

多重トピックテキストの確率モデル ーテキストモデル研究の最前線ー (1)



上田 修功 ueda@cslab.kecl.ntt.co.jp

日本電信電話(株)
NTTコミュニケーション科学基礎研究所

齊藤 和巳 saito@cslab.kecl.ntt.co.jp

日本電信電話(株)
NTTコミュニケーション科学基礎研究所

テキストモデルと言語モデルの違いは？

言語モデルとは、単語の系列 $w = \langle w_1, \dots, w_M \rangle$ が観測される確率モデル $P(w)$ を指す。たとえば、“京都には多くの…”という文に対し、…の部分には、“寺がある”、“山がある”、“研究所がある”などを確率的に予測する問題を扱う。このようなモデルを考えることにより、たとえば音声認識で、認識が不十分な個所を、言語モデルで、よりもっともらしい候補を絞り込むことができる。代表的な言語モデルとして、単語の系列の規則性を直前の $N-1$ 単語組の生起確率というかたちで定式化する N グラムモデルがある¹⁾。

一方、テキスト検索、テキスト分類などの電子テキストを知識源として有効利用するためのテキストマイニング関連の応用では、主として、1センテンスの理解というよりは、文章の大意、概念等を反映したより大まかなトピック^{☆1}解析が必要となる。テキストモデルは、このような応用への貢献を意図したモデル化であり、1つの文章 d がどのようなトピックについて書かれているのかを考慮した確率モデル $P(d)$ を取り扱う。この意味で言語モデルと異なる。言語モデリングに対し、テキストモ

デリングは、まだ歴史の浅い研究分野である。

国語辞典によると、文書とは、“文よりも大きい言語単位で、それ自身完結し、統一ある言語表現をなすもの”、“文字を連ねてまとまった思想を表現したもの”と説明されている。一方、テキストは、“原文”、“書かれたもの”と定義され、必ずしもトピックや心情などが背後にあるとは限らない文字列と解釈できる。document (文書) が可算名詞で、text (テキスト) が不可算名詞であるのは、“まとまり”の有無による違いと思われる。この観点では、トピックを考慮したモデルは、テキストモデルというより文書モデルと呼ぶのが正確かもしれない。しかし、テキスト分類、テキストモデルという表現が多く関連研究者により一般的に用いられているので、本稿でも慣習に従い“テキストモデル”とするが、その真意は“文書モデル”を意図している。

言語モデルにおいても、単語系列のみならず、クラスという概念を導入したモデルもすでに提案されている。たとえば、 N グラムモデルにクラスを導入した N クラスモデルや、隠れマルコフモデルに基づいて、単語の系列を確率的オートマトンとしてモデル化する方法がある¹⁾。しかし、そこでのクラスは、単語の品詞情報であって、テキストモデルで取り扱う文書のトピックといっ

☆1 トピックとは、スポーツ、音楽、政治といったテキストの内容を指す。

たものとは次元が異なる。

以上が言語モデルとテキストモデルの違いである。本稿では、テキストモデルに焦点をあてる。

多重トピックをモデル化する

トピックを考慮したテキストモデリングとして、**ナイーブベイズ** (Naive Bayes : NB) モデルが著名である。NBモデルでは文書 d が1つのトピックについて書かれていることを前提とする。しかし、Webページを例にとると、あるページはスポーツと音楽といった2つ以上のトピックについて書かれている場合がある。つまり、テキストモデルでは1つの文書に対し、多重のトピックを取り扱えることが重要で、単一トピックを仮定するNBモデルでは不十分である。

潜在的意味解析 (latent semantic analysis : LSA) を確率モデル化した**確率的潜在意味解析モデル** (probabilistic latent semantic analysis: pLSA)²⁾ は多重の潜在トピックを仮定している。しかし、pLSAは、与えられた文書のモデル化であり、**未知文書を生成するモデル**となっておらず、その意味でテキストモデルとはいえない。

筆者らは、近年、1つの文書で多重トピックを取り扱えるテキストモデル、**パラメトリック混合モデル** (parametric mixture model : PMM) を考案し、Webページの多重トピック分類実験でその有効性を確認した^{4), 5)}。ところが、最近、多重トピックのテキストモデルとして、**潜在ディレクレ割り当てモデル** (Latent Dirichlet Allocation: LDA)³⁾ の存在を知った。そして、LDAとPMMとを詳細に比較することにより興味深い知見を得ることができた。本稿の主眼は、上記pLSA, LDA, PMMを直観的に説明し、従来の言語モデルとの差異、さらには、pLSA, LDA, PMMのモデルの本質的な差異を解説し、読者にテキストモデルの基礎を理解していただくことである。

