Pivotal

Manish Tripathi <mtripathi@pivotal.io>

Thursday Thunder: Week 4, Learning 2

Manish Tripathi <mtripathi@pivotal.io> Fri, Oct 30, 2015 at 3:46 PM
To: PDS-ALL <PDS-ALL@pivotal.io>

TL;DR

Hi

Back again!

This week I am keeping it very brief with no question.

Why LDA should not be used with tf-idf data

Yesterday, during the Knowledge Sharing , April did a fantastic job. I as usual was a bit rough on her regarding my comment on LDA that it should not be used with tf-idf data and should actually be used with only term frequency (pure counts).

I didn't want to digress the discussion yesterday so didn't provide a reasonable answer. Doing it here and keeping it as short as possible. Original requires Bayesian understanding something which Woo can better explain. He is the only Bayesian amongst us I guess :D.

So what is the purpose of LDA?.

LDA stands for Latent Dirichlet Allocation. Everyone knows that. What does it try to do?

Given a set of documents LDA would help you find hidden topics in your documents. Hence it is also used in Topic Modeling.

Ok , that's fine but why shouldn't I use tf-idf data?.

Well, it has to do with the way LDA algorithm is devised. The assumption is that there are say M documents. Each of these M documents is represented as word vector where the vector size is the no. of unique words also known as Vocabulary of the document. So each document is represented as a word vector with N unique words.

The idea of LDA is that there is a multinomial distribution where the no. of categories happen to be the no. of words (N) such that each word is sampled from this distribution. You can have multiple samples to get multiple counts of each word.

Ohh.. But what is this multinomial distribution?
Simple way to understand this:

If a Random Variable has:

Two categories, One trial---> Bernoulli Distribution

Two outcomes, Multiple Trails--> Binomial Distribution
Multiple Outcomes, Single trial--> Categorical Distribution
Multiple Outcomes, Multiple Trials--> Multinomial Distribution

As you see, multinomial distribution is a categorical or discrete distribution with N categories. Having a real value or continuous value(which would happen if you tf-idf your data) would theoretically not make sense with the algorithm.

Ok. That sounds fine, but why is it called Latent Dirichlet Allocation?
Good question. So the idea is that each word is generated from a multinomial distribution whose parameter is a dirichlet distributed. Each different value of a dirichlet distribution represents a hidden(latent) topic in the document. And each word is generated conditioned on the fact that a particular value of the parameter , hence a particular topic is already being chosen. So it is basically allocating each word to a topic which is part of the whole document.

Hmm.. So but how can the distribution parameters be a random variable? Didn't we study that Gaussian distribution has fixed mean and sigma parameters?
This is where Frequentist and Bayesian have been fighting for ages. Frequentist says, the distribution parameter is a fixed value once you estimate it using Maximum Likelihood Estimation or Method of Moment. Bayesian says that even the parameter is a random variable and you use Maximum a Posterior (MAP) estimation where you assume a value to the random variable (prior) and then change it based on the evidence in the data and get a new estimate (posterior).

Optional Reading on Bayesian and Frequentist:
An interview question I was asked long back.

You go to a casino and play a game of roulette with only two number options. You decide if I get heads in the coin I would bet one number else I would bet the other. You toss a coin 10 times. 8 out of 10 you get Heads. Your friends suggests you to bet on heads on the 11th trial. Is this a good strategy?.

A **naive person** would answer that it's a good strategy and one would expect heads on the 11th trial.

A **little statistics** educated would say it's a bad strategy as the coin flips are independent in nature and the next outcome would not depend on the previous outcome so it has the same 0.50 chance of seeing either heads and tails.

A **Bayesian** would argue that the first naive person is actually right and you should bet on heads.

Apparently the Bayesian and the naive guy both are correct. Coin flips are independent conditioned on the fact that the coin is fair. If you don't know that the coin is fair apriori, then the second guy doesn't stand a chance in his reasoning.

Thanks

PS: If you are wondering which guy was I, then I was the second guy... and still cleared the round. :O. But that's wrong. Yes,it is but only if the other person knows it is wrong. Interviewer was also the second guy :D.