CIL Series 5

The Nonstandard Deviations

April 7, 2015

1 Problem 1 (Semi-NMF does K-means clustering)

1. *Proof.* $\mathbf{Z}$ being orthogonal implies $\mathbf{Z}\mathbf{Z}^T = \mathbf{I}$, or $\mathbf{z}_i\mathbf{z}_j^T = [i = j]$ for rows i and j. Assume that there exists a column c with multiple non-zero entries, let z_{ic} and z_{jc} with $i \neq j$ be some of them. Due to the orthogonality constraint we have $\mathbf{z}_i\mathbf{z}_j^T = \sum_{k=1}^{K} z_{ik}z_{jk} = 0$. However, combining this with the non-negativity constraint this only holds if $z_{ik}z_{jk} = 0 \quad \forall k \in \{1, \ldots, K\}$. But we know $z_{ic} \neq 0$ and $z_{jc} \neq 0$, thus $z_{ic}z_{jc} \neq 0$, which is a contradiction. Hence there can be no column with multiple non-zero entries. □

2 Problem 1 (Singularities in Gaussian Mixture Models)

1.

$$p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$\iff \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

2.

$$\ln p(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

3.

$$\begin{aligned}\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &= \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_j)\right) \\ &= \frac{1}{(2\pi)^{\frac{D}{2}} \left|\sigma_j^2 \boldsymbol{I}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{x}_n)^T (\sigma_j^2 \boldsymbol{I})^{-1}(\boldsymbol{x}_n - \boldsymbol{x}_n)\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma_j)^D}\end{aligned}$$

4. $\lim_{\sigma_j \to 0} \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \infty$, the same holds true for the log likelihood.

5. Yes, most definitely. It is enough to have the single cluster center equal to one of the data points. When the variance tends to zero, the normal distribution transforms into the dirac delta function. Given that the likelihood is not normalized, it will also approach infinity even with $K = 1$.

6. The most simple idea would be to avoid variances of zero. This could be predetermined and not changed during the EM-steps.

3 Problem 2 (Identifiability)

1. Since the actual label of the cluster does not matter, there are $K!$ equivalent solutions.

2. Since we usually do not care about the label of the cluster (just the set of data points belonging to a given cluster) this is not a problem.