まず、第1回目の本稿では、テキストモデリングの基礎であるNBモデルの考え方、学習法、応用、さらにはNBモデルの上記問題点について解説する。第2回目で、その問題点を解決するための新たなモデル化 (pLSA, LDA, PMM) について解説する。

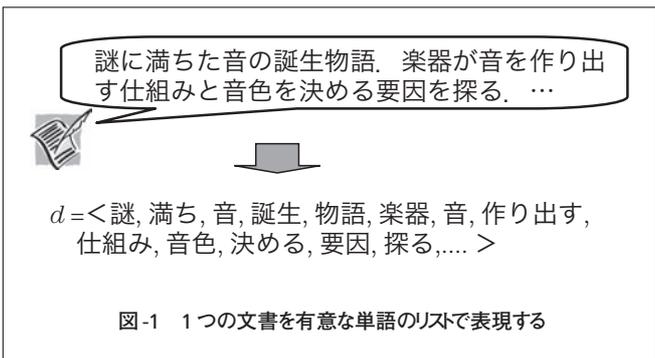


図-1 1つの文書を有意な単語のリストで表現する

確率モデルとは何?

テキストモデルの説明の前に、言語モデルおよびテキストモデルで共通する単語生成の確率モデルとは何かについて説明する。

文書情報はその文書に有用と思われる単語 w_m を抜き出して羅列したリスト：

$$d = \langle w_1, w_2, \dots, w_M \rangle \quad (1)$$

で表現できる。有意な単語数 M は文書ごとに異なる。図-1の例で“音”が2回現れているように、 $w_m, m=1, \dots, M$ の各々は必ずしも異なる単語とは限らないことに注意。以下では、このリストを文書 d と同一視し、文書 d といえ、式(1)のリストを指すものとする。

日本語の場合、英語と違って単語が空白で分割されていないので、**形態素解析**という処理により、品詞情報を手がかりに単語に分割する。次いで、コーパス全体での低頻度語やstop wordsと呼ばれる文書の内容に関与しない語の削除、さらに英語の場合、3人称単数形や過去形などで語尾変化した語を同一視するための語末処理を施し d を得る。コーパス全体から抽出された単語群から解析対象の有意な語彙集合 \mathcal{W} が定まる^{☆2}。

$$\mathcal{W} = \{t_1, \dots, t_V\} \text{ ただし, } t_i \neq t_j (i \neq j)$$

t_i は第 i 語彙 (term) を表す。 V は語彙総数 (コーパスに渡る異なる単語の総数) とする。

単語生成の確率モデルとは何か? 単語生成確率モデルとは、 w_1, \dots, w_M が同時に共起する確率分布、すなわち、1文書に出現する単語の同時分布： $P(w_1, \dots, w_M)$ を意味する。換言すれば、図-1に示した文書は、この分布から

☆2 たとえば、Yahoo!ドメイン内のWebページでは1文書あたりの有意な単語数 (N) は約100で、約1万ページでの語彙総数 V は数万単語に及ぶ。

生成された実現値と見なせる。その意味で、確率モデルはより正確には、**確率的生成モデル** (probabilistic generative model) と呼ばれる。

何の制約もなくこの同時分布を推定するのは無茶というもので、何らかの妥当な仮定を設ける必要がある。前述した N グラムモデルでは、“ある単語の生起はその直前の $N-1$ 個の単語がどのような単語であったのかに依存する” という仮定を用いる。

どのような仮定を設けるかは、何のためのモデル化を目指すのかに依存する。また、モデルはできる限りシンプルなものが多い。その理由は、複雑なモデルは学習データに過度に適応し、未学習データに対するモデルの予測能力 (汎化能力) が著しく低下するからである。文書全体の潜在的な意味を解析する応用では、各単語の出現頻度のみで十分と考え、各単語の生起は統計的に独立と仮定してモデル化する。この仮定に基づく文書表現が以下に述べる **BOW 表現** である。

BOW 表現と NB モデル

■ BOW 表現とは

テキスト分類などの応用では単語の順序は必ずしも必要ではなく、文書中にどのような単語がどのような頻度で出現するかの情報で十分な場合が多い。単語の順序を無視し、文書を**単語の集合**としてとらえるテキスト表現として **bag-of-words: BOW**^{☆3} が用いられる。

