**Practical Session 11**

# Clustering

## 1 K-medians

**Problem 1:** Consider a modified version of the $K$-means objective, where we use $L_1$ distance instead.

$$J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_1$$

This variation of the algorithm is called $K$*-medians*. Derive the Lloyd's algorithm for this model.

1. Updating the cluster assignments $z_{ik}$ is the same as for the $K$-means algorithm:

$$z_{ik}^{new} = \begin{cases} 1 & \text{if } \; k = \arg\min_j \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_1 \\ 0 & \text{else.} \end{cases}$$

2. The updates for $\boldsymbol{\mu}_k$'s should solve

$$\boldsymbol{\mu}_k^{new} = \arg\min_{\boldsymbol{\mu}_k} \sum_{i=1}^{N} z_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_1$$

The objective for each single centroid $\boldsymbol{\mu}_k$ can be rewritten as

$$\begin{aligned} J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}_k) &= \sum_{i=1}^{N} z_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_1 \\ &= \sum_{i=1}^{N} z_{ik} \sum_{d=1}^{D} |x_{id} - \mu_{kd}| \end{aligned}$$

Clearly, this is a convex function of $\boldsymbol{\mu}_k$, as it is a sum of piecewise linear functions. We can actually solve for each $\mu_{kd}$ separately, as they do not interact in the objective, by finding the roots of the derivatives.

Observe, that

$$\frac{\partial}{\partial \mu_{kd}} |x_{id} - \mu_{kd}| = \begin{cases} 1 & \text{if } \; \mu_{kd} > x_{id} \\ -1 & \text{if } \; \mu_{kd} < x_{id} \\ 0 & \text{if } \; \mu_{kd} = x_{id}. \end{cases}$$

(Note: actually the absolute value function is not differentiable at 0, so the derivative is undefined. A rigorous treatment of this problem would require us to use subgradients (see `https://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf`), but just "pretending" that the gradient is 0 suffices for our purpose.)

Hence, the derivative of the entire objective is

$$\begin{aligned}\frac{\partial}{\partial \mu_{kd}} J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}) &= \sum_{i=1}^{N} z_{ik} |x_{id} - \mu_{kd}| \\ &= \sum_{i=1}^{N} z_{ik} \mathbb{I}[x_{id} < \mu_{kd}] - \sum_{i=1}^{N} z_{ik} \mathbb{I}[x_{id} > \mu_{kd}] \stackrel{!}{=} 0\end{aligned}$$

The first sum represents "number of points $\boldsymbol{x}_i$ assigned to class $k$, such that $x_{id} < \mu_{kd}$". Each of these sums represents the number of points in class $k$, that are located to the left (right) of the given value of $\mu_{kd}$. Because we want to set the gradient to zero, we are looking for such a $\mu_{kd}$, that along the axis $d$ exactly $N_k/2$ points are to left of it, and another $N_k/2$ points are to the right (where $N_k = \sum_{i=1}^{N} z_{ik}$). This is exactly the definition of a *median*.

Therefore, the optimal update is given as

$$\mu_{kd} = \text{median} \{x_{id}, \text{ such that } z_{ik} = 1\}$$

## 2 Gaussian mixture model

**Problem 2:** Derive the E step update for Gaussian mixture model.

See discussion surrounding Eq. (9.13) in Bishop.

**Problem 3:** Derive the M step update for Gaussian mixture model.

Bishop: Section 9.2.2.

Also, Problems 2 & 3 from Practical 4 solve the same task.

## 3 Expectation Maximization algorithm

**Problem 4:** Consider a mixture model where the components are given by independent Bernoulli variables. This is useful when modelling, e.g., binary images, where each of the $D$ dimensions of the image $\boldsymbol{x}$ corresponds to a different pixel that is either black or white. More formally, we have

$$p(\boldsymbol{x} | \boldsymbol{z} = k) = \prod_{d=1}^{D} \theta_{kd}^{x_d} (1 - \theta_{kd})^{1 - x_d}.$$

That is, for a given mixture index $\boldsymbol{z} = k$, we have a product of independent Bernoullis, where $\theta_{kd}$ denotes the Bernoulli parameter for component $k$ at pixel $d$.

