# Homework 4

Aleck Zhao

April 27, 2018

## 1 Analytical (60 points)

**1) Clustering (14 points)** We want to cluster the data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ using the K-means algorithm. The K-means algorithm partitions the $n$ observations into $k$ sets ($k < n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares:

$$\underset{\mathbf{S}=\{S_1,\ldots,S_k\}}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in S_j} \|\mathbf{x}_i - \mu_j\|_2^2$$

where $\mu_j$ is the mean of the examples in $S_j$.

(a) Prove that the objective is non-increasing after each E/M step.

*Proof.* Suppose at some EM step, we have $\theta = \{\mu_1, \cdots, \mu_k\}$ is a vector containing the centroid values. Then $Z = \{r_{nk}\}$ is a binary matrix where $r_{ij} = 1$ if $x_i$ is in cluster $j$, and 0 otherwise. Thus our objective function can be written as

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|x_i - \mu_j\|_2^2$$

Then using the update rules

$$r_{ij}^{new} = \begin{cases} 1 & j = \operatorname{argmin}_\ell \|x_i - \mu_\ell\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

we have that

$$\sum_{j=1}^{k} r_{ij}^{new} \|x_i - \mu_j\|_2^2 = \min_j \|x_i - \mu_j\|_2^2 \leq \|x_i - \mu_\ell\|_2^2, \quad \forall \ell$$

and thus since $\sum_{j=1}^{k} r_{ij} \|x_i - \mu_j\|_2^2 = \|x_i - \mu_\ell\|$ for some $\ell$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij}^{new} \|x_i - \mu_j\|_2^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|x_i - \mu_j\|_2^2$$

so the objective is non-increasing after each E step. Next, using the update rule

$$\mu_j^{new} = \frac{\sum_i r_{ij}^{new} x_i}{\sum_i r_{ij}^{new}}$$

we have the objective (after switching order of summation)

$$\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}^{new} \left\|x_i - \mu_j^{new}\right\|_2^2 = \sum_{j=1}^{k}\sum_{i=1}^{n} r_{ij}^{new} \left\|x_i - \mu_j^{new}\right\|_2^2 = \sum_{j=1}^{k}\sum_{x_i \in S_j} \left\|x_i - \mu_j^{new}\right\|_2^2$$

where $\mu_j^{new} = \mu(S_j)$ as defined. From the previous homework, we know that $\mu(S_j)$ minimizes the sum of squared errors, so

$$\sum_{i=1}^{n} r_{ij}^{new} \left\|x_i - \mu_j^{new}\right\| = \sum_{x_i \in S_j} \left\|x_i - \mu_j^{new}\right\|_2^2 \leq \sum_{x_i \in S_j} \left\|x_i - \mu_j\right\|_2^2 = \sum_{i=1}^{n} r_{ij}^{new} \left\|x_i - \mu_j\right\|_2^2$$

and thus

$$\sum_{j=1}^{k}\sum_{x_i \in S_j} \left\|x_i - \mu_j^{new}\right\|_2^2 = \sum_{j=1}^{k}\sum_{i=1}^{n} r_{ij}^{new} \left\|x_i - \mu_j^{new}\right\| \leq \sum_{j=1}^{k}\sum_{i=1}^{n} r_{ij}^{new} \left\|x_i - \mu_j\right\|_2^2$$

□

(b) One variant is called K-medoids. K-medoids is similar to K-means: both K-means and K-medoids minimize the squared error. However, unlike K-means, K-medoids chooses a training example as a cluster center (medoid). It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances.

Give an example to illustrate that K-medoids will be more robust to outliers than K-means.

**Answer.** Consider the dataset $\{1, 2, 3, 4, 1000\}$ with 1 cluster. Then under K-means, the center is at $\frac{1+2+3+4+1000}{5} = 202$, but under K-medoids, the center is at 3. Thus K-medoids is more robust to outliers because 3 is a better approximation for the cluster center than 202 is.

(c) Another variant of K-means is to change the distance metric to the L1 distance, which is more robust to outliers. In this case, we are minimizing:

$$\min_{S=\{S_1,\ldots,S_k\}} \sum_{j=1}^{k}\sum_{\mathbf{x}_i \in S_j} \left\|\mathbf{x}_j - \mu_j\right\|_1 \tag{1}$$

where $\mu_j$ is the center of $S_j$ w.r.t. to L1 distance. It turns out that computing the medians is involved when computing the cluster center. That is why this algorithm is call K-medians. Does the cluster center have to be an actual instance from the dataset? Explain why.

**Answer.** The median does not need to be an actual instance from the dataset. If there were an even number of points, then the median would be the average between the two "middle" points, which might not be in the dataset.

**2) Expectation-Maximization (12 points)** As we discussed in class, clustering objectives are non-convex. Therefore, different initializations will lead to different clustering solutions.