1つの文書 d が単一トピックからなると仮定するテキストモデルである NB モデル^{☆4} もこの BOW を土台とする。NB モデルでは、トピック c を持つ文書 d にて、各単語 w_m の生起を統計的に独立と見なすので、独立性の定義から次式：

$$P(d|c) = P(w_1, \dots, w_M|c) = \prod_{m=1}^M P(w_m|c) \quad (2)$$

が成り立つ。ここで、 $P(w_m|c)$ はトピック c における $w_m \in \mathcal{W}$ の生起確率を表す。このように M 個の確率変数の同時分布を式 (2) に示す M 個の分布の積に分解できれば、各単語の出現順序は無意味となる。これが BOW 表現である。以下では、特に明記する必要がない限り、 c に関する条件を省略して記述を簡略化する。

次に、単語頻度ベクトル $\mathbf{x} = (x_1, \dots, x_V)$ を導入する。 V は前述したように、語彙集合 \mathcal{W} の要素数 (語彙総数) を表す。 x_i は $t_i \in \mathcal{W}$ が文書 d 中に出現した回数 (頻度) を表すものとする。出現単語を語彙 t_i $i=1, \dots, V$ ごとに整理することにより、 $P(w_1) \times P(w_2) \times \dots \times P(w_M) = P(t_1)^{x_1} \times P(t_2)^{x_2} \times \dots \times P(t_V)^{x_V}$ が成り立つことが分かる。 $P(t_i)$ は $t_i \in \mathcal{W}$ が生起する確率を表す。文書の d のリストの長さは M 故、 $x_1 + \dots + x_V = M$ が成り立つ。

さらに、表記の便宜上、 $P(t_i) = \theta_i$ と書くと、結局式 (2) は次式のように書ける。

$$P(w_1, \dots, w_M; \boldsymbol{\theta}) = \prod_{i=1}^V \theta_i^{x_i} \quad (3)$$

これが BOW 表現に基づくテキストの NB モデルである。式 (2) の表現では、文書ごとに M の値が異なるのに対し、式 (3) では全文書に共通の V で表現できる利点がある。 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_V)$ は未知パラメータである。出現単語 w_m は語彙集合 \mathcal{W} の要素故、 t_1, \dots, t_V のいずれかであるので、 $\sum_{i=1}^V \theta_i = 1$ が成り立つ。

