CS7DS3 Assignment 2

February 25, 2019

To be handed into the SCSS Office by 12 noon on Friday 12th April, 2019. Please remember to print your name and student number on the front of your script.

Question 1 [25 marks]

Suppose y follows a Poisson distribution, so that $y \sim P(\theta)$, and

$$\mathbb{P}(Y = y) = \frac{1}{y!}\theta^y \exp(-\theta).$$

.

a) Represent the Poisson distribution in exponential family form. Explicitly identify the sufficient statistics $s(y)$, natural parameter $\phi(\theta)$, link function $g(\theta)$, and normalising constant $h(y)$. [15 marks]

b) A financial company asks you to construct a model that predicts the number of times a client will default on loan payments based on covariate information such as number of existing loans, size of company, etc. The company believes that a Poisson regression would be suitable to model the data. However, for reasons relating to privacy and data volume, they cannot provide the raw data, and can only provide summaries.

What summaries do you require to analyse the data? How will the summaries affect your analysis? Justify your answer. [10 marks]

Question 2 [25 marks]

Latent Dirichlet allocation (LDA) is a statistical model popular in natural language processing applications. The model assumes that the data consist of a corpus made up of n documents. Each document consists of a number of words.

The model assumes that the words in each document can be modelled as coming from a combination of "topics". Each topic has a probability distribution over a dictionary of words. For example, a newspaper article could be about finance ("money", "invest") and technology ("data", "artificial", "intelligence).

Assume there are K topics hidden in the data. Let β_k denote the word distribution for topic k. The model has the following additional structure:

- Document-topic proportion parameter $\gamma_i = \gamma_{i1}, \dots, \gamma_{iK}$ denotes the topic distribution for each document i;
- Word-topic indicator z_{ij} denotes the topic that word j in document i belongs to;
- $\alpha = \alpha_1, \dots, \alpha_K$ are hyper-parameters for the distribution of γ.

The generative process for the model is as follows:

- For each document $i = 1, \dots, n$, generate $\gamma_i \sim \mathcal{D}(\alpha)$;
- For each word w_j :
  - Choose a topic $z_{ij} \sim \mathcal{M}(\gamma_i)$;
  - Conditional on this topic, choose a word $w_{ij} | z_{ij} = k \sim \mathcal{M}(\beta_k)$.

Here $\mathcal{D}$ refers to a Dirichlet distribution and $\mathcal{M}$ a multinomial distribution.

a) Identify which elements of the model are latent variables and which are parameters.

[10 marks]

b) Based on the generative description above, sketch the graph of the LDA model. Compare the graph of this model to that of a mixture model. Using the graphs, or otherwise, comment on any similarities or differences between the two models that you notice.

[15 marks]