Homework 0 - Introduction to Probabilistic Graphical Models

kipngeno koech - bkoech

April 5, 2025

2 K-means [20 points]

Given a set of data points $\{\mathbf{x}_n\}_{n=1}^{N}$, k-means clustering minimizes the following distortion measure (also called the "objective" or "clustering cost"):

$$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 \tag{1}$$

where $\boldsymbol{\mu}_k$ is the prototype of the k-th cluster and r_{nk} is a binary indicator variable. If $\mathbf{x}_n$ is assigned to the cluster k, r_{nk} is 1, and otherwise r_{nk} is 0. For each cluster, $\boldsymbol{\mu}_k$ is the prototype representative for all the data points assigned to that cluster.

1. In lecture, we stated but did not prove that $\boldsymbol{\mu}_k$ is the mean of all points associated with the kth cluster, thus motivating the name of the algorithm. You will now prove this statement. Assuming all r_{nk} are known (i.e., assuming you know the cluster assignments of all N data points), show that the objective D is minimized when each $\boldsymbol{\mu}_k$ is chosen as the mean of all data points assigned to cluster k, for any k. This justifies the iterative procedure of k-means[1].
   Let us denote the set of data points assigned to cluster k as $C_k = \{n : r_{nk} = 1\}$. The objective function can be rewritten as:

$$\begin{aligned} D &= \sum_{k=1}^{K} \sum_{n \in C_k} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 && (2) \\ &= \sum_{k=1}^{K} \sum_{n \in C_k} \left(\|\mathbf{x}_n\|_2^2 - 2\mathbf{x}_n^T \boldsymbol{\mu}_k + \|\boldsymbol{\mu}_k\|_2^2\right) && (3) \\ &= \sum_{k=1}^{K} \left(\sum_{n \in C_k} \|\mathbf{x}_n\|_2^2 - 2\boldsymbol{\mu}_k^T \sum_{n \in C_k} \mathbf{x}_n + |C_k| \|\boldsymbol{\mu}_k\|_2^2\right) && (4) \end{aligned}$$

   where $|C_k|$ is the number of data points assigned to cluster k. Now, we can differentiate D with respect to $\boldsymbol{\mu}_k$ and set it to zero to find the optimal $\boldsymbol{\mu}_k$:

$$\frac{\partial D}{\partial \boldsymbol{\mu}_k} = -2 \sum_{n \in C_k} \mathbf{x}_n + 2|C_k| \boldsymbol{\mu}_k = 0 \tag{5}$$

$$\sum_{n \in C_k} \mathbf{x}_n = |C_k| \boldsymbol{\mu}_k \tag{6}$$

$$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{n \in C_k} \mathbf{x}_n \tag{7}$$

   Thus, the optimal $\boldsymbol{\mu}_k$ is the mean of all data points assigned to cluster k.

2. As discussed in lecture, sometimes we wish to scale each feature in order to ensure that "larger" features do not dominate the clustering. Suppose that each data point $\mathbf{x_n}$ is a d-dimensional feature vector and that we scale the jth feature by a factor $w_j > 0$. Letting $\mathbf{W}$ denote a $d \times d$ diagonal matrix with the j'th diagonal entry being w_j, $j = 1, 2, \ldots, d$, we can write our transformed features as $\mathbf{x}' = \mathbf{W}\mathbf{x}$.

   Suppose we fix the r_{nk}, i.e., we take the assignment of data points $\mathbf{x}_n$ to clusters k as given. Our goal is then to find the cluster centers $\boldsymbol{\mu}_k$ that minimize the distortion measure

$$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2. \tag{8}$$

   Show that the cluster centers $\{\boldsymbol{\mu}_k\}$ that do so are given by $\boldsymbol{\mu}_k = \frac{1}{\sum_{n=1}^{N} r_{nk}} \mathbf{W} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n$.
   Let us denote the set of data points assigned to cluster k as $C_k = \{n : r_{nk} = 1\}$. The objective function can be rewritten as:

$$\begin{aligned} D &= \sum_{k=1}^{K} \sum_{n \in C_k} \|\mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 && (9) \\ &= \sum_{k=1}^{K} \sum_{n \in C_k} \left(\|\mathbf{W}\mathbf{x}_n\|_2^2 - 2(\mathbf{W}\mathbf{x}_n)^T \boldsymbol{\mu}_k + \|\boldsymbol{\mu}_k\|_2^2\right) && (10) \\ &= \sum_{k=1}^{K} \left(\sum_{n \in C_k} \|\mathbf{W}\mathbf{x}_n\|_2^2 - 2\boldsymbol{\mu}_k^T \sum_{n \in C_k} \mathbf{W}\mathbf{x}_n + |C_k| \|\boldsymbol{\mu}_k\|_2^2\right) && (11) \end{aligned}$$