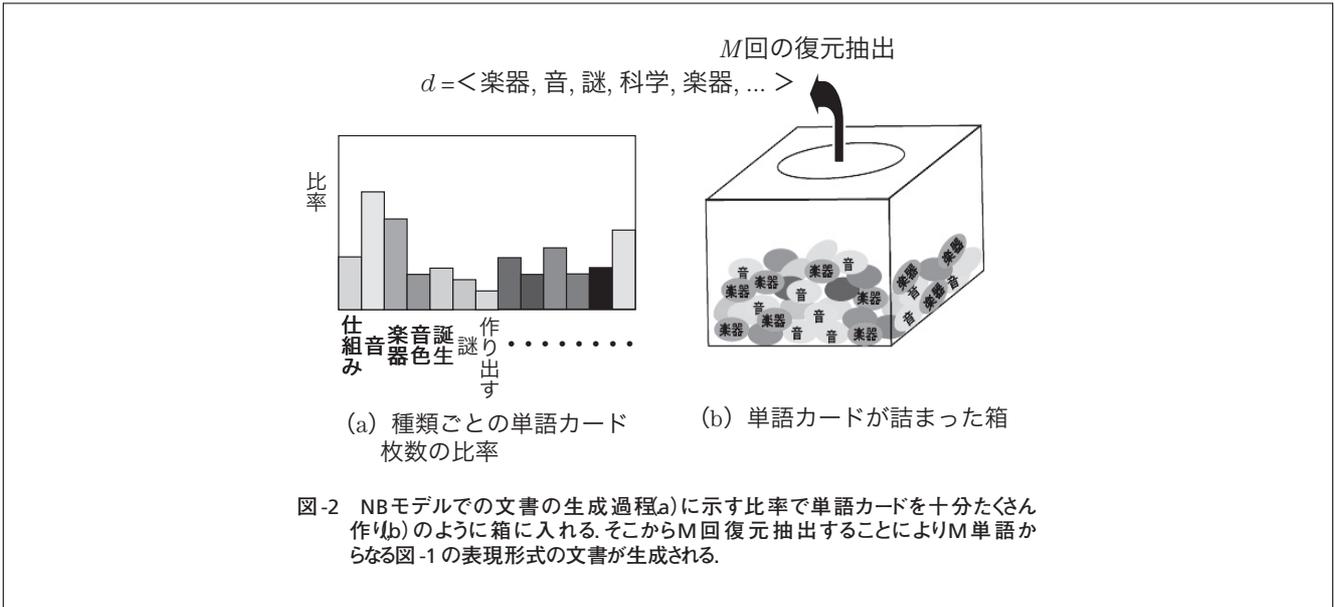
■ NB モデルでのテキスト生成過程

NB モデルによる文書の生成過程を図-2 を用いて説明しよう。今、“音楽” というトピックにおける単語の生起分布が図-2 (a) に示す分布に従うとする。この分布に従って多数の単語カードを作り、それらを箱の中に入れる (図-2 (b))。この箱から無作為に 1 枚カードを取り出しそのカードに書かれた単語をメモし、そのカードを箱に戻す。上記操作を M 回繰り返すことにより、式 (1) のような M 単語からなる単語のリストが得られる。以上が、式 (3) でパラメータの値が既知としたときの NB モデルによる文書生成過程である。式 (3) のパラメータベクトル $\boldsymbol{\theta}$ の各要素が図-2 (a) の各単語の枚数の比率に相当する。したがって、式 (3) での V は図-2 (a) の単語の種類数に相当する。

もちろん、NB モデルは BOW を仮定し単語の順番には意味を持たせていないので、生成された文書は我々が書く文書とは程遠い。しかし、このモデルで多くの文書を生成すると、それら文書群にはよく出現する単語群が観測され、それに伴いそれら文書群に共通する概念 (音楽というトピック) が浮かび上がる。つまりランダムな単語の生起では得られない何らかの規則性が観測され、

☆3 テキストを“単語が無秩序につまったかばん”と比喩表現している。

☆4 日本語では、“あの人はナイーブだ”という風にナイーブという言葉は“純粋で素直な”という意味で用いられることが多いが、英語では“大雑把な、いい加減な”という意味で用いられることが多い。したがって、NB モデルは同時分布を独立性を仮定して積表現する大雑把なモデルということになるが、それで応用上十分であればむしろ“好い加減”なモデルといえよう。なお、NB モデルはテキストモデルで誕生したのではなく、むしろテキストモデルへの応用と位置づけられる。



その意味でテキストの生成モデルといえる。文書数を増やしていくにつれてそれら文書群に渡る単語の種類と頻度分布は図-2 (a) に漸近する。

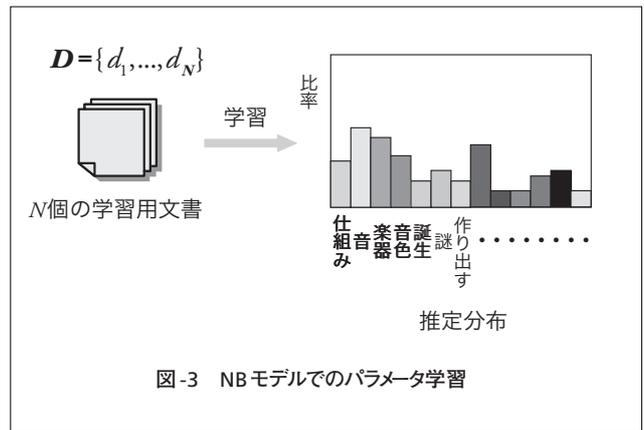
パラメータ学習

前章でパラメータの値が既知の下でのNBモデルによるテキスト生成過程について説明した。パラメータ学習はまさにこの逆問題である。今、図-3に示すように N 文書($d_n, n=1, \dots, N$)が与えられたとする。ただし、 d_n は式(1)の表現形式とする。これらがNBモデルから生成されたと仮定したらNBモデルのパラメータ $\theta=(\theta_1, \dots, \theta_V)$ はどのような値が妥当か？これがNBモデルの学習(パラメータの推定)の問題である。

一般に、確率モデルのパラメータ推定値の妥当性の基準としてよく用いられるのが、尤度(ゆうど)最大化および事後分布最大化基準で、各々、最尤(ML)学習、事後分布最大化(MAP)学習と呼ばれる^{☆5}。以下では、NBモデルに対する各々の学習法を説明する。

■最尤(ML)学習

最尤(さいゆう) (ML) 学習では尤度関数を最大化するパラメータを推定値とする。NBモデルの場合、単語の出現分布は式(3)で与えられた。今、このモデルから



$D=\{d_1, \dots, d_N\}$ が得られたとする。ただし、 $d_n, n=1, \dots, N$ は式(1)の形式とする。リスト長(式(1)の M)は文書ごとに異なる。

