rightarrow$ for one data point $x_{ni} = \sum_{i=1}^{M} z_{ni} u_i$.

$\rightarrow$ then same as (12.9)

$$x_n^T \cdot u_i = z_1 u_1 + \cdots + z_i u_i + \cdots + z_M u_M$$

$$x_n^T \cdot u_i = z_i$$

$\rightarrow z_{ni} = x_n^T \cdot u_i$.

2. (12.13) $b_j = \bar{x}_n^T u_j$.

$\rightarrow \partial J / \partial b_j = 2/N \sum_{n=1}^{N} (x_n - \sum_{i=1}^{M} z_{ni} u_i - \sum_{i=M+1}^{D} b_i u_i) \sum_{i=M+1}^{D} u_i = 0$

when $i = M+1$ to D.

$$\sum_{n=1}^{N} (\bar{x}_n - \sum_{i=M+1}^{D} b_i u_i) = 0 \quad \text{(}b_i\text{ is constant)}$$

$\rightarrow$ for one data point $\bar{x}_{ni} = \sum_{i=M+1}^{D} b_i u_i$.

$\rightarrow \bar{x}_n^T \cdot u_i = b_{M+1} u_1 + \cdots b_i u_i + \cdots + b_D u_D$

$$\bar{x}_n^T \cdot u_i = b_i.$$

$\rightarrow$ ~~b_i~~ $b_{ni} = \bar{x}_n^T \cdot u_i$.

3. When $D=2$

$\tilde{J} = u_2^T S u_2 + \lambda_2 (1 - u_2^T u_2)$ to minimize $\tilde{J}$. take derivate of u_2, $\frac{\partial \tilde{J}}{\partial u_2}$ and set to zero.

$\frac{\partial \tilde{J}}{\partial u_2} = 2Su_2 - 2\lambda_2 u_2 = 0$

$\rightarrow Su_2 = \lambda_2 u_2$.

Then replace Su_2 by $\lambda_2 u_2$.

$\tilde{J} = \cancel{u_2^T \lambda u_2} + \lambda_2 - \cancel{\lambda_2 u_2^T u_2}$

$\rightarrow \tilde{J} = \lambda_2$