マルチタスク最大マージントピックモデルによる収益予測

Revenue Prediction based on Multi-task Max-margin Topic Models

中川　雄太[1*]　　上野　良輔[1]　　江口　浩二[1†]

[1] 神戸大学大学院システム情報学研究科

Abstract: Researchers and practitioners in the economic and financial field recently have a keen interest in discovering new ideas by making full use of large-scale data, such as in the form of document data of company valuation in online news and the form of numerical data of company financial indices. One promising approach to analyzing such large-scale data is topic modeling, typically by Maximum Entropy Discrimination LDA (MedLDA). MedLDA is a supervised topic model that can improve accuracy of latent topic estimation by making use of the side information associated with each document. In this paper, we generalize Multi-task MedLDA (MultiMedLDA) that simultaneously addresses classification and regression tasks in an extension of MedLDA. In this paper, we evaluate the effectiveness of MultiMedLDA through experiments with enterprise evaluation documents associated with continuous labels of change rate of operating incomes and discrete labels of categories of business, and discuss it compared with single-task MedLDA.

1　はじめに

近年，情報技術の発達によって，情報の発信形態は多様化しており，世の中に存在するテキストデータの量は増加し続けている．これに伴って大規模データを解析する重要性が増しており，代表的なモデルの 1 つとして潜在的ディリクレ配分法 (Latent Dirhlet Allocation: LDA)[1] を代表とするトピックモデルが挙げられる．トピックモデルでは，文書に存在するテキストデータを bag-of-words 形式で離散表現している．bag-of-words とは文書中に現れる単語の語順を無視して単語の出現頻度に着目した表現形式のことである．また，文書中には複数の潜在的な潜在トピックが存在していると仮定しており，この潜在トピックは文書集合中に出現する語彙からなる分布で表現される．更に，LDA では文書ごとに潜在トピック分布が存在していると仮定している．これら 2 つの分布によって文書集合に含まれる単語を表現している．

また，企業や金融機関を中心に経済・金融分野の情報解析への関心が高まっている．経済データは社会のあらゆる事象を反映し，刻一刻と変化する情勢の中で生み出されており，このデータから新たな知見を得る取り組みがなされ始めている．経済・金融分野での文書データに存在する特徴として，企業データや数値を含むデータのような多種多様な付加情報が存在することが挙げられる．この付加情報を考慮した，より有望な解析手法が望まれている．

このように，実世界に存在する文書データは付加情報として離散値（たとえば，ジャンルや著者の性別）や連続値（たとえば，著者の年齢や評価点）を持つことが一般的である．付加情報を学習に活かすことで潜在トピック推定の精度を向上させているモデルとして，教師有りトピックモデル (Supervised topic models)[2] や最大マージントピックモデル（Maximum Entropy Discrimination LDA : MedLDA）[3] が存在する．MedLDA には，離散値を伴う文書データに対応した MedLDA regression と，連続値を伴う文書データに対応した MedLDA classification が存在する．文書の付加情報を潜在トピックの推定に活かすことによって，予測精度の向上を図っている．また，サポートベクターマシン（Support Vector Machine : SVM）[4, 5] におけるマージン最大化法を取り入れ，推定した潜在トピックを特徴量として利用している．ただし，MedLDA は単一の付加情報を持つ文書データを対象としており，離散値ラベルと連続値ラベルの両方を持つデータに対して適用できない．

本論文では，金融・経済テキストデータから企業の収益性に関する指標を予測するモデルを実現することを目的とする．この目的のもと，MedLDA を拡張して分類タスクと回帰タスクを同時に解決するマルチタスク最大マージントピックモデル（MultiMedLDA）に一般化する．MultiMedLDA では複数種類のラベルが付与された文書データを対象としており，複数種類の付加情

*連絡先：神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: nakagawa@cs25.scitec.kobe-u.ac.jp
†連絡先：神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: eguchi@port.kobe-u.ac.jp

報を同時に考慮しながら潜在トピックの推定を可能にしている．また，複雑化した最適化問題を解く為に双対分解 [7] と呼ばれる手法を導入している．双対分解は効率的に解くことができない目的関数に対するアプローチの 1 つであり，目的関数がいくつかの関数に分解でき，それぞれの関数の最適解が効率的に解くことができる場合に適用することが可能である．MultiMedLDA ではより多くの付加情報から潜在トピックを推定しており，予測精度の改善が期待される．本論文では，業種の離散ラベル，営業利益変化率の連続ラベルを伴う企業評価テキストを用いて MultiMedLDA の有効性を評価し，MedLDA の分類タスクおよび回帰タスクと比較して議論する．

2 関連研究

提案手法の基礎研究として，教師ラベルを考慮して潜在トピックの推定を行う MedLDA，および双対分解について説明する

2.1 最大マージントピックモデル

Maximum Entropy Discriminated LDA(MedLDA)[3] は，最大エントロピー識別 [6] と呼ばれる教師付き学習の枠組みにおいて，潜在トピックを特徴として用いつつ，その潜在トピックを推定する手法である．MedLDA のグラフィカルモデルを図 1 に示す．図 1 中の y は各文書に付与された教師ラベルである．また，η はラベル評価時の各トピックに対する重み係数であり，トピックに対する関係度を表す．MedLDA の生成過程を以下に示す．

