Homework 7

Applied Machine Learning–CS 498

Mengyu Xie, Adrian Bandolon

April 9, 2018

Problem 1–EM Topic Model

- We used the NIPS dataset found at **https://archive.ics.uci.edu/ml/datasets/Bag+of+Words**
- This dataset is composed of two sets: a table of word counts per document and a vocabulary list.
- We clustered to 30 topics using a simple mixture of multinomial topic model.
- The EM-Topic Model has two steps:

1. The E-step:
   - The initial probabilities (0.0 to 1.0) of π_j and p_{jk} were randomly generated (`numpy.random.uniform()`). This helped simplify initialization.
   - To handle underflow issues we used the `expsumlog` trick (see code) as suggested in Piazza **(https://piazza.com/class/jchzguhsowz6n9?cid=1121)**.

$$p(\delta_{ij} = 1|\theta^{(n)}, x) = \frac{\left[\prod_k p_{j,k}^{x_{ik}}\right]\pi_j}{\sum_l \left[\prod_k p_{l,k}^{x_{ik}}\right]\pi_l}$$

2. The M-Step:
   - We added a very small number ($\epsilon = 0.00001$) for smoothing to avoid zero probability words.
   - The parameters we are trying to estimate are:

$$p_j^{n+1} = \frac{\sum_i x_i w_{ij}}{\sum_i x_i^T 1 w_{ij}} \textbf{ and } \pi_j^{n+1} = \frac{\sum_i w_{ij}}{N}$$

- Convergence was evaluated using $|\textbf{mean } \pi_{\mathbf{j}}^{(\mathbf{n+1})} - \textbf{mean } \pi_{\mathbf{j}}^{\mathbf{current}}| < \mathbf{0.0001}$.
- We set a limit of 1000 iterations, in case the convergence criteria above was not met.

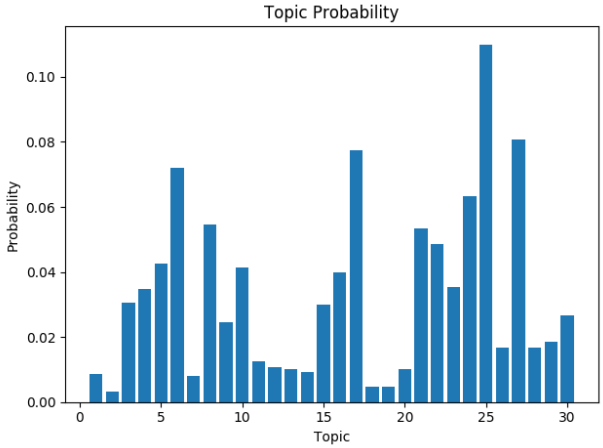

Figure 1: *Probability with which a topic is selected. Sum of the probabilities equal to* 1.0000

Table 1. *Top ten words (columns) with the highest probability for a topic (rows).*

	Top.1.Word	Top.2.Word	Top.3.Word	Top.4.Word	Top.5.Word	Top.6.Word	Top.7.Word	Top.8.Word	Top.9.Word	Top.10.Word
1	input	direction	network	unit	learning	motion	set	neural	weight	function
2	evidence	speaker	approximation	set	network	condition	system	data	module	states
3	network	algorithm	function	unit	learning	output	input	model	set	weight
4	network	model	function	learning	input	neuron	algorithm	system	pattern	set
5	network	learning	algorithm	system	function	input	weight	neural	set	training
6	model	network	function	learning	data	neural	unit	set	system	training
7	function	network	number	result	method	bound	threshold	neural	unit	problem
8	network	model	neuron	input	function	neural	cell	system	learning	set
9	model	data	algorithm	function	set	parameter	point	problem	result	learning
10	model	function	network	algorithm	data	learning	set	input	training	error
11	network	output	recognition	system	neural	set	model	context	distribution	word
12	cell	model	network	field	synaptic	firing	david	word	input	rat
13	learning	network	model	algorithm	mean	input	parameter	function	data	gaussian
14	learning	algorithm	function	input	model	set	training	unit	pattern	number
15	network	model	training	input	system	learning	neural	set	data	output
16	data	model	learning	network	function	set	input	neural	algorithm	distribution
17	model	network	function	learning	input	data	algorithm	neuron	cell	set
18	network	weight	algorithm	layer	node	input	nodes	output	training	neural
19	function	input	threshold	response	visual	correlation	map	spike	component	contour
20	network	student	learning	input	neural	teacher	system	function	error	vector
21	network	model	function	algorithm	learning	system	unit	set	input	weight
22	model	network	data	learning	algorithm	function	input	set	neural	system
23	network	learning	model	neural	input	algorithm	unit	function	output	system
24	network	input	model	learning	neural	neuron	function	weight	output	pattern
25	network	neural	set	model	learning	error	training	function	weight	data
26	network	system	training	classifier	set	neural	recognition	algorithm	output	model
27	network	model	learning	input	unit	function	system	neural	output	training
28	network	learning	training	system	set	neural	word	unit	result	object
29	learning	network	data	algorithm	function	field	problem	system	set	method
30	network	model	learning	system	input	unit	neural	problem	pattern	algorithm