As an example, take Gaussian mixture model (GMM) clustering. Suppose all the parameters of the GMM are initialized such that all components/clusters have the same mean $\mu_k = \hat{\mu}$ and same covariance $\Sigma_k = \hat{\Sigma}$ for all $k = 1, \ldots, K$.

(a) Prove that the EM algorithm will converge after a single iteration for any choice of the initial mixing coefficients $\pi$.

*Proof.* In the E step, using the fact that $\mu_k = \hat{\mu}$ and $\Sigma_k = \hat{\Sigma}$ for all $k$, we have

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}\left(x_n \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_n \mid \mu_k, \Sigma_k\right)} = \frac{\pi_k \mathcal{N}\left(x_n \mid \hat{\mu}, \hat{\Sigma}\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_n \mid \hat{\mu}, \hat{\Sigma}\right)} = \frac{\pi_k}{\sum_{j=1}^{K} \pi_j} = \pi_k$$

Then in the M step, we have

$$\begin{aligned}
N_k &= \sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} \pi_k = N\pi_k \\
\implies \mu_k^{new} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n = \frac{1}{N\pi_k} \sum_{n=1}^{N} \pi_k x_n = \frac{1}{N} \sum_{n=1}^{N} x_n = \bar{x} \\
\implies \Sigma_k^{new} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(x_n - \mu_k^{new}\right) \left(x_n - \mu_k^{new}\right)^T = \frac{1}{N\pi_k} \sum_{n=1}^{N} \pi_k \left(x_n - \bar{x}\right) \left(x_n - \bar{x}\right)^T \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(x - \bar{x}\right) \left(x - \bar{x}\right)^T = \Sigma \\
\implies \pi_k^{new} &= \frac{N_k}{N} = \frac{N\pi_k}{N} = \pi_k
\end{aligned}$$

Now, we are in the exact same situation as the initialization, where both means and covariances are equal for all $k$ since $\mu_k = \bar{x}$ and $\Sigma_k = \Sigma$. Thus, if we were to perform another EM step, we would find that the parameters $\mu_k^{new} = \bar{x}, \Sigma_k^{new} = \Sigma$ and $\pi_k^{new} = \pi_k$. Thus since $\pi$ was arbitrary, the algorithm will have converged after a single iteration. □

(b) Show that this solution has the property $\mu_k = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$ and $\Sigma_k = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$, where $N$ is the total number of examples. Note that this represents a degenerate case of the mixture model in which all of the components are identical, and in practice we try to avoid such solutions by using an appropriate initialization.

*Proof.* From part (a), at convergence of the EM algorithm, we have

$$\mu_k = \frac{1}{N} \sum_{n=1}^{N} x_n, \qquad \Sigma_k = \frac{1}{N} \sum_{n=1}^{N} \left(x_n - \mu_k\right) \left(x_n - \mu_k\right)^T$$

□

**3) Dimensionality Reduction (9 points)** Show that in the limit $\sigma^2 \to 0$, the posterior mean (2) (3) for the probabilistic PCA model becomes an orthogonal projection onto the principal subspace, as in conventional PCA (4).

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}) \tag{2}$$

$$\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I} \tag{3}$$

$$\mathbf{y} = \mathbf{L}^{-1/2}\mathbf{U}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}) \tag{4}$$

*Proof.* We have

$$E\left[z \mid x\right] = M^{-1}W_{ML}^{T}(x - \bar{x}) = (W^{T}W + \sigma^2 I)^{-1}\left[U\left(L - \sigma^2 I\right)^{1/2} R\right]^{T}(x - \bar{x})$$

In the limit as $\sigma^2 \to 0$, this becomes

$$\begin{aligned} E\left[z \mid x\right] &= \left(W^{T}W\right)^{-1}\left(UL^{1/2}R\right)^{T}(x - \bar{x}) = (W^{T}W)^{-1}\left(R^{T}L^{1/2}U^{T}\right)(x - \bar{x}) \\ &= \left(W^{T}W\right)^{-1}R^{-1}L^{1/2}U^{T}(x - \bar{x}) = \left(RW^{T}W\right)^{-1}L^{1/2}U^{T}(x - \bar{x}) \end{aligned}$$

where we have used the fact that $L$ is diagonal and $R$ is orthogonal, so $R^{T} = R^{-1}$. Now $RW^{T}W = L$, so we finally get

$$E\left[z \mid x\right] = L^{-1}L^{1/2}U^{T}(x - \bar{x}) = L^{-1/2}U^{T}(x - \bar{x}) = y$$

□

**4) Probabilistic PCA (10 points)** Draw a directed probabilistic graphical model representing a discrete mixture of probabilistic PCA models in which each PCA model has its own values of $\mathbf{W}$, $\boldsymbol{\mu}$, and $\sigma^2$. Then draw a modified graph in which these parameter values are shared between the components of the mixture. The graph should represent the model for a single data point $\mathbf{x}$. (Hint: refer to slide 24 of the Dimensionality Reduction lecture as a starting point.)

