Topic Modelling

13-12-2020

Q How do you interpret $P(t/d)$ where d is a document and t is a term?

$P(t/d)$ = How likely is it that document d generated the term t.

Q Why do we need Latent variables?

We wish to capture as much information as possible about the documents that we have. Just using $P(t/d)$ as variables will be lot of variables to estimate.

no. of variables = no. of terms × no. of documents.

Solution :- Any document is considered to

have an underlying mixture of topics associated with it. Similarly topic is considered to be a mixture of terms it is likely to generate.

LDA MODEL

$p(z/d)$

z

$p(t/z)$

t space climate tax rule cure vote play

$$p(t/d) = \sum_z p(t/z) \cdot p(z/d)$$

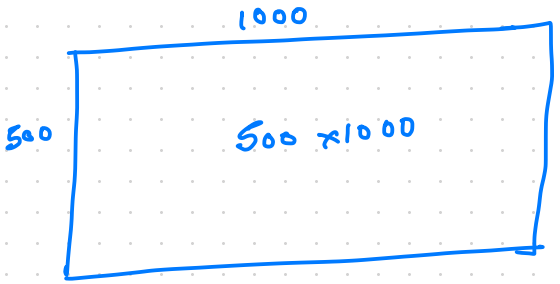
Q How many parameters do we need to estimate now ?

assuming $d = 500$, $z = 10$, $t = 1000$

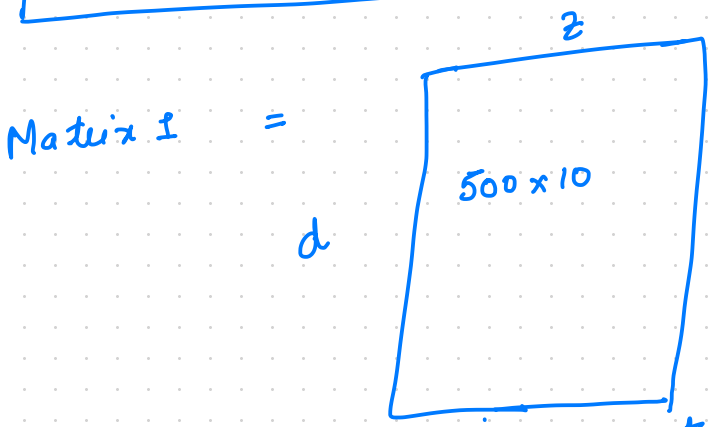
A: $500 \times 10 + 10 \times 1000 = 15000$

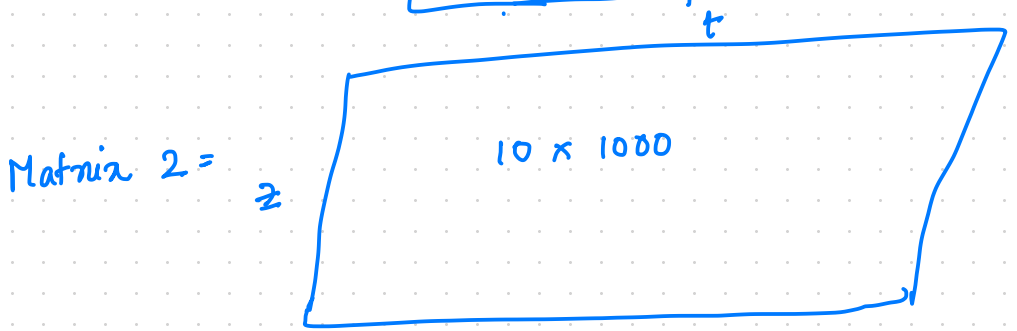
Q. How is LDA a matrix factorization technique ?

$$P(t|d) = \sum_z P(z|d) \cdot P(t|z)$$



= Matrix 1 and Matrix 2





Beside above matrix factorization, LDA has a huge advantage that it gives us bunch of topics that can divide the documents on.

So in general, LDA helps in factoring BOW matrix on the left into two matrices.

Matrix 1 : Indexing documents by topics

Matrix 2 : " topics " word.

Q How to estimate matrix 1 and matrix 2?

We make an assumption that these two topic modelling matrices come from some special distributions.

Q Explain Dirchlet Distributions?

Dirchlet distributions have parameters at the corner.

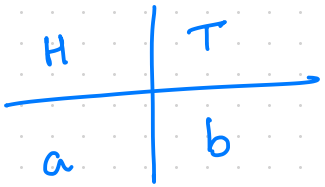
When parameter values are small (0,1), they attract. When they are = 1, they don't do anything. When they are > 1, they repel. Parameters can be thought of as repelling

factors.