Problem 2–Image Segmentation

- We were provided 3 test images for this exercise.

Part 1:

- Here we segment the test images (`RobertMixed03.jpg`, `smallsunset.jpg` and `smallstrelitzia.jpg`) to 10, 20 and 50 segments.

1. For the E-step we computed:
   - Initial values for μ and π were obtained using the *K-means Algorithm* as implemented in `sklearn.cluster.KMeans()`.

$$p(\delta_{ij} = 1|\theta^{(n),x}) = \frac{\Big[exp\big(-\frac{1}{2}(x_i - \mu_j)^T(x_i - \mu_j)\big)\Big]\pi_j}{\sum_k \Big[exp\big(-\frac{1}{2}(x_i - \mu_j)^T(x_i - \mu_j)\big)\Big]\pi_k}$$

2. The values for μ and π were updated at the M-step using:

$$\mu_j^{n+1} = \frac{\sum_i x_i w_{ij}}{\sum_i w_{ij}} \textbf{ and } \pi_j^{n+1} = \frac{\sum_i w_{ij}}{N}$$

- We used the convergence criteria of $|\mu^{\mathbf{n-1}} - \mu^{\mathbf{current}}| < \mathbf{0.0001}$.
- We set a limit of 1000 iterations in case the convergence criteria above was not met.
- After convergence, original pixels were replaced with the mean color of the closest segment .

Figure 1: Robert Original Image

Figure 2: Strelitzia Original Image

Figure 3: Sunset Original Image

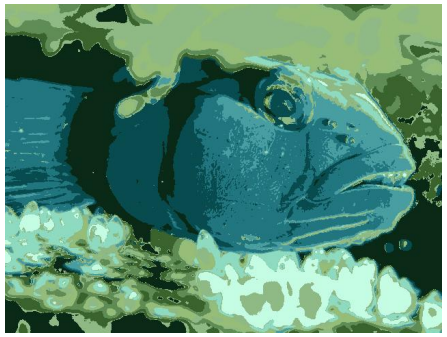

Figure 4: Robert Mixed 10 Segments

Figure 5: Robert Mixed 20 Segments

Figure 6: Robert Mixed 50 Segments

Figure 7: Strelitzia 10 Segments

Figure 8: Strelitzia 20 Segments

Figure 9: Strelitzia 50 Segments

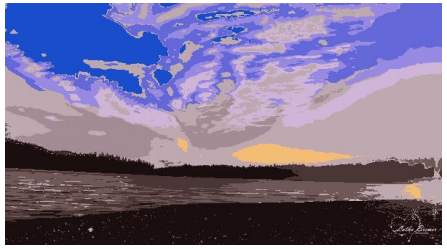

Figure 10: Sunset 10 Segments

Figure 11: Sunset 20 Segments

Figure 12: Sunset 50 Segments

Part 2:

- The 'smallsunset.jpg' image was used in this exercise. It was segmented to 20 segments using different start points.
- There is no significant variation between the images which can be attributed to the robustness of this algorithm. There are however, minor differences (e.g. Figures 14, 15, and 16 look similar to each other, and Figures 17 and 18 look similar to each other) which could be due to the sensitivity of the EM algorithm to the starting and being only able to achieve a local mininum (versus a global minimum).

Figure 13: Sunset Original Image

Figure 14: Start-Point 1-20 Segments

Figure 15: Start-Point 2-20 Segments

Figure 16: Start-Point 3-20 Segments

Figure 17: Start-Point 4-20 Segments

Figure 18: Start-Point 5-20 Segments