*Solution.* The graphical models are shown below using plate notation.

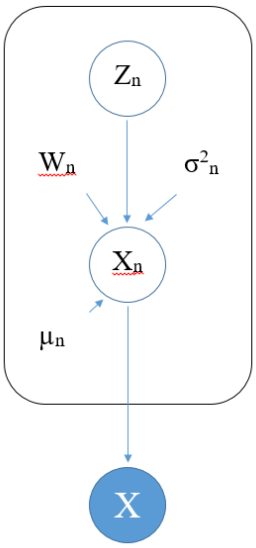


Figure 1: Each PCA model has its own parameters

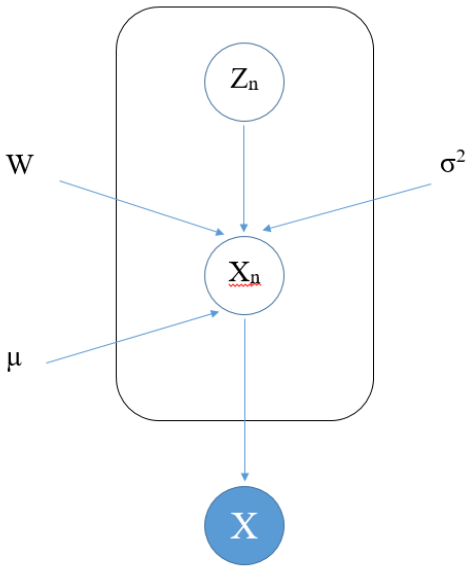


Figure 2: Each PCA shares the parameters

□

**5) Graphical Models (15 points)** Consider the Bayesian Network given in Figure 3. Are the sets **A** and **B** d-separated given set **C** for each of the following definitions of **A**, **B** and **C**? Justify each answer.

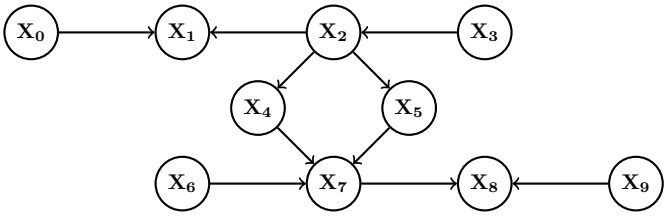


Figure 3: A directed graph

a. $\mathbf{A} = \{X_3\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_1, X_2\}$

**Answer.** Yes. In the path $X_3 \to X_2 \to X_4$, the arrows meet head to tail at $X_2 \in C$. In the path $X_3 \to X_2 \to X_5 \to X_7 \to X_4$, the arrows meet head to tail at $X_2 \in C$. These are the only paths.

If this was an MRF, then $A$ and $B$ would be d-separated because both paths pass through $X_2 \in C$.

b. $\mathbf{A} = \{X_5\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_2, X_7\}$

**Answer.** No. The path $X_5 \to X_7 \to X_4$ has arrows meeting head to head at $X_7 \in C$.

If this was an MRF, then $A$ and $B$ would be d-separated because the two possible paths are $X_5 \to X_2 \to X_4$ and $X_5 \to X_7 \to X_4$, which pass through $X_2 \in C$ and $X_7 \in C$, respectively.

c. $\mathbf{A} = \{X_4\}$, $\mathbf{B} = \{X_6\}$, $\mathbf{C} = \{X_8\}$

**Answer.** No. The path $X_4 \to X_7 \to X_6$ has arrows meeting head to head at $X_7$, which has descendent $X_8 \in C$.

If this was an MRF, then $A$ and $B$ would not be d-separated because there is no path from $X_4$ to $X_6$ passing through $X_8$.

d. $\mathbf{A} = \{X_5\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_2\}$

**Answer.** Yes. In the path $X_5 \to X_2 \to X_4$, the arrows meet tail to tail at $X_2 \in C$. In the path $X_5 \to X_7 \to X_4$, the arrows meet head to head at $X_7$, where $X_2$ is not a descendent.

If this was an MRF, then $A$ and $B$ would not be d-separated because the path $X_5 \to X_7 \to X_4$ does not pass through any node in $C$.

e. $\mathbf{A} = \{X_6, X_9\}$, $\mathbf{B} = \{X_8\}$, $\mathbf{C} = \{X_3, X_1\}$

**Answer.** No. The path $X_6 \to X_7 \to X_8$ has the arrows meet head to tail at $X_7 \notin C$.

If this was an MRF, then $A$ and $B$ would not be d-separated because the path $X_6 \to X_7 \to X_8$ does not pass through any node in $C$.

Now assume that Figure 1 is a Markov Random Field, where each edge is undirected (just drop the direction of each edge.) Re-answer each of the above questions with justifications for your answers.