Dirichlet-Multinomial Mixture Model and EM algorithm

Saki Kuzushima

August 17, 2019

This document introduces the Dirichlet-Multinomial Mixture Model and its inference using EM algorithm. The Dirichlet-Multinomial Mixture Model is often used in text analysis, so I use an example of document clustering. The supervised version of this model was introduced in [1].

1 Notation

1.1 Data

- $D = \{d_i\}_{i=1}^{|D|}$: a set of documents
- $d_i = [N_{i1}...N_{i|V|}]$: a document is a vector of the number of each unique word appeared in the document
- N_{it}: the number of the unique word t appeared in the document i
- $V = \{W_t\}_{t=1}^{|V|}$: a set of vocabulary. W_t is the tth unique word.

1.2 Latent parameters

- Z_i: $Z_i = 1$ if the document i has positive class; $Z_i = 0$ otherwise.

1.3 Parameters

- π: the probability of a document having the positive class
- $\eta = \{\eta_{1,t}, \eta_{2,t}\}_{t=1}^{|V|}$: the probability of a word t being drawn out of the set of vocabulary, to write a document with class 1 or 2. $\sum_{t=1}^{|V|} \eta_{tj} = 1, j = 1, 2$

[1]

[1] I use $i = 1...|D|$ for document index, $j = 1, 2$ for class index, and $t = 1...|V|$ for unique-word index.

2 Unsupervised Dirichlet Mixture Model

2.1 Model

We assume that there are two groups, or clusters in the documents.

2.1.1 Prior

$$\pi \sim Beta(2, 2) \tag{1}$$

$$Z_i \overset{i.i.d}{\sim} Bernoulli(\pi), \quad \text{for } i \text{ in } 1..|D| \tag{2}$$

2.1.2 Likelihood

$$\eta_j \overset{i.i.d}{\sim} Dirichlet(2, ..., 2), \quad j = 1, 2 \tag{3}$$

$$d_i | Z_i = 1 \sim Multinomial(\eta_1), \quad \text{for } i \text{ in } 1...|D| \tag{4}$$

$$d_i | Z_i = 0 \sim Multinomial(\eta_2), \quad \text{for } i \text{ in } 1...|D| \tag{5}$$

We used $Beta(2, 2)$ and $Dirichlet(2...2)$ because this results in the equivalence to the Laplace smoothing in the M step.

2.2 Maximand (Marginal Posterior)

To be updated.

2.3 E step

First we write down the joint posterior density of the parameters η, π and the latent variable Z, given data D.

2.3.1 Joint Posterior

$$p(Z, \eta, \pi | D) \propto p(D|Z, \eta)p(Z|\pi)p(\eta)p(\pi) \tag{6}$$

Decomposing the RHS,

$$p(D|Z,\eta) = \prod_{i=1}^{|D|} p(d_i|Z_i,\eta) \tag{7}$$

$$= \prod_{i=1}^{|D|} \{p(d_i|Z_i=1,\eta)^{Z_i} p(d_i|Z_i=0,\eta)^{1-Z_i}\} \tag{8}$$

$$= \prod_{i=1}^{|D|} \{\prod_{t=1}^{|V|} \eta_{1t}^{N_{it}Z_i} \times \prod_{t=1}^{|V|} \eta_{2t}^{N_{it}(1-Z_i)}\} \tag{9}$$

$$= \prod_{t=1}^{|V|} \eta_{1t}^{\sum_{i=1}^{|D|} N_{it}Z_i} \times \prod_{t=1}^{|V|} \eta_{2t}^{\sum_{i=1}^{|D|} N_{it}(1-Z_i)} \tag{10}$$

$$p(Z|\pi) = \prod_{i=1}^{|D|} p(Z_i|\pi) \tag{11}$$

$$= \prod_{i=1}^{|D|} \{p(Z_i=1|\pi)^{Z_i} p(Z_i=0|\pi)^{1-Z_i}\} \tag{12}$$

$$= \prod_{i=1}^{|D|} \{\pi^{Z_i}(1-\pi)^{1-Z_i}\} \tag{13}$$

$$= \pi^{\sum_{i=1}^{|D|} Z_i}(1-\pi)^{\sum_{i=1}^{|D|}(1-Z_i)} \tag{14}$$

$$p(\eta) = \eta_1 \times \eta_2 \tag{15}$$

$$= \prod_{t=1}^{|V|} \eta_{1t} \times \prod_{t=1}^{|V|} \eta_{2t} \tag{16}$$

$$p(\pi) = \pi \times (1-\pi) \tag{17}$$

Combining all terms,

$$p(Z,\eta,\pi|D) \propto p(D|Z,\eta)p(Z|\pi)p(\eta)p(\pi) \tag{18}$$

$$= \prod_{t=1}^{|V|} \eta_{1t}^{\sum_{i=1}^{|D|} N_{it}Z_i} \times \prod_{t=1}^{|V|} \eta_{2t}^{\sum_{i=1}^{|D|} N_{it}(1-Z_i)} \times \pi^{\sum_{i=1}^{|D|} Z_i}(1-\pi)^{\sum_{i=1}^{|D|}(1-Z_i)} \tag{19}$$

