

## EM k-means algorithm (statistical interpretation)

We are given set of measurements  $\{\mathbf{x}_i\}_{i=1}^N$  and we would like to assign a cluster for each measurement and find the parameters of all clusters. Assume for now that the number of clusters is given. We have  $N$  measurements and  $n$  clusters. The index of a cluster will be modeled as a discrete random variable  $z = j$ , such that the probability of each cluster is

$$p(z = j) = \pi_j \text{ for } j = 1, \dots, n \text{ s.t. } \pi_1 + \dots + \pi_n = 1$$

which is the standard multinomial distribution. We further assume that  $p(\mathbf{x}|z = j) \sim \mathcal{N}(\mu_j, \sigma_j I)$  is a multivariate isotropic Gaussian distribution. We are faced with the problem of estimating unknown parameters of our model, denoted by  $\theta = \{\mu_j, \Sigma_j, \pi_j\}_{j=1}^n$ . Given the parameters of the model the likelihood function of the data has the following form

$$p(\mathbf{x}|\theta) = \sum_{z=1}^n p(\mathbf{x}|z, \theta)p(z|\theta) = \sum_{j=1}^n p(\mathbf{x}|z = j, \theta)\pi_j,$$

where  $z$  is hidden unknown variable. Note that since  $z$  is not observed we have to marginalize over all possible values of  $z$ . One possibility would be to write down the total log likelihood of all data and find the value of parameters of the model which maximize that likelihood. Assuming that the samples are i.i.d the total likelihood has the following form

$$l(\theta, D) = \log \prod_{i=1}^N \sum_{j=1}^n \pi_j \exp\left(-\frac{\|\mathbf{x}_i - \mu_j\|^2}{2\sigma_j^2}\right).$$

This is a complicated optimization problem, involving both discrete and continuous variables. The problem can be simplified by assuming that some of unknowns are known, while estimating the others and vice versa. Here we will provide only an informal description of this algorithm.

For each class, we can compute the conditional expectation of the  $z = j$  given the data and the parameters.

$$w_j = p(z = j|\mathbf{x}, \theta) = \frac{p(\mathbf{x}|z = j, \theta)p(z = j|\pi_j)}{p(\mathbf{x}|\theta)} = \frac{\pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)}{\sum_{i=1}^n \pi_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)}$$

Since each point  $\mathbf{x}$  contributes to  $w_j$  in some proportion, for particular point  $\mathbf{x}_i$  we have

$$w_{ij} = \frac{\pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)}{\sum_{i=1}^n \pi_i \mathcal{N}(\mathbf{x}_i|\mu_i, \Sigma_i)}$$

The optimization algorithm also called EM then proceeds in the following steps: Assume that we have some (random) initial estimates of the means and variances of our models  $\{\mu_j^{(0)}, \Sigma_j^{(0)}, \pi_j^{(0)}\}$ .

1. **Expectation:** Using current estimate of the parameters  $\theta^{(t)} = \{\mu_j^{(t)}, \Sigma_j^{(t)}, \pi_j^{(t)}\}$ , compute estimates of  $w_{ij}$ :

$$w_{ij}^{(t)} = p(z = j | \mathbf{x}_i, \theta^{(t)}) = \frac{\pi_j^{(t)} p(\mathbf{x}_i | z_i = j, \theta^{(t)})}{\sum_{k=1}^n \pi_k^{(t)} p(\mathbf{x}_i | z_i = k, \theta^{(t)})}$$

2. **Maximization:** Using estimates of  $w_{ij}^{(t)}$ , update the estimates of the model parameters

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^N w_{ij}^{(t)} \mathbf{x}_i}{\sum_{i=1}^N w_{ij}^{(t)}} \tag{1}$$

$$\sigma_j^{(t+1)} = \frac{\sum_{i=1}^N w_{ij}^{(t)} \|\mathbf{x}_i - \mu_j\|^2}{\sum_{i=1}^N w_{ij}^{(t)}} \tag{2}$$

$$\pi_i^{(t+1)} = \frac{1}{N} \sum_{i=1}^N w_{ij}^{(t)} \tag{3}$$

Repeat steps 1 and 2 until the change of the parameters is small enough.