WORKSHOP SOLUTION - POINT 10

What are the underlying mathematical principles behind LDA?

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

At a high-level, the tensor decomposition algorithm follows this process:

The goal is to calculate the spectral decomposition of a **V** x **V** x **V** tensor, which summarizes the moments of the documents in our corpus. **V** is vocabulary size (in other words, the number of distinct words in all of the documents).

The spectral components of this tensor are the LDA parameters α and β, which maximize the overall likelihood of the document corpus. However, because vocabulary size tends to be large, this **V** x **V** x **V** tensor is prohibitively large to store in memory.

Instead, it uses a **V** x **V** moment matrix, which is the two-dimensional analog of the tensor from step 1, to find a whitening matrix of dimension **V** x **k**. This matrix can be used to convert the **V** x **V** moment matrix into a **k** x **k** identity matrix. k is the number of topics in the model.

This same whitening matrix can then be used to find a smaller **k** x **k** x **k** tensor. When spectrally decomposed, this tensor has components that have a simple relationship with the components of the **V** x **V** x **V** tensor.

Alternating Least Squares is used to decompose the smaller **k** x **k** x **k** tensor. This provides a substantial improvement in memory consumption and speed. The parameters α and β can be found by “unwhitening” these outputs in the spectral decomposition.

After the LDA model’s parameters have been found, you can find the topic mixtures for each document. You use stochastic gradient descent to maximize the likelihood function of observing a given topic mixture corresponding to these data.

MATHEMATICAL PRINCIPLES:

- Manifold assumption

- Fuzzy simplicial set

- Riemannian geometry

- Stochastic optimization

What are the underlying mathematical principles behind LDA?

You can use LDA for a variety of tasks, from clustering customers based on product purchases to automatic harmonic analysis in music. However, it is most commonly associated with topic modeling in text corpuses.