Assignment 1: EM for Categorical Data
Advanced Signal Processing

Antonio Artés-Rodríguez
Pablo Moreno-Muñoz

Due: Wednesday December 12, 2018

1 Problem description

Consider the following probabilistic mixture model that can be applied, for example, to a set of documents

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \mathrm{Cat}(x_j|\boldsymbol{\theta}_k),$$

where K is the number of mixture components, D is the number of words that appear in a certain document which is itself a random variable, and $\mathrm{Cat}(x_j|\boldsymbol{\theta}_k)$ is the categorical distribution with I categories.

The aim of this assignment is to develop the Expectation-Maximization (EM) algorithm for this specific mixture model and evaluate it over the data that is provided.

2 Data description

The dataset can be found in the `LDAdata.mat` file, and it contains 2 variables: *LDAdata* and *dictionary*. The former (i.e. *LDAdata*) is a struct that contains 600 rows, each one of them having information about a single abstract that has been downloaded from arXiv[1]. Such information can be accessed for the i-th abstract as follows:

`LDAdata(i).title` extracts a char array that contains the title of the scientific publication.

`LDAdata(i).abstract` extracts a char array that contains the raw abstract.

`LDAdata(i).processed` extracts a cell array that contains the processed abstract. The processing consists on the following steps: 1) Removal of non-alphabetical symbols, e.g. ",",'-' 2) "Lowerization": all characters are set to

[1] arXiv is an on-line repository for scientific papers. Refer to https://arxiv.org/

lowercase, e.g. "Computation"→"computation" 3) "Stemming" extraction of the stem of each word, e.g. "illustrated" →"ilustr" 4) Removal of the so-called stop words, that is, words that don't intrinsically bear any specific meaning and thus are likely to appear in all types of texts, e.g. "the". This pre-processing aims at removing the elements that are not useful for determining the topics that are present in a document, and to aggregate several related terms into a reduced number of entities that can stand out more easily in the model -rather than treating them separately as independent units. The final number of different words is 4,061.

`LDAdata(i).corpus` extracts a two-column matrix whose first column contains an integer number starting at 1 to uniquely represent every word that has appeared in all the processed abstracts. On the other hand, the right column contains the number of times that such a word has appeared on the i-th abstract.

Notice that **only the corpus is needed** for the implementation of the LDA model, the rest of the information is given for completeness.

The latter (i.e.*dictionary*) is a cell array whose i-th entity extracts the word that has been assigned to i when creating `LDAdata(i).corpus`.

3 Work description

1. Write down the expression for the complete data log-likelihood for the mixture of multinomials model

$$l_c(\boldsymbol{\theta}) = \ln p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln(p(\mathbf{x}_i|z_i, \boldsymbol{\theta})p(z_i|\boldsymbol{\theta}))$$

where z_i are the hidden variables and $p(z_i = k) = \pi_k$.

2. Write down the expression for the expected complete data log-likelihood

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = E\{l_c(\boldsymbol{\theta})|\mathcal{D}, \boldsymbol{\theta}^{t-1}\}$$

3. Derive the expression of the ML estimates of the new set of parameters $\boldsymbol{\theta}^t$

$$\boldsymbol{\theta}^t = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$$

4. Implement the EM algorithm for a mixture according to the probability model. The algorithm should take as parameter, at least, the number of components K, the minimum increment in the log-likelihood for convergence, the maximum number of iterations and the parameters of the prior distribution (if needed). Hand in code and a high level explanation of what you algorithm does, including the initialization strategy. **(This part can be done in groups)**.

5. Run your algorithm on the data sets for varying $K = 2, 3, 4, 5$. Verify that the log-likelihood increases at each step of EM. Report the log-likelihood values obtained and display the parameters found. Comment the performances of the algorithm in finding good clusters for the different values of K in comparison with some model selection indicator.

6. Define a conjugate prior for the model parameters and derive the MAP estimates of the new set of parameters $\boldsymbol{\theta}^t$.

7. Modify the implementation of the EM algorithm for MAP estimation, including as an additional input the prior's parameters, and checking the convergence using the posteriors instead of the likelihoods. **(This part can be done in groups)**

8. Run your algorithm on the data sets for the values of K you consider "optimum" in the EM-ML and varying the values of the prior distribution. You must use, at least, flat and non-informative (Jeffreys) priors. Comment the results you obtain and compare them with results obtained using the EM-ML.