[1] More rigorously, one would also need to show that if all $\boldsymbol{\mu}_k$ are known, then r_{nk} can be computed by assigning $\mathbf{x}_n$ to the nearest $\boldsymbol{\mu}_k$. You are not required to do so.

where $|C_k|$ is the number of data points assigned to cluster k. Now, we can differentiate D with respect to $\boldsymbol{\mu}_k$ and set it to zero to find the optimal $\boldsymbol{\mu}_k$:

$$\frac{\partial D}{\partial \boldsymbol{\mu}_k} = -2 \sum_{n \in C_k} \mathbf{W}\mathbf{x}_n + 2|C_k|\boldsymbol{\mu}_k = 0 \tag{12}$$

$$\sum_{n \in C_k} \mathbf{W}\mathbf{x}_n = |C_k|\boldsymbol{\mu}_k \tag{13}$$

$$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{n \in C_k} \mathbf{W}\mathbf{x}_n \tag{14}$$

We can rewrite the sum over C_k as a sum over all points using the indicator variables r_{nk}, and note that $|C_k| = \sum_{n=1}^{N} r_{nk}$:

$$\boldsymbol{\mu}_k = \frac{1}{\sum_{n=1}^{N} r_{nk}} \sum_{n=1}^{N} r_{nk}\mathbf{W}\mathbf{x}_n \tag{15}$$

Since $\mathbf{W}$ is a diagonal matrix that doesn't depend on the summation index n, we can factor it out of the summation:

$$\boldsymbol{\mu}_k = \frac{1}{\sum_{n=1}^{N} r_{nk}} \mathbf{W} \sum_{n=1}^{N} r_{nk}\mathbf{x}_n \tag{16}$$

Thus, the cluster centers $\boldsymbol{\mu}_k$ that minimize the distortion measure are given by:

$$\boldsymbol{\mu}_k = \frac{1}{\sum_{n=1}^{N} r_{nk}} \mathbf{W} \sum_{n=1}^{N} r_{nk}\mathbf{x}_n \tag{17}$$

1 3-Dimensional Principal Component Analysis [20 points]

In this problem, we will perform PCA on 3-dimensional data step by step. We are given three data points:

$$\mathbf{x}_1 = [0, -1, -2] \tag{18}$$
$$\mathbf{x}_2 = [1, 1, 1] \tag{19}$$
$$\mathbf{x}_3 = [2, 0, 1] \tag{20}$$

and we want to find 2 principal components of the given data.

1. First, find the covariance matrix $\mathbf{C}_X = \mathbf{X}^T\mathbf{X}$ where $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 - \bar{\mathbf{x}} \\ \mathbf{x}_2 - \bar{\mathbf{x}} \\ \mathbf{x}_3 - \bar{\mathbf{x}} \end{bmatrix}$, where $\bar{\mathbf{x}} = \frac{1}{3}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3)$ is the mean of the data samples. Then, find the eigenvalues and the corresponding eigenvectors of $\mathbf{C}_X$. Feel free to use any numerical analysis program such as numpy, e.g., `numpy.linalg.eig` can be useful. However, you should explain what you inputted into this program.
   Finding the mean $\bar{\mathbf{x}}$:

$$\begin{aligned} \bar{\mathbf{x}} &= \frac{1}{3}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3) && (21)\\ &= \frac{1}{3}([0, -1, -2] + [1, 1, 1] + [2, 0, 1]) && (22)\\ &= \frac{1}{3}([3, 0, 0]) && (23)\\ &= [1, 0, 0] && (24) \end{aligned}$$

   let us find $\mathbf{X}$:

