CUED - Engineering Tripos Part IIB 2019-2020 **Module Coursework**

Module	4F13	Title of report	**Latent Dirichlet Allocation**

Date submitted: **04/12/2020**	Assessment for this module is ☑ **100%** / ☐ 25% coursework of which this assignment forms **33** %

UNDERGRADUATE STUDENTS ONLY		**POST GRADUATE STUDENTS ONLY**			
Candidate number:	**5562E**	Name:		College:	

Feedback to the student

☐ **See also comments in the text**

		Very good	**Good**	Needs improvmt
CONTENT	**Completeness, quantity of content:** Has the report covered all aspects of the lab? Has the analysis been carried out thoroughly?			
	Correctness, quality of content Is the data correct? Is the analysis of the data correct? Are the conclusions correct?			
	Depth of understanding, quality of discussion Does the report show a good technical understanding? Have all the relevant conclusions been drawn?			
	Comments:			
PRESENTATION	**Attention to detail, typesetting and typographical errors** Is the report free of typographical errors? Are the figures/tables/references presented professionally?			
	Comments:			

Overall assessment (circle grade)		A*	**A**	B	C	D
Guideline standard		>75%	**65-75%**	55-65%	40-55%	<40%
Penalty for lateness:		*20% of marks per week or part week that the work is late.*				

Marker: Date:

4F13 Probabilistic Machine Learning - Latent Dirichlet Allocation

Candidate: 5562E

November 26, 2020

Contents

1 Introduction

We have a document test set $\mathcal{A}$, which consists of D documents indexed by $d \in \{1 \dots D\}$. A document is an ordered list of words. The vocabulary $\mathcal{M}$ has $M = |\mathcal{M}|$ unique words. We denote the n'th word in the d'th document by $w_{nd} \in \{1 \dots N_d\}$. Where N_d is the length of document d. For simplicity we denote the count of occurences of word m in the test set by c_m.

We hold back a test set $\mathcal{B}$ to calculate the performance of our approaches.

2 Questions

a Maximum Likelihood

We begin by assuming that each word is drawn independently from a categorical distribution with parameter β: $w_{nd} \overset{iid}{\sim} Cat(\beta)$. In this case β is a $M \times 1$ vector with the conditions that $\sum_{m=1}^{M} \beta_m = 1$ and $\beta_i \geq 0$. The likelihood of the parameter β is the probability of the dataset given β:

$$L(\beta) = P(\mathcal{A}|\beta) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} P(w_{nd}|\beta) = \prod_{m \in \mathcal{M}} \beta_m^{c_m} \tag{1}$$

Where c_m is the count of word m in the training set. We wish to obtain the Maximum-Likelihood estimate $\hat{\beta}^{ML} = \arg\max L(\beta)$. We prefer to maximise the log-likelihood as this is more tractable:

$$\mathcal{L}(\beta) = \log L(\beta) = \sum_{m \in \mathcal{M}} c_m \log \beta_m \tag{2}$$

We can now take derivatives and include a Lagrange multiplier to respect the sum to 1 constraint:

$$\frac{\partial}{\partial \beta_i} \left\{ \mathcal{L}(\beta) + \lambda \left(1 - \sum_{m=1}^{M} \beta_m \right) \right\}_{\beta = \hat{\beta}^{ML}} = \frac{c_i}{\hat{\beta}_i^{ML}} - \lambda = 0$$

$$\therefore \hat{\beta}_i^{ML} = \frac{c_i}{\lambda} = \frac{c_i}{\sum_{m \in \mathcal{M}} c_m} = \frac{c_i}{C} \tag{3}$$

Therefore, the ML estimate is simply the empirical frequency of each word (normalised by the sum of all counts C).

b Bayesian Inference

Words: XX

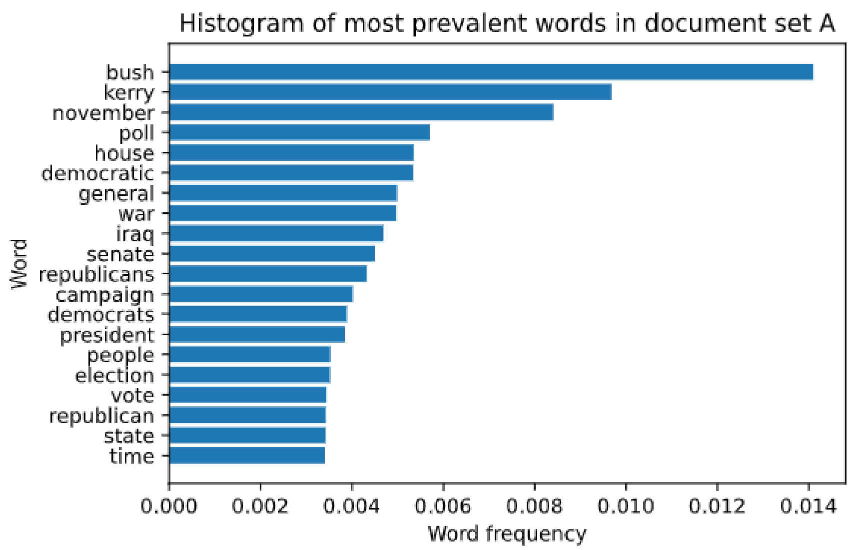

Figure 1: Histogram of top 20 most prevalent words in test set A