このとき、観測データの分布は $P(D; \theta)$ と書け、 D は観測された固定値(定数)故、 $P(D; \theta)$ はもはやパラメータの関数と見なせる。これを尤度関数と呼ぶ。ML学習ではこの尤度関数を最大化するパラメータをML推定値と呼び、最適なパラメータ値とする。尤度関数は観測データ D を生じさせる分布と見ることができ、パラメータ値を変えたとき、この分布の値が最大となるパラメータ値を“最も尤もらしい”という意味でML推定値とする^{☆6}。

さて、 D 中のデータ中の統計的独立性(つまり、各文書は独立に生起)を仮定すると、 $P(D; \theta) = \prod_n P(D_n; \theta)$ が

☆5 最尤学習は Maximum Likelihood (ML) learning, 事後分布最大化学習は Maximum A Posteriori (MAP) learning と呼ばれる。参考までに、“A”は冠詞ではなく副詞。
 ☆6 ML推定値の理論的な妥当性についてはやや難解なので省略する。ここではML推定値の直観的な意味を理解しておけば十分である。

成り立つ。また、表記 $x_{n,i}$ が d_n 中で $t_i \in \mathcal{W}$ が出現した回数を表すものとする、式(3)より $P(D_n; \theta) = \prod_{i=1}^V \theta_i^{x_{n,i}}$ と書ける。以上よりML推定値は次の尤度関数を最大化するパラメータとなる。

$$P(D; \theta) = \prod_{n=1}^N P(d_n; \theta) = \prod_{n=1}^N \prod_{i=1}^V \theta_i^{x_{n,i}} \quad (4)$$

対数関数の単調増加性より、式(4)の両辺の対数をとっても尤度関数を最大化する θ は不変なので、結局、形式的には、次の**対数尤度関数**：

$$\log P(D; \theta) = \sum_{n=1}^N \sum_{i=1}^V x_{n,i} \log \theta_i \quad (5)$$

を制約条件 $\sum_{i=1}^V \theta_i = 1$ の下で最大化する $\theta_i, i=1, \dots, V$ を求める問題となる。 θ_i のML推定値を $\hat{\theta}_i$ と書くこととすると、この最大化問題はラグランジュ乗数法を用いて容易に解け(付録参照)、次式を得る。

$$\hat{\theta}_i = \frac{\sum_{n=1}^N x_{n,i}}{\sum_{i=1}^V \sum_{n=1}^N x_{n,i}}, i=1, \dots, V \quad (6)$$

式(6)の分母が D 中に現れた総出現単語数を表すことに注意すると、単語 t_i の生起確率である θ_i のML推定値は、**出現した総単語数に対する t_i の出現回数の比**で求まることを意味し、直観的に自然である。

しかし、ML学習は学習データ数が少ない場合、推定値の信頼性に問題がある。 $N \ll V$ の場合、後述する“テキスト分類への応用”で述べるように各トピックごとにNBモデルを学習するとき、明らかに文書中に一度も出現しないような単語が多数発生する。このとき、ある i に対応する式(6)の分子が零となり、それらの単語に対応する $\hat{\theta}_i$ は零となる。つまり、このモデルで図-2に示した過程で文書を生成すると $\hat{\theta}_i = 0$ となる単語 t_i がまったく生成されないという**零頻度問題**が生じる。少数の学習データで出現しなかったからといって、未来永劫出現しないと考えるのは問題である。次節で述べる事後分布最大化(MAP)学習はこの問題を避けるための学習法で、通常、NBモデルの学習法として用いられている。

■事後分布最大化(MAP)学習

MAP学習では、ML学習と異なり、未知パラメータを**確率変数**と見なす。そして観測データ D が得られた下で

θ の**事後分布**(posterior distribution) $p(\theta|D)$ を最大化するパラメータを最適な推定値とする。

ML推定のと異なり、パラメータは確率変数として扱われているのでパラメータの事前分布 $p(\theta)$ を持つ \star^8 。ベイズの定理から、事前分布はデータ D を観測することにより事後分布 $P(D|\theta)$ と変化する。

$$p(\theta|D) = \frac{P(D|\theta)p(\theta)}{P(D)} \quad (7)$$

MAP推定では、この事後分布は D を生成する θ の分布と見なせる。それ故、その分布のピークを与えるパラメータ値がパラメータの最適な推定値とするのは直観的に自然といえよう。