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{x}_1 - \bar{\mathbf{x}} \\ \mathbf{x}_2 - \bar{\mathbf{x}} \\ \mathbf{x}_3 - \bar{\mathbf{x}} \end{bmatrix} && (25)\\ &= \begin{bmatrix} [0, -1, -2] - [1, 0, 0] \\ [1, 1, 1] - [1, 0, 0] \\ [2, 0, 1] - [1, 0, 0] \end{bmatrix} && (26)\\ &= \begin{bmatrix} [-1, -1, -2] \\ [0, 1, 1] \\ [1, 0, 1] \end{bmatrix} && (27) \end{aligned}$$

   Now, we can find the covariance matrix $\mathbf{C}_X$:

$$\begin{aligned} \mathbf{C}_X &= \mathbf{X}^T\mathbf{X} && (28)\\ &= \begin{bmatrix} [-1, 0, 1] \\ [-1, 1, 0] \\ [-2, 1, 1] \end{bmatrix} \begin{bmatrix} [-1, -1, -2] \\ [0, 1, 1] \\ [1, 0, 1] \end{bmatrix} && (29)\\ &= \begin{bmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 3 & 3 & 6 \end{bmatrix} && (30) \end{aligned}$$

   Now, we can find the eigenvalues and eigenvectors of $\mathbf{C}_X$ using numpy:

```
import numpy as np
C_X = np.array([[2, 1, 3], [1, 2, 3], [3, 3, 6]])
eigenvalues, eigenvectors = np.linalg.eig(C_X)
print("Eigenvalues:", eigenvalues)
print("Eigenvectors:", eigenvectors)
```

   The output will give us the eigenvalues and eigenvectors of the covariance matrix. The eigenvalues are:

$$\lambda_1 = 9.00 \tag{31}$$
$$\lambda_2 = 1.00 \tag{32}$$
$$\lambda_3 = 0.00 \tag{33}$$

   The corresponding eigenvectors are:

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \mathbf{u}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \mathbf{u}_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} \tag{34}$$

   I used the numpy function `numpy.linalg.eig` to compute the eigenvalues and eigenvectors of the covariance matrix. The input to this function was the covariance matrix $\mathbf{C}_X$ that we computed above.

2. Using the result above, find the first two principal components of the given data.
   The first two principal components are the eigenvectors corresponding to the two largest eigenvalues. In this case, the first two principal components are:

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \quad \text{(corresponding to } \lambda_1 = 9.00\text{)} \tag{35}$$

$$\mathbf{u}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad \text{(corresponding to } \lambda_2 = 1.00\text{)} \tag{36}$$

3. Now we want to represent the data $\mathbf{x}_1, \cdots, \mathbf{x}_3$ using a 2-dimensional subspace instead of a 3-dimensional one. PCA gives us the 2-D plane which minimizes the difference between the original data and the data projected to the 2-dimensional plane. In other words, $\mathbf{x}_i$ can be approximated as:

$$\tilde{\mathbf{x}}_i = a_{i1}\mathbf{u}_1 + a_{i2}\mathbf{u}_2 + \bar{\mathbf{x}}, \tag{37}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the principal components we found in 3.b. Figure 1 gives an example of what this might look like.

Figure 1: Example of 2-D plane spanned by the first two principal components.

Find a_{i1}, a_{i2} for $i = 1, 2, 3$. Then, find the $\tilde{\mathbf{x}}_i$'s and the difference between $\tilde{\mathbf{x}}_i$ and $\mathbf{x}_i$, i.e., $||\tilde{\mathbf{x}}_i - \mathbf{x}_i||_2$ for $i = 1, 2, 3$. (Again, feel free to use any numerical analysis program to get the final answer. But, show your calculation process.)

To find a_{i1} and a_{i2}, we can project the original data points onto the principal components:
Let us project $\mathbf{x}_1$ onto the first two principal components:

$$\mathbf{x}_1 = a_{11}\mathbf{u}_1 + a_{12}\mathbf{u}_2 + \bar{\mathbf{x}} \tag{38}$$
$$\mathbf{x}_1 - \bar{\mathbf{x}} = a_{11}\mathbf{u}_1 + a_{12}\mathbf{u}_2 \tag{39}$$
$$[-1, -1, -2] = a_{11}\begin{pmatrix}1\\1\\2\end{pmatrix} + a_{12}\begin{pmatrix}-1\\1\\0\end{pmatrix} \tag{40}$$
$$[-1, -1, -2] = \begin{pmatrix}a_{11} - a_{12}\\a_{11} + a_{12}\\2a_{11}\end{pmatrix} \tag{41}$$

