データ解析入門

by

Hitoshi Nakanishi

Powered by OpenBook

§1 データ解析: LDA

1.1 データ解析: LDA

LDAとは、

LDA とは "Latent Dirichlet Allocation"。文書中の単語の「トピック」を確率的に求める言語モデル。

目的と特徴

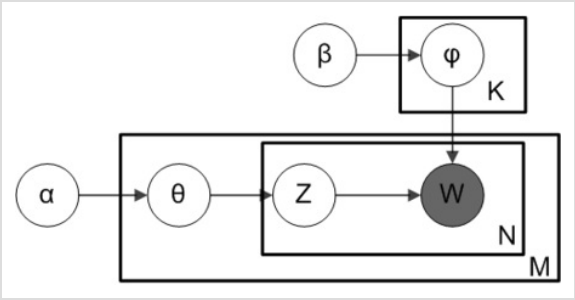
各単語が「隠れトピック」(話題、カテゴリー)から生成されている、と想定して、そのトピックを文書集合から教師無しで推定することが目的。 果物のappleと音楽のappleとコンピュータ関連のappleを区別することが出来る(ことが期待される)という特徴がある。

LDA 概要

一言で言えば、

単語は独立に存在しているのではなく、潜在的なトピックを持ち、同じトピックを持つ単語は同じ文章に出現しやすい。

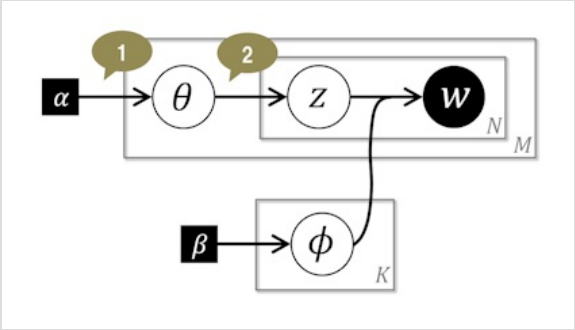
に着目している。



LDA の初出は [Blei+ 2003] Latent Dirichlet Allocation 。

w が単語、z がその単語のトピック、θが文書に対してどのトピックが生起しやすいかの多項分布。 φがトピックに対してどの単語が生起しやすいかの多項分布。 それぞれの多項分布には(対称な)ディリクレ事前分布が入っていて、α&βはそれぞれのハイパーパラメータ。

一般に言うLDA での推論とは、



M 個の文書集合(m 番目の文書は N=N_m 個の単語 (w_mn) をもつ)が与えられたとき、 もっともらしい z とかφとかθとか、場合によってはαとかβとかを推定すること。

ハイパーパラメータからディリクレ分布に従って『文書の数だけ』が生成される。 このは以下のような『文書内の単語ごとの』トピックを決める、いびつなK面サイコロ。文書ごとに形が変わる（分布が変化する）。 LDAのも「トピック分布」と呼ばれる。文書内のトピックの構成比を決めるために「トピック混合比」とも呼ばれる。 トピック混合比が決定により文書内の単語の構成比はおおよそ決定される。

教師なしとは

データ・クラスタリングとかでよく使われる手法で、データを外的基準なしに分類する。出力を先に見えない、決めていない点で異なる。

pythonでの**LDA**の一般的な使い方

1. MeCabとかで、文章から日本語、例えば名詞だけを抽出した配列を作る。
2. Gensimで特徴語辞書を作る。この特徴語辞書にfilterをかけて、高頻度or低頻度すぎるものを削除する。
3. GensimのBoWで各文章の特徴語をカウントして特徴ベクトルを作る。
4. 何個か正解の特徴ベクトルでfitさせてtrainingさせる

具体的には、

・日本語については、形態素解析して名詞だけ取り出す、英語では大文字を全て小文字に変換を施すよう

・辞書作りで、助詞(「の」とか)をfilterで削除する。またstop wordを定義して、”I” “He”とかを削除しておく。辞書形式でidつきで保存(重要:頻度も一緒に)

・特徴ベクトルで表現された文章をcorpusとよぶ。

参考サイト

scikit-learnとgenesimでニュース記事を分類する
http://qiita.com/yasunori/items/31a23eb259482e4824e2

Python用のトピックモデルのライブラリgensimの使い方
http://sucrose.hatenablog.com/entry/2013/10/29/001041

協賛企業募集