式(7)の分布の $P(D)$ は θ に無関係故、MAP推定値は、 $p(D|\theta)p(\theta)$ の最大化、すなわち、等価的に

$$\log p(D|\theta) + \log p(\theta) \quad (8)$$

を最大化する θ として求める。ここに $p(\theta)$ はパラメータの**事前分布**(prior distribution)を表す。非負で和が1の V 個の確率変数の同時分布として**ディリクレ分布**(Dirichlet distribution)が知られている。ディリクレ分布を事前分布とすると、 $p(\theta)$ は次式となる。

$$P(\theta) = p(\theta_1, \dots, \theta_V) = f(\alpha) \prod_{i=1}^V \theta_i^{\alpha-1} \quad (9)$$

$\alpha (> 0)$ はハイパーパラメータで定数とする。 $f(\alpha)$ は $\int p(\theta) d\theta = 1$ となるための規格化定数で α のみに依存する。

式(4)、(9)を式(8)に代入すると、 θ のMAP推定値は次式の目的関数：

$$\sum_{n=1}^N \sum_{i=1}^V x_{n,i} \log \theta_i + (\alpha-1) \sum_{i=1}^V \log \theta_i \quad (10)$$

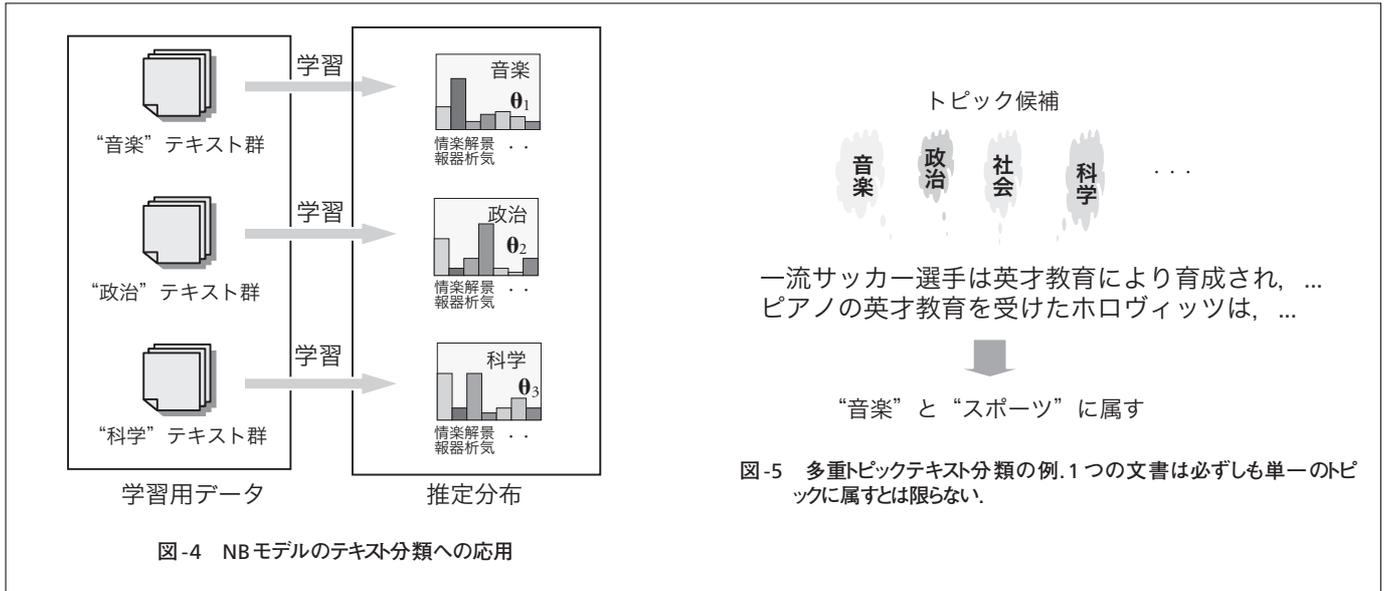
を制約条件： $\sum_{i=1}^V \theta_i = 1$ の下で最大化することにより求まる。ラグランジュ乗数法より以下のMAP推定値を得る。

$$\hat{\theta}_i = \frac{\sum_{n=1}^N x_{n,i} + \alpha - 1}{\sum_{i=1}^V \sum_{n=1}^N x_{n,i} + V(\alpha - 1)}, i=1, \dots, V \quad (11)$$

式(6)と比較すると、式(11)では、分母分子に新たな項が付加されていることが分かる。これは、パラメータの事前分布を考慮したことによる。ハイパーパラメータ α が推定値に加味されることにより、どの単語も最低