Now, we can set up a system of equations:

$$-1 = a_{11} - a_{12} \tag{42}$$
$$-1 = a_{11} + a_{12} \tag{43}$$
$$-2 = 2a_{11} \tag{44}$$

From the third equation, we can find a_{11}:

$$2a_{11} = -2 \tag{45}$$
$$a_{11} = -1 \tag{46}$$

Now, we can substitute a_{11} into the first two equations to find a_{12}:

$$-1 = -1 - a_{12} \tag{47}$$
$$-1 = -1 + a_{12} \tag{48}$$
$$a_{12} = 0 \tag{49}$$

So, for $\mathbf{x}_1$, we have:

$$a_{11} = -1 \tag{50}$$
$$a_{12} = 0 \tag{51}$$

Now, we can find $\tilde{\mathbf{x}}_1$:

$$\tilde{\mathbf{x}}_1 = a_{11}\mathbf{u}_1 + a_{12}\mathbf{u}_2 + \bar{\mathbf{x}} \tag{52}$$
$$= -1\begin{pmatrix}1\\1\\2\end{pmatrix} + 0\begin{pmatrix}-1\\1\\0\end{pmatrix} + [1, 0, 0] \tag{53}$$
$$= [-1, -1, -2] + [1, 0, 0] \tag{54}$$
$$= [0, -1, -2] \tag{55}$$

Now, we can find the difference between $\tilde{\mathbf{x}}_1$ and $\mathbf{x}_1$:

$$||\tilde{\mathbf{x}}_1 - \mathbf{x}_1||_2 = ||[0, -1, -2] - [0, -1, -2]||_2 \tag{56}$$
$$= ||[0, 0, 0]||_2 \tag{57}$$
$$= 0 \tag{58}$$

Now, we can repeat the process for $\mathbf{x}_2$:

$$\mathbf{x}_2 = a_{21}\mathbf{u}_1 + a_{22}\mathbf{u}_2 + \bar{\mathbf{x}} \tag{59}$$
$$\mathbf{x}_2 - \bar{\mathbf{x}} = a_{21}\mathbf{u}_1 + a_{22}\mathbf{u}_2 \tag{60}$$
$$[0, 1, 1] = a_{21}\begin{pmatrix}1\\1\\2\end{pmatrix} + a_{22}\begin{pmatrix}-1\\1\\0\end{pmatrix} \tag{61}$$
$$[0, 1, 1] = \begin{pmatrix}a_{21} - a_{22}\\a_{21} + a_{22}\\2a_{21}\end{pmatrix} \tag{62}$$

Now, we can set up a system of equations:

$$0 = a_{21} - a_{22} \tag{63}$$
$$1 = a_{21} + a_{22} \tag{64}$$
$$1 = 2a_{21} \tag{65}$$

From the third equation, we can find a_{21}:

$$2a_{21} = 1 \tag{66}$$
$$a_{21} = \frac{1}{2} \tag{67}$$

Now, we can substitute a_{21} into the first two equations to find a_{22}:

$$0 = \frac{1}{2} - a_{22} \tag{68}$$
$$0 = \frac{1}{2} + a_{22} \tag{69}$$
$$a_{22} = \frac{1}{2} \tag{70}$$

So, for $\mathbf{x}_2$, we have:

$$a_{21} = \frac{1}{2} \tag{71}$$
$$a_{22} = \frac{1}{2} \tag{72}$$

Now, we can find $\tilde{\mathbf{x}}_2$:

$$\begin{aligned}
\tilde{\mathbf{x}}_2 &= a_{21}\mathbf{u}_1 + a_{22}\mathbf{u}_2 + \bar{\mathbf{x}} && (73)\\
&= \frac{1}{2}\begin{pmatrix}1\\1\\2\end{pmatrix} + \frac{1}{2}\begin{pmatrix}-1\\1\\0\end{pmatrix} + [1,0,0] && (74)\\
&= \begin{pmatrix}\frac{1}{2}-\frac{1}{2}\\ \frac{1}{2}+\frac{1}{2}\\ \frac{1}{2}\end{pmatrix} + [1,0,0] && (75)\\
&= [0,1,1] + [1,0,0] && (76)\\
&= [1,1,1] && (77)
\end{aligned}$$

