Assignment 2: HMM for Categorical Data Sequences
Advanced Signal Processing

Antonio Artés
Pablo Moreno-Muñoz

Due: Monday Jan 29, 2019

1 Problem description

Consider a HMM with categorical observations that can be applied, for example, to sequences of documents,

$$p(\mathbf{y}_t|s_t = k, \mathbf{B}) = \prod_{j=1}^{D_t} \text{Cat}(y_{tj}|\mathbf{b}_k)$$

where D_t is the number of words in the document $\mathbf{y}_t$, and $\text{Cat}(y_{tj}|\mathbf{b}_k)$ is the categorical distribution with I categories.

The aim of this assignment is to develop a Baum-Welch algorithm for training HMMs for this type of observations, and to evaluate it in one set of data.

2 Data description

The first data set is in the file `observed.mat` and it contains 10 sequences generated by an HMM using a corpus of real documents. Each document is represented by 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 documents, and the second column contains the number of times that such a word has appeared on the document.

3 Work description

1. Write down the expression for the complete data log likelihood for N sequences

$$\log p(S, Y|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} \left(p(s_1^n|\boldsymbol{\pi}) \prod_{t=2}^{T_n} p(s_t^n|s_{t-1}^n, \mathbf{A}) \right) \left(\prod_{t=1}^{T_n} p(\mathbf{y}_t^n|s_t^n, \mathbf{B}) \right)$$

where $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$ are the model parameters.

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 $\boldsymbol{\theta}^t$

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

4. Implement the Baum-Welch algorithm for this HMM using MATLAB or PYTHON code. The algorithm should take as input, at least, the number of hidden states, I, a cell of cells containing the data set, the minimum increment in the log likelihood for convergence, and the maximum number of iterations. Hand in code and a high level explanation of what you algorithm does. **(This part can be done in groups)**

5. Implement a state-by-state MAP decoder based on the Forward-Backward algorithm and a ML sequence decoder based on the Viterbi algorithm. **(This part can be done in groups)**

6. Run your algorithm on the data set for varying $I = 2, 3, 4, 5$. Verify that the log likelihood increases at each step of EM. Report the log likelihoods obtained and display the parameters and the hidden sequences found by the state-by-state MAP decoder and the ML sequence decoder. Comment the performances of the algorithms for finding the hidden sequences.