Derive the EM algorithm for the parameters $\theta = \{\theta_{kd} \mid k = 1, \dots, K, d = 1, \dots, D\}$ of a mixture of Bernoullis.

Assume here for simplicity, that the distribution of components $p(\boldsymbol{z})$ is uniform: $p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k} = \prod_{k=1}^{K} (\frac{1}{K})^{z_k}$.

Using our uniformity assumption, we have

$$p(\boldsymbol{z}|\boldsymbol{x}, \theta) \propto p(\boldsymbol{x}|\boldsymbol{z}, \theta).$$

so we obtain the responsibilities as

$$r_{nk}(\theta) = \frac{p(\boldsymbol{x}^{(n)}|\boldsymbol{z} = k, \theta)}{\sum_{j=1}^{K} p(\boldsymbol{x}^{(n)}|\boldsymbol{z} = j, \theta)}.$$

( which is the E-step ) .

It remains to derive the M-step. Similiar to mixture of Gaussians:

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{z}|\mathcal{D},\theta^{(t)})}[\ln p(\mathcal{D}, \boldsymbol{z}|\theta)] &= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}\left(\theta^{(t)}\right) \ln\left(\frac{1}{K} \prod_{d=1}^{D} \theta_{kd}^{x_d^{(n)}} (1 - \theta_{kd})^{1-x_d^{(n)}}\right) \\
&= C + \sum_{n=1}^{N} \underbrace{\sum_{k=1}^{K} r_{nk}\left(\theta^{(t)}\right) \sum_{d=1}^{D} \left(x_d^{(n)} \ln \theta_{kd} + \left(1 - x_d^{(n)}\right) \ln(1 - \theta_{kd})\right)}_{=:\mathcal{L}_n}
\end{aligned}$$

The constant $C$ is independent of $\theta$ and hence irrelevant for further optimization.

We now need to take derivatives w.r.t. $\theta$. We observe that the $\theta_{kd}$ do not interfere nonlinearly, i.e., we can handle the gradients individually:

$$\begin{aligned}
\frac{\partial \mathcal{L}_n}{\partial \theta_{k',d'}} &= \sum_{k=1}^{K} r_{nk}\left(\theta^{(t)}\right) \sum_{d=1}^{D} \left(x_d^{(n)} \frac{\partial \ln \theta_{kd}}{\partial \theta_{k',d'}} + \left(1 - x_d^{(n)}\right) \frac{\partial \ln(1 - \theta_{kd})}{\partial \theta_{k',d'}}\right) \qquad \text{lots of zeros} \\
&= r_{nk'}\left(\theta^{(t)}\right) \left(\frac{x_{d'}^{(n)}}{\theta_{k',d'}} - \frac{1 - x_{d'}^{(n)}}{1 - \theta_{k',d'}}\right)
\end{aligned}$$

Setting this to zero, we obtain the optimal update in a similar fashion as in the standard Bernoulli MLE:

$$\begin{aligned}
&\frac{\partial\, \mathbb{E}_{p(\boldsymbol{z}|\mathcal{D},\theta^{(t)})}[\ln p(\mathcal{D}, \boldsymbol{z}|\theta)]}{\partial \theta_{kd}} = \sum_{n=1}^{N} \frac{\partial \mathcal{L}_n}{\partial \theta_{kd}} = \sum_{n=1}^{N} r_{nk}\left(\theta^{(t)}\right) \left(\frac{x_d^{(n)}}{\theta_{kd}} - \frac{1 - x_d^{(n)}}{1 - \theta_{kd}}\right) \stackrel{!}{=} 0 \\
&\Leftrightarrow \theta_{kd} = \frac{\sum_{n=1}^{N} r_{nk}\left(\theta^{(t)}\right) x_d^{(n)}}{\sum_{n=1}^{N} r_{nk}\left(\theta^{(t)}\right)}
\end{aligned}$$