\star^7 θ_i は t_i の生起確率故、 $\sum_i \theta_i = 1$ を満たす。

\star^8 本稿では、離散(連続)の確率変数に対する確率分布は大文字 P (小文字 p)を用いて区別する。



$\alpha-1$ 回は出現することになり、ML推定における零頻度問題を緩和できる。 α は一種の平滑化(スムージング)パラメータと見なせる。なお、 $\alpha=2$ とした場合、ラプラススムージングと呼ばれる。

テキスト分類への応用

NBモデルはテキスト分類に容易に応用できる。たとえば、テキストを3クラス(c_1 =音楽, c_2 =政治, c_3 =科学)のいずれかに分類する問題を考える。学習用データは、その文書が所属するクラスのラベルが付与されているとする。このとき、図-4に示すように、各クラスごとに独立にNBモデルを学習する。これにより3つのNBモデル $P(d|c_1)$, $P(d|c_2)$, $P(d|c_3)$ が得られることになる。そして、クラスが未知の文書 d^* に対し、クラス事後確率 $P(c_i|d^*)$ を最大化するクラス c_i^* がベイズ誤り確率最小化の観点で最適なクラス分類となる。

ベイズの定理から、 $P(c_i|d^*) \propto P(c_i)P(d^*|c_i)$ 故、 $P(c_i|d^*)$ の最大化は $P(c_i)P(d^*|c_i)$ の i に関する最大化となる。ここに、 $P(c_i)$ はクラス c_i の事前確率を表す^{☆9}。以上が、NBモデルのテキスト分類への応用である。

一般に、テキストのクラスラベル付けは専門家により成され、労力がかかる。一方、ラベルなしテキストは大量に入手できる。大量のクラスラベルなしテキストを少

量のクラスラベルありテキストに混在させて学習し、少量のクラスラベルありデータのみによる分類精度を向上させるNBモデルの学習法も提案されている。詳しくは文献6)を参照されたい。

NBモデルの限界

■多重性の欠如

NBモデルはBOWテキスト表現を土台とする自然かつ妥当なモデルといえる。しかし、NBモデルでは1つの文書がある単一のトピックに関して書かれていることを仮定しているので、表現能力の観点で限界がある。たとえば、テキスト分類の場合、図-5に示す文書(の一部)は、音楽とスポーツの両方のトピックに帰属する文章といえる。すなわち、前述したように、一般に文書は単一トピックについて書かれているというよりはむしろ、トピックが遷移しながら**多重**のトピックについて書かれていると考えるのが自然であり、多重性を扱えない点がNBモデルの限界といえる。

■ならば混合分布では？

確率モデルについて知識のある読者の中には、“それならばNBモデルを複数合わせた**混合NBモデル**とすればよいのでは？”と思うかもしれない。しかし、**単一ト**

☆9 事前確率は、学習データでの各クラスのデータ数の比とするか、あるいは、単純に等確率とする。

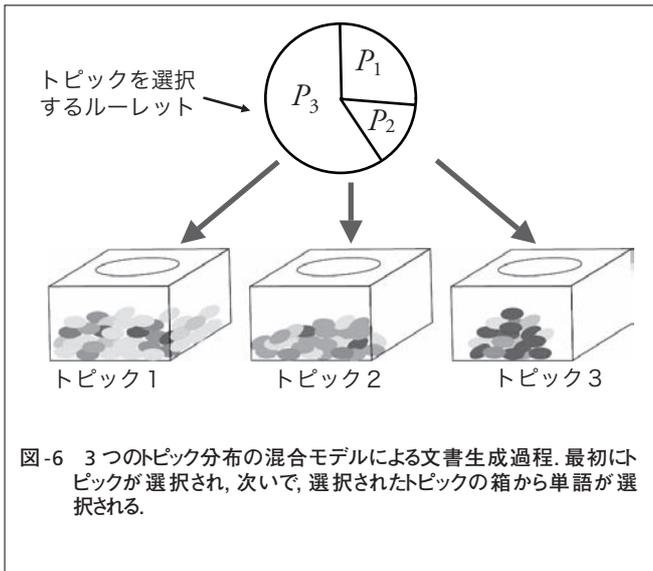


図-6 3つのトピック分布の混合モデルによる文書生成過程. 最初にトピックが選択され, 次いで, 選択されたトピックの箱から単語が選択される.

トピックの要素分布を混合化した従来の混合モデルでは多重性は表現できない. 以下にその理由を説明しよう.

