Roman Shaikh

Student number: 18300989

CS7DS3-Applied Statistical Modelling

Short-Assignment-2

# Question 1:

Suppose follows a Poisson distribution so that and

1. General Exponential family of distribution form is given by:

Where is called “**natural parameter**”

is called “**sufficient static**”

is called “**underlying measure**”

is called “**log normalizer**”

Given,

Therefore, on comparing above equation with standard form, we get

When we have enough samples / observations , this PDF becomes as follows,

Here, is **Natural parameter.()**

is the **link function / log normalizer**. **()**

Is **the normalizing constant. ()**

is the **sufficient static. ()**

1. We build a Poisson regressing model using

Here is a count Response variable.

We can also consider y/t (i.e. the incident / response), where is the time interval or space or grouping.

is the exploratory variable.

Therefore, we consider as our outcome.

We require, the No. of times a customer defaults a loan - ? | Given some variables.

For this, We first group customers based on defaulter count.

We consider variables like the annual income & No. of Loans of a customer.

We can group this variable in different brackets like:

* **Annual income {**Low income (0 to 30k), High income ( 30 to 50k)}
* **No. of Loans** {<2, 2-5, >5}

The customers with same values of feature variable belong to same group. i.e. defaulter or not defaulter.

E.g. 0 – No defaulter & 1 - Yes defaulter.

|  |  |  |
| --- | --- | --- |
| Total Loan | Income group | Defaulter status |
| <2 | High | 0 |
| 3-5 | Low | 1 |
| >5 | Low | 1 |
| >5 | High | 0 |
| <2 | Low | 0 |
| … | … | … |

Therefor as from above example we can identify a defaulting customer based on the No. Of loans and the income group.

**Effects on analysis:**

Interpretation of parameter estimates

1. When , we get i.e. mean of [ expected count of y is ]. This implies X is not related to y
2. If , then and expected count is times larger than when
3. Similarly, if , then and expected count is times smaller than when

If the time t is not considered, then we can apply linear regression instead of Poisson.

# Question 2:



Latent Variable:

= Parameter of the Dirichlet prior on the pre-document topic distribution

= Dirichlet prior on the pre-topic word distribution.

= topic distribution of document i

= topic of word j in document i

Parameters:

= specific word

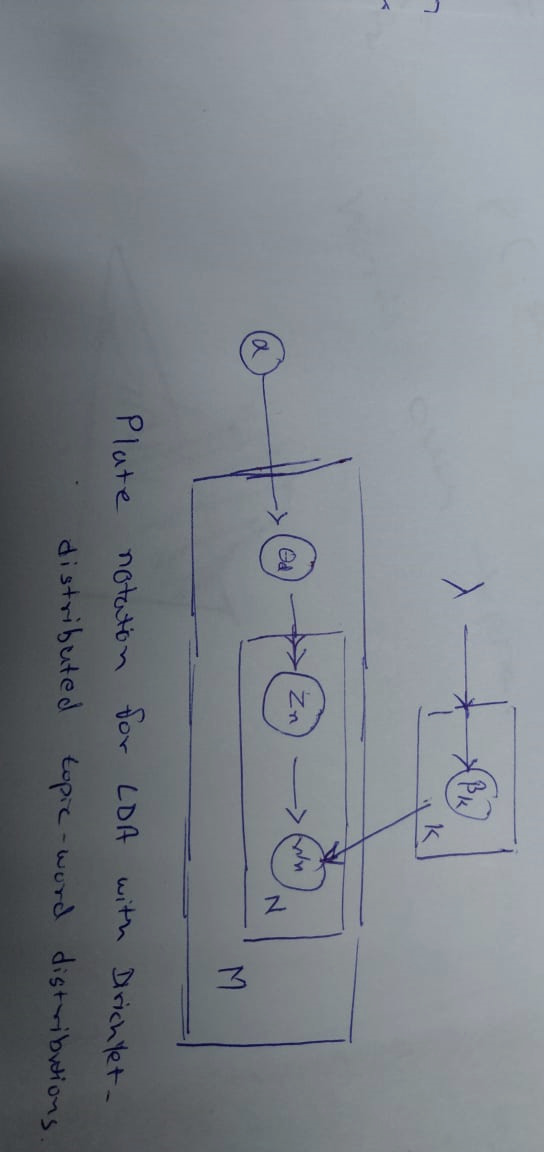
**Latent Dirichlet allocation (LDA):**

* LDA is a generative model in statics which allows set of observed values to be explained by unobserved groups.
* LDA is an example of topic model, widely used in NLP for classifying words in a document and identifying topics.