$$\times \prod_{t=1}^{|V|} \eta_{1t} \times \prod_{t=1}^{|V|} \eta_{2t} \times \pi \times (1-\pi) \tag{20}$$

$$= \eta_{1t}^{\sum_{t=1}^{|V|}(1+\sum_{i=1}^{|D|} N_{it}Z_i)} \times \eta_{2t}^{\sum_{t=1}^{|V|}(1+\sum_{i=1}^{|D|} N_{it}(1-Z_i))} \tag{21}$$

$$\times \pi^{1+\sum_{i=1}^{|D|} Z_i} \times (1-\pi)^{1+\sum_{i=1}^{|D|} Z_i} \tag{22}$$

Taking log to get log joint posterior distribution,

$$\log p(Z, \eta, \pi | D) \tag{23}$$

$$= \sum_{t=1}^{|V|} (1 + \sum_{i=1}^{|D|} N_{it} Z_i) \log \eta_{1t} + \sum_{t=1}^{|V|} (1 + \sum_{i=1}^{|D|} N_{it} (1 - Z_i)) \log \eta_{2t} \tag{24}$$

$$+ (1 + \sum_{i=1}^{|D|} Z_i) \log \pi + (1 + \sum_{i=1}^{|D|} (1 - Z_i)) \log(1 - \pi) + constant \tag{25}$$

2.3.2 Expectation over Z

We take the expectation over Z, given the old parameters, η^{old}, π_{old}, to find the Q function. Let $p_i = P(Z_i = 1 | \eta^{old}, \pi^{old}, D)$.

$$Q \equiv \mathbb{E}[\log p(Z, \eta, \pi | D)] \tag{26}$$

$$= \sum_{t=1}^{|V|} (1 + \sum_{i=1}^{|D|} N_{it} p_i) \log \eta_{1t} + \sum_{t=1}^{|V|} (1 + \sum_{i=1}^{|D|} N_{it} (1 - p_i)) \log \eta_{2t} \tag{27}$$

$$+ (1 + \sum_{i=1}^{|D|} p_i) \log \pi + (1 + \sum_{i=1}^{|D|} (1 - p_i)) \log(1 - \pi) + constant \tag{28}$$

2.4 M step

We maximize the Q function w.r.t π and η. Taking derivative w.r.t π,

$$\frac{\partial Q}{\partial \pi} = \frac{1 + \sum_{i=1}^{|D|} p_i}{\pi} - \frac{1 + \sum_{i=1}^{|D|} (1 - p_i)}{1 - \pi} = 0 \tag{29}$$

Solving this w.r.t π,

$$\pi^* = \frac{1 + \sum_{i=1}^{|D|} p_i}{2 + |D|} \tag{30}$$

Because $\sum_{t=1}^{|D|} \eta_{1t} = 1$, form a Lagrange,

$$L = Q - \lambda (\sum_{t=1}^{|D|} \eta_{1t} - 1) \tag{31}$$

Taking derivative w.r.t η_{1t}

$$\frac{\partial L}{\partial \eta_{1t}} = \frac{1 + \sum_{i=1}^{|D|} N_{it} p_i}{\eta_{1t}} - \lambda = 0 \tag{32}$$

$$\iff \eta_{1t} = \frac{1 + \sum_{i=1}^{|D|} N_{it} p_i}{\lambda} \tag{33}$$

Taking a derivative w.r.t λ

$$\frac{\partial Q}{\partial \lambda} = \sum_{t=1}^{|V|} \eta_{1t} - 1 = 0 \tag{34}$$

Solving them w.r.t. η_{1t},

$$\eta_{1t}^{*} = \frac{1 + \sum_{i=1}^{|D|} N_{it} p_i}{|V| + \sum_{t=1}^{|V|} \sum_{i=1}^{|D|} N_{it} p_i} \tag{35}$$

Likewise,

$$\eta_{2t}^{*} = \frac{1 + \sum_{i=1}^{|D|} N_{it} (1 - p_i)}{|V| + \sum_{t=1}^{|V|} \sum_{i=1}^{|D|} N_{it} (1 - p_i)} \tag{36}$$

References

[1] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.