今, 3つのトピックからなる混合NBモデルを考える. このモデルは

$$P(d) = \sum_{i=1}^3 P(c_i) P(d|c_i) \quad (12)$$

と書ける. ただし, $P(c_i)$ は第 i 要素分布 $P(d|c_i)$ の混合比で $P(c_i) \geq 0, \sum_{i=1}^3 P(c_i) = 1$ を満たす. $P(d|c_i)$ はトピック c_i のNBモデルに相当する. 式(12)を見ると, 文書 d が3つのトピックのテキストモデル $P(d|c_1), P(d|c_2), P(d|c_3)$ について各々重み $P(c_1), P(c_2), P(c_3)$ 付きの平均を求めているので, 一見, 多重性を表現できているように思えるかもしれない. しかし, 式(12)のモデルの生成過程を考えるとその誤解は解消される.

式(12)のモデルの生成過程は以下となる.

Step 1. 確率 $P_i = P(c_i), i=1, 2, 3$ に従ってトピックを1つ選択する.

Step 2. 選択されたトピックのNBモデルを用いて M 個の単語を生成し, M 単語からなる式(1)の形式の文書を生成する.

Step 1での確率的トピック選択は, たとえば, P_1, P_2, P_3 の大きさに比例する図-6に示すルーレットを用いればよい. また, Step 2は, 選択されたトピックの箱から, 図-2に示した要領でカードを復元抽出により M 回抽出すれば文書 d を作成することができる.

以上から明らかなように, 混合NBモデルでは3つのトピックを取り扱うものの, ある文書ではそのうちの1

つが確率的に選択される. すなわち, 混合NBモデルで生成された文書集合全体は3つのトピックからなるが, 1つの文書に注目すると, あくまで単一トピックの文書となっており, 多重性を持つことが原理的にできない. ではどうすれば多重性を取り扱うことができるのか? その解決策とそれに関連する確率モデル (pLSA, LDA, PMM) については次回で詳しく解説するので, 乞う, ご期待!

付録: ラグランジュの未定乗数法

ラグランジュの未定乗数法 (簡単に, ラグランジュ乗数法とも呼ばれる) とは, 制約条件付きの最適化問題を制約条件なしの最適化問題に帰着させる一般的手法で, 最適化問題で用いられる常套手段である.

今, $g_k(x)=0, j=1, \dots, K$ を満たす x に対して, 関数 $f(x)$ を最大化 (最小化) する x を求める問題を考える. ただし, x はスカラーとは限らず, ベクトル, 行列でもよい. 上記問題は, 次の目的関数 J の制約なし最大化 (最小化) する問題に変換できる. ただし, 探索空間は, $(x, \lambda_1, \dots, \lambda_K)$ となる. λ_K はラグランジュ乗数と呼ばれる.

$$J(x, \lambda_1, \dots, \lambda_K) = f(x) + \sum_{k=1}^K \lambda_k g_k(x)$$

たとえば, 式(5)の最尤推定の場合, 上式で, x を $(\theta_1, \dots, \theta_V)$ とし, $g(\theta_1, \dots, \theta_V) = \sum_{i=1}^V \theta_i - 1$ とすればよく, このとき,

$$J(\theta_1, \dots, \theta_V, \lambda) = \sum_n \sum_i x_{n,i} \log \theta_i + \lambda (\sum_i \theta_i - 1) \quad (13)$$

となる. 故に, J を $\theta_1, \dots, \theta_V, \lambda$ に関して最大化すべく, J を各 θ_i に関し微分して零とおくことにより, $\partial J / \partial \theta_i = 0$ より, $\theta_i = -1 / \lambda \sum_n x_{n,i}$ を得る. さらに, $\partial J / \partial \lambda = 0$ と連立させて λ を消去すると式(6)を得る.

参考文献

- 1) 北 研二: 確率的言語モデル, 東京大学出版会 (1999).
- 2) Hofmann, T.: Probabilistic Latent Semantic Indexing, Proc. International Conference on Information Retrieval (SIGIR'99), pp.50-57 (1999).
- 3) Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, Advances in Neural Information Processing Systems (NIPS'01) (2001).
- 4) Ueda, N. and Saito, K.: Single-shot Detection of Multiple Topics Using Parametric Mixture Models, Proc. International Conference on Knowledge Discovery and Data Mining (SIGKDD'02), pp.626-631 (2002).
- 5) Ueda, N. and Saito, K.: Parametric Mixture Models for Multi-topic Text, to appear Advances in Neural Information Processing Systems (NIPS'02) (2002).
- 6) Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM, Machine Learning, Vol.39, pp.103-134 (2000).

(平成 15 年 12 月 24 日受付)