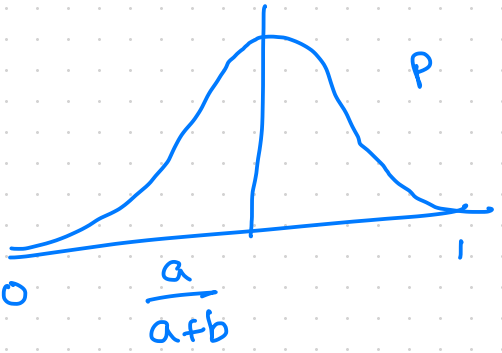
In dirchelt distribution, probability of picking a point on the triangle depends on the height of the probability distribution at that point.

Q What is Beta distribution ?

Coin



Beta distribution



$$\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, x^{a-1}\, y^{b-1}$$

Γ = gamma function → it is a continuous version of factorial function

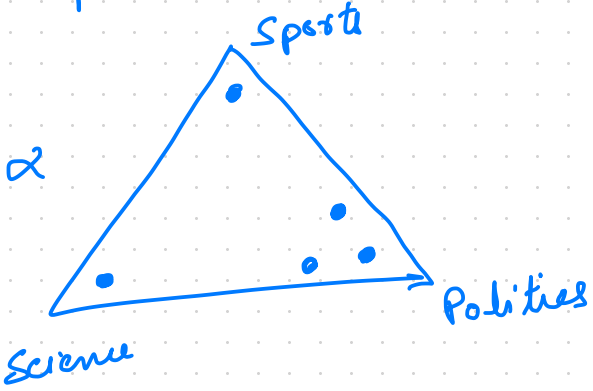
$$\Gamma(a) = (a-1)!$$

Q How to find two matrices ?

Entries for matrix 1 come from picking points in the distribution alpha.

Entries for matrix 2 come from picking points in the distribution Beta.

1) We start with dirchlet distribution for topics alpha.



2) We draw some points corresponding to all documents

3) Lets draw one point from above. It will be a multivariate distribution theta.

θ

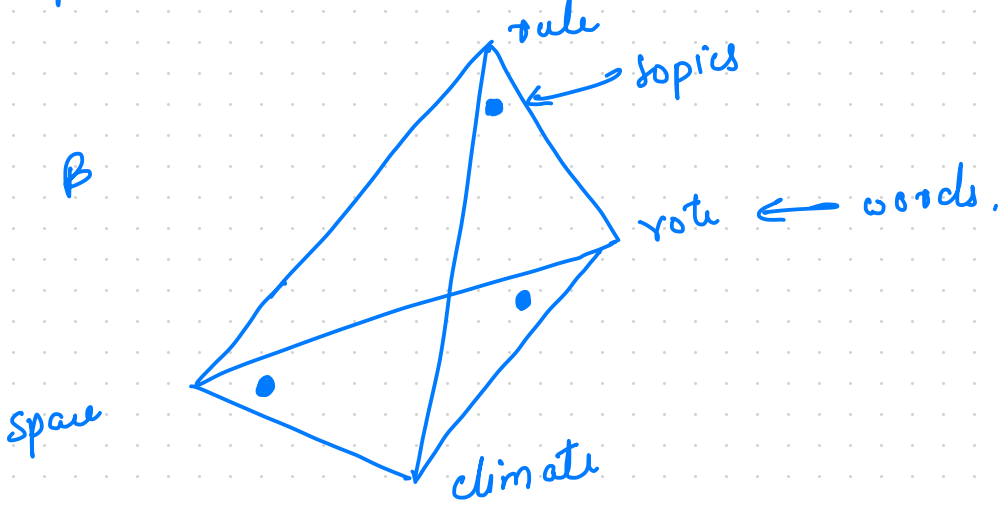
science	politics	sports
0.7	0.2	0.1

This is a mixture of topics corresponding to

document one.

4) From θ, we draw some topics. No. of topics will be decided by the poisson variable.

5) Now we will assign words to these topics using words dirichlet distribution β.



6) From each these data (topics), we generate a distribution of the words, generated by each of the topic.

space	climate	vote	rule
0.4	0.4	0.1	0.1

space	climate	vote	rule
0.1	0.2	0.5	0.2

These distributions are called ϕ.

7) For each of the topic we have chosen, we will pick a word associated with it using multivariate distribution ϕ.

8) This way we will generate document. (fake).

9). Then we will use MLE to figure out the arrangements of the points which will give us the real articles with the highest probability.