Now, we can find the difference between $\tilde{\mathbf{x}}_2$ and $\mathbf{x}_2$:

$$\begin{aligned}
||\tilde{\mathbf{x}}_2 - \mathbf{x}_2||_2 &= ||[1,1,1] - [1,1,1]||_2 && (78)\\
&= ||[0,0,0]||_2 && (79)\\
&= 0 && (80)
\end{aligned}$$

Now, we can repeat the process for $\mathbf{x}_3$:

$$\begin{aligned}
\mathbf{x}_3 &= a_{31}\mathbf{u}_1 + a_{32}\mathbf{u}_2 + \bar{\mathbf{x}} && (81)\\
\mathbf{x}_3 - \bar{\mathbf{x}} &= a_{31}\mathbf{u}_1 + a_{32}\mathbf{u}_2 && (82)\\
[1,0,1] &= a_{31}\begin{pmatrix}1\\1\\2\end{pmatrix} + a_{32}\begin{pmatrix}-1\\1\\0\end{pmatrix} && (83)\\
[1,0,1] &= \begin{pmatrix}a_{31}-a_{32}\\ a_{31}+a_{32}\\ 2a_{31}\end{pmatrix} && (84)
\end{aligned}$$

Now, we can set up a system of equations:

$$1 = a_{31} - a_{32} \tag{85}$$
$$0 = a_{31} + a_{32} \tag{86}$$
$$1 = 2a_{31} \tag{87}$$

From the third equation, we can find a_{31}:

$$2a_{31} = 1 \tag{88}$$
$$a_{31} = \frac{1}{2} \tag{89}$$

Now, we can substitute a_{31} into the first two equations to find a_{32}:

$$1 = \frac{1}{2} - a_{32} \tag{90}$$
$$0 = \frac{1}{2} + a_{32} \tag{91}$$
$$a_{32} = -\frac{1}{2} \tag{92}$$

So, for $\mathbf{x}_3$, we have:

$$a_{31} = \frac{1}{2} \tag{93}$$

$$a_{32} = -\frac{1}{2} \tag{94}$$

Now, we can find $\tilde{\mathbf{x}}_3$:

$$\begin{aligned}
\tilde{\mathbf{x}}_3 &= a_{31}\mathbf{u}_1 + a_{32}\mathbf{u}_2 + \bar{\mathbf{x}} && (95)\\
&= \frac{1}{2}\begin{pmatrix}1\\1\\2\end{pmatrix} - \frac{1}{2}\begin{pmatrix}-1\\1\\0\end{pmatrix} + [1,0,0] && (96)\\
&= \begin{pmatrix}\frac{1}{2}+\frac{1}{2}\\ \frac{1}{2}-\frac{1}{2}\\ \frac{1}{2}\end{pmatrix} + [1,0,0] && (97)\\
&= [1,0,1] + [1,0,0] && (98)\\
&= [2,0,1] && (99)
\end{aligned}$$

Now, we can find the difference between $\tilde{\mathbf{x}}_3$ and $\mathbf{x}_3$:

$$\begin{aligned}
||\tilde{\mathbf{x}}_3 - \mathbf{x}_3||_2 &= ||[2,0,1] - [2,0,1]||_2 && (100)\\
&= ||[0,0,0]||_2 && (101)\\
&= 0 && (102)
\end{aligned}$$

So, we have:

$$\begin{aligned}
&\tilde{\mathbf{x}}_1 = [0,-1,-2] \quad ||\tilde{\mathbf{x}}_1 - \mathbf{x}_1||_2 = 0 && (103)\\
&\tilde{\mathbf{x}}_2 = [1,1,1] \quad ||\tilde{\mathbf{x}}_2 - \mathbf{x}_2||_2 = 0 && (104)\\
&\tilde{\mathbf{x}}_3 = [2,0,1] \quad ||\tilde{\mathbf{x}}_3 - \mathbf{x}_3||_2 = 0 && (105)
\end{aligned}$$

Thus, the differences between the projected data points and the original data points are all zero. This means that the PCA projection perfectly represents the original data in the 2-dimensional subspace spanned by the first two principal components.