582744 Advanced Course in Machine Learning

Exercise 3 **Due April 12, 10:00 AM**

Rules:

1. Your submission is composed of two parts: (a) A single PDF file containing your answers to all questions, including both the pen and paper questions and the written answers and various plots for the programming questions. (b) a single compressed file (zip or tar.gz) containing your code (and nothing else). If your code is in a single file, it can be sent also a plain source code.
2. The submission should be sent directly to BOTH Arto (`arto.klami@cs.helsinki.fi`) and Aditya (`aditya.jitta@cs.helsinki.fi`).
3. All material must be renamed to your student ID. Always mention your name and student ID also in the written report.
4. The subject field of your email should be [AML][your student ID][exercise 3].
5. Please typeset your work using appropriate software such as LaTeX. However, there is no need to typeset the pen and paper answers – you can also include a scanned hand-written version.
6. The pen and paper exercises can alternatively be returned in paper form during the Tuesday lecture.

This set of exercises is due on Tuesday April 12th, before 10:00 AM.

1 Plate diagrams and independence (6 points)

Draw the plate diagrams corresponding to the following two factorizations of a joint distribution:

(a)

$$p(\{x_t\}, \{z_t\}, \boldsymbol{\phi}, \boldsymbol{\pi}) = p(\boldsymbol{\phi}|\phi_0)p(\boldsymbol{\pi}|\pi_0)p(z_0)\prod_{t=1}^{3}\Big[p(x_t|z_t, \boldsymbol{\phi})p(z_t|z_{t-1}, \boldsymbol{\pi})\Big],$$

where x_t are observed.

(b)

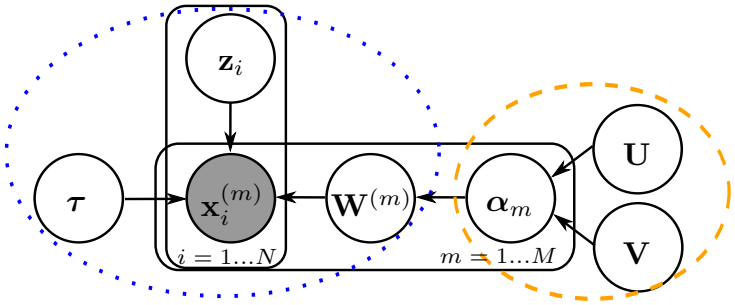
$$p(\{y_n\}, \{\mathbf{x}_n\}, \{z_n\}, \boldsymbol{\pi}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\mathbf{w}_k\}) = p(\boldsymbol{\pi})\prod_{k=1}^{K}\left[p(\boldsymbol{\mu}_k)p(\boldsymbol{\Sigma_k})p(\mathbf{w}_k)\prod_{n=1}^{N}\Big(p(z_n|\boldsymbol{\pi})p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(y_n|\mathbf{x}_n, \mathbf{w}_k)\Big)^{\mathbb{I}(z_n=k)}\right],$$

where both $\mathbf{x}_n$ and y_n are observed.

The curly brackets $\{y_n\}$ denote a collection of N random variables y_n, and $\mathbb{I}(\cdot)$ is the indicator function, obtaining value 1 if the argument is true and 0 otherwise.

Now do the opposite: Write down the joint probability density for the variables depicted by the following plate diagram. You can ignore the blue and orange circles; they were included in the original version of this image (included here with permission) to help understanding the model.

(c)



Can you name these models? (Correct answers are not required for maximum points, so feel free to guess as well)

2 Mixture model for binary data (6 points)

Consider the model

$$\begin{aligned} p(z_n) &= \text{Categorical}(\boldsymbol{\pi}), \\ p(\boldsymbol{\pi}) &= \text{Dirichlet}(\alpha), \\ p(\mathbf{x}_n|z_n = k, \boldsymbol{\mu}) &= \prod_{d=1}^{D}\text{Bernoulli}(\mu_{kd}), \\ p(\mu_{kd}) &= \text{Uniform}(0, 1), \end{aligned}$$

which defines a mixture model for binary vectors $\mathbf{x}_n \in [0, 1]^D$. In verbal terms, the generative process is such that we first pick a cluster index k with probabilities π_k and then generate D independent observations

based on the parameters μ_{kd}. The notation for the Bernoulli distribution means that the dth element of $\mathbf{x}_n$ has probability μ_{kd} to be one if the nth sample belongs to the kth cluster.

Write down the observed data log-likelihood of the model, as well as the complete data log-likelihood. Then derive the expectation maximization algorithm for inferring the parameters $\theta = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\pi}\}$, treating α as a known hyper-parameter.

Do you think this would be a useful clustering model in practice?

3 Spectral clustering (programming, 12 points)

Read in the data set ($N = 120$ samples represented in two dimensions) from http://www.cs.helsinki.fi/u/aklami/teaching/AML/exercise_3_3.csv and compute the pairwise distance matrix d containing all Euclidean distances between the samples (remember to take the square-root). Draw a scatter-plot of the data to see how it looks like. Run also a k-means algorithm with $K = 2$ clusters, using some publicly available code (all programming environments should have one), and color the dots in the scatter-plot according to the cluster indices.

Next you should implement the spectral clustering algorithm:

(a) Create two types of adjancency matrices W based on the data:

- Connect each sample to all other samples that are within distance $e = 0.5$ ($W_{i,j} = 1 \quad \text{if} \quad d_{i,j} \leq e$).
- Connect each sample to its A closest neighbors (not counting the sample itself), using $A = 8$. Do this in a symmetric fashion, so that two nodes are connected if either one of them is within the A closest neighbors of the other one.

Then perform the following steps for both alternatives.

(b) Find the eigenvalues and eigenvectors of the graph Laplacian $L = D - W$, where D is a diagonal matrix with the degrees of the nodes on its diagonal. Plot the eigenvectors corresponding to four smallest eigenvalues (preferably in a single plot) – how do they look like?

Hint: The samples in the data matrix are ordered so that the first 60 samples correspond to one of the natural clusters. Furthermore, the samples within each cluster are ordered along the half-circle. This information should help interpreting the eigenvectors.

(c) Now represent the data using the first $M = 4$ eigenvectors, creating new representation $Y \in \mathbb{R}^{120\times 4}$. Draw a scatter-plot of the first two dimensions of this matrix. How does it look like? Can you see the clusters?

(d) Cluster the data with k-means into two clusters using the new representation Y.

Now compare the three clustering solutions you have: One based on the original data and two based on spectral clustering with different adjacency matrices. What is the difference? Do all three methods solve the clustering problem equally well?

Finally, play around with the numbers e, A and M above to see how things change, answering briefly the following questions. No need to produce separate plots for these, unless you feel it is necessary for understanding your answer.

1. What happens if e is too small? What if it is too big?
2. What happens if A is too small? What it it is too big?
3. Would $M = 2$ be enough? What happens if you use too big M? Why?