Homework 3: LDA & Gibbs sampling

TDA231 - Algorithms for Machine Learning & Inference

Theodor Åstrand, theast@student.chalmers.se, 931109-9114
Linnéa Otterlind, linott@student.chalmers.se, 921126-3620

Chalmers University of Technology

February, 2016

Problem 1.1

(a)

The state space can assume the values from the set $\mathbb{R}^2$ since the distribution is a 2D Gaussian. Although, a vast majority of the values will be distributed around the mean.

(b)

$\mu = [\mu_1; \mu_2] = [1; 1]$

$\Sigma = [\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22}] = [1, -0.5; -0.5, 1]$

$\Lambda = \Sigma^{-1} = [\Lambda_{11}, \Lambda_{12}; \Lambda_{21}, \Lambda_{22}] = [1.33, 0.66; 0.66, 1.33]$

$p(x_1|x_2) = \mathcal{N}(\alpha|\mu_{1|2}, \Sigma_{1|2})$

$\mu_{1|2} = \mu_1 + \Sigma_{12} \cdot \Sigma_{22}^{-1} \cdot (x_2 - \mu_2) = 1 + -0.5 \cdot 1^{-1} \cdot (x_2 - 1) = \frac{3-x_2}{2}$

$\Sigma_{1|2} = \Lambda_{11}^{-1} = 1.33^{-1} = \frac{3}{4}$

The transition rule is thus:

$p(x_1|x_2) = \mathcal{N}(\alpha|\frac{3-x_2}{2}, \frac{3}{4})$

(c)

$\mu = [\mu_1; \mu_2] = [1; 1]$

$\Sigma = [\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22}] = [1, -0.5; -0.5, 1]$

The marginal $p(x_1)$ is a 1D Gaussian, obtained by projecting the joint distribution on the x_1 line:

$p(x_1) = \mathcal{N}(x_1|\mu_1, \Sigma_{11}^2)$

Since x_1 is dependent on x_2 we must obtain a sample of x_2. If we assume that we have observed x_2 the conditional $p(x_1|x_2)$ is given by:

$p(x_1|x_2) = \mathcal{N}(x_1|\mu_1 + \frac{\Sigma_{12}}{\Sigma_{22}^2}(x_2 - \mu_2), \Sigma_{11}^2 - \frac{\Sigma_{12}^2}{\Sigma_{22}^2})$

Problem 2.2

(a)

We want to calculate the probability of the topic given the words, which becomes a formula with an intractable sum in the denominator. If we disregard this sum we get the probability of a word given the topics times the probability of a topic, this is proportional to the probability we wanted to compute to begin with. Using LDA joint probability and then assuming symmetric Dirichlets (α, η uniform) this becomes the formula given in the assignment.

(b)

β represents the probability that a topic contains a specific word.

θ represents the distribution of topics in a document.

(c)

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
classifier	model	model	learning	unit
model	data	neuron	network	network
cell	recognition	system	function	input
object	system	signal	error	layer
function	component	control	weight	model
training	speech	dynamic	problem	neuron
set	network	network	algorithm	visual

Table 1: Most common words for each topic, sorted in descending order.

From the data we computed which is presented in Table 1 we can see that it is not obvious if there are any distinct topics. A few words are shared between different topics although they do not have the same probability in each topic. We can see that topic 4 and 5 could for example be *neural networks*, but it is not very clear. Topic 3 could be *signal processing*, topic 2 could be *neuroscience* and topic 1 could be *probability*.

(d)

Topic 1	Topic 2	Topic 3
input	model	network
unit	neuron	learning
network	cell	error
learning	network	data
function	system	algorithm
system	input	set
output	object	model

Table 2: Most common words for each topic, sorted in descending order.

#	Topic 1	Topic 2	Topic 3
Document 20	**0.4909**	0.1132	0.3959
Document 40	**0.5311**	0.0503	0.4186
Document 60	0.1602	0.0354	**0.8043**
Document 80	0.1825	0.0693	**0.7482**
Document 100	0.1897	0.0494	**0.7609**

Table 3: Probability distribution of topics in specific documents.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
system	model	neuron	model	system	network	error
network	component	cell	unit	circuit	learning	learning
classifier	dynamic	input	object	signal	unit	set
neural	distribution	model	network	frequency	weight	data
training	probability	visual	hidden	output	neural	vector
input	parameter	network	control	input	order	training
performance	input	firing	movement	word	activation	function

Table 4: Most common words for each topic, sorted in descending order.

#	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Document 20	0.0744	0.2633	0.0468	0.0169	0.0205	0.0868	**0.4913**
Document 40	**0.4254**	0.2165	0.0139	0.0331	0.0562	0.0577	0.1973
Document 60	0.1010	0.0594	0.0203	0.0236	0.0165	0.0342	**0.7451**
Document 80	0.2402	0.2249	0.0226	0.0353	0.0226	0.1664	**0.2881**
Document 100	0.1393	**0.4844**	0.0118	0.0184	0.0307	0.0347	0.2806

Table 5: Probability distribution of topics in specific documents.

Looking at Table 2 and Table 4 we can see that more topics seem to result in more variation in words and also a bit more distinction between topics, since they do not share as many words between them. When running with fewer topics we can see that words that are very common often show up in multiple topics.

If we look at Table 3 and Table 5 we can see that less topics seem to give higher probability to a specific topic in comparison than running with more topics.

(e)

Because finding the exact inference algorithm involves computing a probability distribution over a large discrete state space, i.e. it is intractable.

(f)

Because we can approximate a solution by sampling instead. Sampling based inference is exact in the limit of infinite samples.