Machine Learning Fundamentals
Final Exam
2021-2022

Massih-Reza Amini

Duration: 2 hours, authorized documents: lecture slides

We are interested in clustering techniques for document collections using mixture models. The learning algorithm used is the EM algorithm. First, we examine a basic method, then we examine a more sophisticated method that creates clusters of documents while associating a level of *abstraction* to the words of the documents. To derive the formulas for re-estimating the parameters one will need to solve minimization problems under equality constraints – the method to be used for this is indicated in the appendix at the end of the exam text.

Question 1. Explain the differences between classification and clustering.

Question 2. What is a mixture model?

We consider a set of documents D, $d \in D$ will designate a document, $d = w_1 w_2 \dots w_{l(d)}$ where w_i is the i^{th} word of d and $l(d)$ is the length in words of d. $\mathbb{P}(d)$ denotes the probability of d and $\mathbb{P}(w)$ the word density. V denotes the vocabulary, i.e. the set of all words that appear in the corpus. We will use a representation of documents in the form of a *bag of words*: a document is assimilated to all of its words, which are assumed to be statistically independent, i.e.

$$\mathbb{P}(d) = \prod_{j=1}^{l(d)} \mathbb{P}(w_j).$$

Part I. For our basic clustering, we want to partition documents into q groups, a given document belongs to exactly one cluster, we consider a mixture model with q components $\mathbb{P}(d) = \sum_{k=1}^{q} \mathbb{P}(k)\mathbb{P}(d \mid k)$, where $\mathbb{P}(k)$ is the prior probability of cluster k and $\mathbb{P}(d \mid k)$ is the density of component k. The likelihood of D is $\mathbb{P}(D, \Theta) = \prod_{d \in D} \mathbb{P}(d)$, the log-likelihood of D with the mixture model is

$$\mathcal{L}(D, \Theta) = \sum_{d \in D} \log \left(\sum_{k=1}^{q} \mathbb{P}(k)\mathbb{P}(d \mid k) \right)$$

where Θ denotes the parameters of the mixture namely $\mathbb{P}(k)$ and $\mathbb{P}(d \mid k)$, the identifiers of the clusters k are here the hidden variables. We recall that the EM algorithm is an iterative algorithm which maximizes the expectation – with respect to the hidden variables, and for fixed a posterior probabilities $\mathbb{P}(k \mid d, \Theta^{(t)})$ at iteration t – of the likelihood of the complete data. These are the $(d, k)_{1 \leq d \leq D; 1 \leq k \leq q}$ with k the cluster of document d. For our mixture model, this expectation writes:

$$Q(\Theta, \Theta^{(t)}) = \sum_{d \in D} \sum_{k=1}^{q} \log(\mathbb{P}(k)\mathbb{P}(d \mid k)).\mathbb{P}(k \mid d, \Theta^{(t)})$$

where $\Theta^{(t)}$ indicates that the $\mathbb{P}(k \mid d, \Theta^{(t)})$ which were estimated in step E, are considered as constants when maximizing Q (step M).

Question 3. Present briefly the EM algorithm in this case.

Question 4. Explain why it is difficult to directly maximize this likelihood?

Question 5. Step E: show that

$$\mathbb{P}(k \mid d, \Theta^{(t)}) = \frac{\prod_{w \in V} \mathbb{P}(w \mid k)^{n(d,w)} \mathbb{P}(k)}{\sum_{k'} \prod_{w \in V} \mathbb{P}(w \mid k')^{n(d,w)} \mathbb{P}(k')}$$

where $n(d, w)$ is the number of times w is present in the document d.

Question 6. Step M: Let $\theta_{w,k} = \mathbb{P}(w \mid k)$ and $\theta_k = \mathbb{P}(k)$, these two quantities will be the parameters of our mixture model that we will have to estimate. They are subject to the following constraints:

$$\sum_{w \in V} \theta_{w,k} = 1, \forall k \in \{1, \ldots, q\}; \text{ and } \sum_{k=1}^{q} \theta_k = 1$$

Express $Q(\Theta, \Theta^{(t)})$ with respect to $\theta_{w,k}$ and θ_k.

Question 7. Write the associated Lagrangian and give the formulas for re-estimating these parameters for step M. For this, we will solve the following constrained optimization problems:

- Maximize $Q(\Theta, \Theta^{(t)})$ with respect to $\theta_{w,k}$ under the constraints $\sum_{w \in V} \theta_{w,k} = 1$, $\forall k \in \{1, \ldots, q\}$;
- Maximize $Q(\Theta, \Theta^{(t)})$ with respect to θ_k under the constraint $\sum_{k=1}^{q} \theta_k = 1$;

Part II. We will now perform a hierarchical clustering by associating each word w of a document d with a hidden variable a which represents its level of "generality" or "abstraction". We suppose given a binary tree which structures the levels of abstraction of the words. Each internal node of this tree will be associated with a value of the hidden variable "level of abstraction" a. The clusters where the documents will be placed will correspond to the leaves of the tree (see figure). These variables a play a similar role with the words of the documents to that of the variables k for the documents of the clusters. We denote by r the number of values that these variables a can take. The learning will assign each document d to a cluster k and each word of the document to a node a of the word hierarchy. Thus, to the document $d = w_1 w_2 \dots w_{l(d)}$, will correspond the sequence of word labels $A = a_1 a_2 \dots a_{l(d)}$,

