K-means clustering

- For each data point x_n, we introduce a corresponding set of ***binary indicator variables*** $r_{nk} \in \{0, 1\}$.
  - where $k = 1, \ldots, K$ describing which of the K clusters the data point x_n is assigned to.
  - so that if data point x_n is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$. This is known as the **1-of-K coding scheme**.

- We can then define an objective function, sometimes called a distortion measure, given by

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

  - which represents the sum of the squares of the distances of each data point to its assigned vector μ_k.

EM algorithm

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

- Our goal is to find values for the $\{r_{nk}\}$ and the $\{\mu_k\}$ so as to minimize J.
  - First we choose some initial values for the μ_k. Then in the first phase we minimize J with respect to the r_{nk}, keeping the μ_k fixed.
  - In the second phase we minimize J with respect to the μ_k, keeping r_{nk} fixed.
  - This two-stage optimization is then repeated until convergence.

- These two stages of updating r_{nk} and updating μ_k correspond respectively to the **E (expectation)** and **M (maximization)** steps of the EM algorithm.

Expectation Step (a.k.a. assignment step)

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

- We can optimize for each n separately by choosing r_{nk} to be *1* for whichever value of *k* gives the minimum value of $||x_n - \mu_k||^2$.
- We simply assign the nth data point to the closest cluster centre.
- More formally, this can be expressed as

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

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

Maximization Step (a.k.a. recompute centroid step)

- Now consider the optimization of the μ_k with the r_{nk} held fixed.
- The objective function J is a quadratic function of μ_k, and it can be minimized by setting its derivative with respect to μ_k to zero giving

$$2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

  - which we can easily solve for μ_k to give

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

* μ_k equal to the mean of all of the data points x_n assigned to cluster k.