The PDF of Dirichlet distribution is given by

, where and

**Graph for Latent Dirichlet allocation (LDA):**

****

*Plate notation for LDA with Dirichlet-distributed topic-word distributions*

Explanation:

1. First, we draw multinomial distribution for each topic
2. Then for each document we generate a multinomial distribution
3. Follow this by selecting for each work a multinomial distribution parametrized by Y

Below we find the visualization of Dirichlet distribution with varying :

🡪 We suppose uniform mean across m =equal probability distribution for all multinomial distribution.

🡪 If is increased to a larger value, this concentrates the probability distribution around mean i.e. in center of the simplex.

🡪 If mean m is different and is adjusted, then the mean can be concentrated around three corners

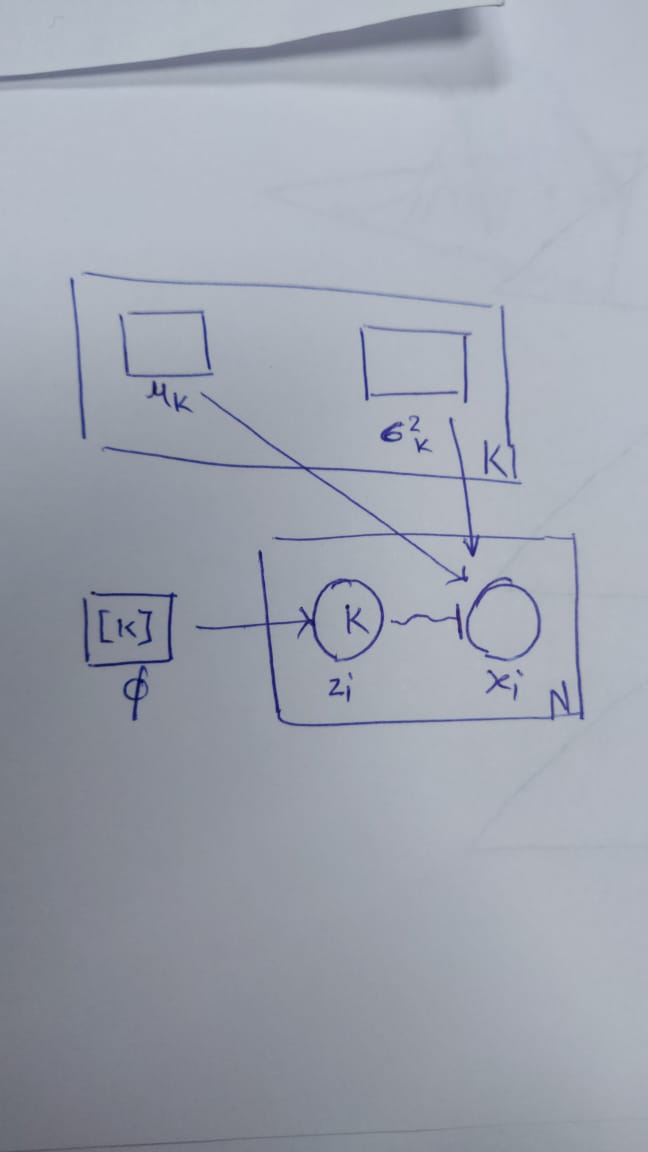
🡪 when m = no. of topic concentrated around edges with very less probability, to belong to any of the topics.

**Mixture model:**

* Mixture model is a probabilistic approach for representing the normally distributed subpopulations within an overall population.
* They generally don’t require knowing which subpopulation a data point belongs to , allowing the model to learn the subpopulation automatically.

**One-dimensional Model is given by:**

**Graph for Mixture Model (Non-Bayesian GMM):**

 Where,

K 🡪 Number of mixture components

N 🡪 number of observations

🡪 component of observation i

🡪 observation i

🡪

🡪 mean component of i

*Non-Bayesian Gaussian mixture model using plate notation.* 🡪 Variance component of i

🡪

**Differences between these two models:**

**LDA:**

* LDA is not a purely mixture model but a **admixture** model.
* Has more than one latent variable.
* Used to model documents and topics
* For each topic we calculate the Multinomial distribution

**Mixture Model:**

* Mixture models can be either Gaussian mixture model, multivariate gaussian mixture model OR a categorical mixture model
* MMs use single latent variable
* This are modeled over number of observations.
* There are two main parts in a mixture model a kernel with parameter represented by and a mixing distribution