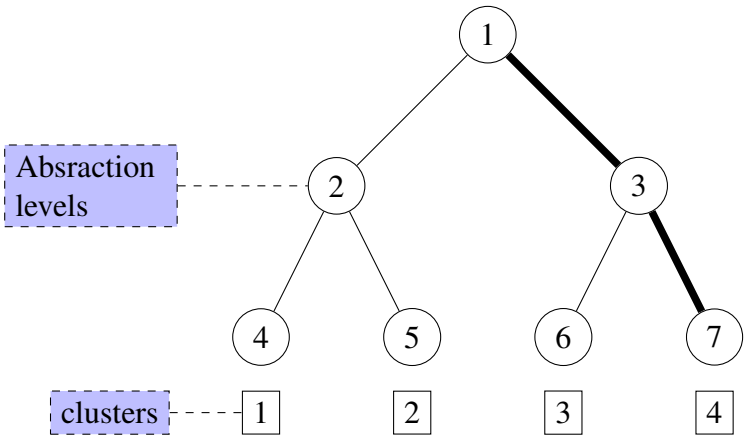

Figure 1: 3 levels of abstraction, $q = 4, r = 7$, squares represent document clusters, circles represent the abstraction variables. A document classified in cluster 3 may have a word w_d, whose abstraction variable is $a_d = 3$. If a node of the tree, a, is not a parent of cluster k then $\mathbb{P}(a \mid k) = 0$. Words in documents in cluster 3 can only be indexed by abstraction variables 1, 3, or 6 (bold path) which are on the path between the root and the cluster. The most specific words of cluster 3 will have the label 6, the words common to all the clusters will have the label 1. When the clustering is finished, the most frequent words of the different nodes offer summaries at different levels of resolution of the clusters which are below.

We introduce the following notations and constraints for the parameters of this model:

$\theta_{w,a} = \mathbb{P}(w \mid a)$, which verify the constraint: $\sum_{w' \in V} \theta_{w',a} = 1; \forall a = 1 \dots r$

$\theta_{a,k} = \mathbb{P}(a \mid k)$, which verify the constraint: $\sum_{a'=1}^{r} \theta_{a',k} = 1; \forall k = 1 \dots q$

$\theta_k = \mathbb{P}(k)$, which verify the constraint $\sum_{k=1}^{q} \theta_k = 1$.

We consider a mixture model with the hidden variables that are; the identifiers of the clusters and the levels of abstraction of the words in the tree. We suppose that:

- (H1): $\mathbb{P}(d \mid k, a_j) = \mathbb{P}(d \mid a_j) = \mathbb{P}(w_j \mid a_j); \forall j \in \{1, \dots, l(d)\}$
- (H2): $\mathbb{P}(d \mid k) = \sum_{j=1}^{l(d)} \sum_{a_j=1}^{r} p(w_j \mid a_j) p(a_j \mid k)$
- (H3): The expectation of the likelihood of the complete data simplifies to the following form:

$$Q(\Theta, \Theta^{(t)}) = \sum_{d \in D} \sum_{k=1}^{q} \sum_{j=1}^{l(d)} \sum_{a_j=1}^{r} \left[\log(\theta_{w_j,a_j} \theta_{a_j,k} \theta_k) \mathbb{P}(k, a_j \mid d, \Theta^{(t)}) \right]$$

Question 8. Step E: Give the expression of conditionnal probabilities $\mathbb{P}(k \mid d, \Theta^{(t)})$ and $\mathbb{P}(a_j \mid k, d, \Theta^{(t)})$ with respect to the parameters of the model: $\theta_{w,a}$, $\theta_{a,k}$ and θ_k.

Question 9. Step M: Give the Lagrangian associated to $Q(\Theta, \Theta^{(t)})$.

Question 10. Derive the re-estimate formulas for the parameters $\theta_{w,a}$ and $\theta_{a,k}$.

Appendix

Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be a real vector and A, B, C functions of $\boldsymbol{x}$. We consider the following constrained optimization problem:

$$\text{Maximize } A(\boldsymbol{x})$$
$$\text{Under the constraints } B(\boldsymbol{x}) = 1 \text{ and } C(\boldsymbol{x}) = 1$$

We define the Lagrangian associated with this problem:

$$L(\boldsymbol{x}, \lambda_B, \lambda_C) = A(\boldsymbol{x}) - \lambda_B(B(\boldsymbol{x}) - 1) - \lambda_C(C(\boldsymbol{x}) - 1)$$

A necessary condition (we will consider that it is sufficient here) for $\boldsymbol{x}$ to be a solution to the maximization problem is: $\frac{\partial L(\boldsymbol{x}, \lambda_B, \lambda_C)}{\partial x_i} = 0, \forall i = 1, \dots, n$.