1. 文書 $d(d \in 1, \ldots, D)$ に対して，$\theta \sim \mathrm{Dirichlet}(\alpha)$ を選択．
2. 文書 d 内の N_d 個の単語 $w_{d,n}(n \in 1, \ldots, N_d)$ に対して，
   (a) トピック $z_{d,n} \sim \mathrm{Multinomial}(\theta_d)$ を選択．
   (b) 単語 $w_{d,n} \sim \mathrm{Multinomial}(\beta_t)$ を選択．
3. D 個の文書に対してラベル $y_d \sim F(\eta, z_d)(y_d \in 1 \ldots R)$ を選択する．

なお，教師ラベル y が連続量であるときの MedLDA は回帰モデル，離散量であるときは分類モデルに位置づけられる．それぞれについて 2.2.1 節と 2.2.2 節にて後述する．

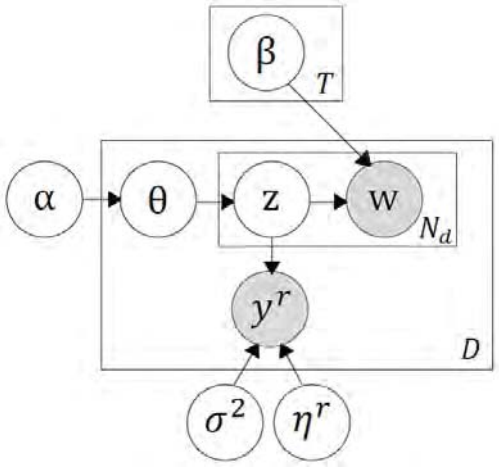

図 1: MedLDA のグラフィカルモデル

2.1.1 回帰問題を想定した最大マージントピックモデル

連続ラベルを持つ文書データを扱う MedLDA Regression (MedLDA-Reg) について説明する．MedLDA-Reg ではマージン最大化を考慮することにより，以下のような最適化問題が定義される．

P1(MedLDA-Reg) :

$$\min_{q(\mathbf{Z},\mathbf{\Theta},\eta),\alpha,\beta,\delta^2,\xi,\xi^\star} \mathcal{L}(q(\mathbf{Z},\mathbf{\Theta},\eta)) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\text{s.t.}\forall \text{d} : \begin{cases} y_d - E[\eta^\top \overline{z}_d] \le \epsilon + \xi_d & [\mu_d] \\ -y_d + E[\eta^\top \overline{z}_d] \le \epsilon + \xi_d^\star & [\mu_d^\star] \\ \xi_d \ge 0 & [v_d] \\ \xi_d^\star \ge 0 & [v_d^\star] \end{cases}$$

$\mu_d, \mu_d^\star, v_d, v_d^\star$ はラグランジュ乗数である．上式中の [ ] はラグランジュ関数を求める際の制約式とラグランジュ乗数の対応を表している．また，各変数は $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_D\}$,$\mathbf{z}_d = \{z_{d,1}, \ldots, z_{d,N_d}\}$, $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_D\}$, $\mathbf{Y} = \{y_1, \ldots, y_D\}$, $\mathbf{W} = \{w_1, \ldots, w_D\}$, $\mathbf{w}_d = \{\mathbf{w}_{d,1}, \ldots, \mathbf{w}_{d,N_d}\}$, $\beta = \{\beta_1, \ldots, \beta_K\}$ を表す．制約式中の $\xi, \xi^\star$ は訓練データの誤差を吸収する程度を示すスラック変数であり，ϵ は許容誤差である．ここからは MedLDA-Reg の更新式の導出を行う．目的関数の $\mathcal{L}$ は

$$\begin{aligned}\mathcal{L}(q(\mathbf{Z},\mathbf{\Theta},\eta)) = &-\mathrm{E}[\log p(\theta, \mathbf{z}, \eta, \mathbf{Y}, \mathbf{W} | \alpha, \beta, \delta^2)] \\ &- \mathcal{H}(q(z), \theta, \eta)\end{aligned} \quad (1)$$

である．$\mathcal{H}$ は変分事後分布 $q(Z, \theta, \eta)$ のエントロピーであり，$\mathcal{H}(q) = -\sum q \log(q)$ である．最適化問題 P1 は一般的に解くことが困難であるため，変分近似を行い

q についての独立性を与える．

$$q(Z,\Theta,\eta|\gamma,\phi)=q(\eta)\prod_{d=1}^{D}q(\theta_d|\gamma_d)\prod_{n=1}^{N}q(z_{d,n}|\phi_{d,n}) \quad (2)$$

ここで，γ_d および $\phi_{d,n}$ は変分パラメータであり，γ_d はディリクレ分布パラメータの K 次元のベクトル，$\phi_{d,n}$ は K トピックの多項分布である．そして変分 EM アルゴリズムを行い，各パラメータを最適化する．変分 EM アルゴリズムでは次の 2 ステップを繰り返し行う．

1. **E-Step**：潜在変数の事後分布を推定
2. **M-Step**：未知変数を推定

更新式の導出では変分下限を最大化する各パラメータを求める．また，最適化問題 P1 の制約式を目的関数に組み込み，ラグランジュ関数 L^r を定義する．

$$\begin{aligned}L^r=&\mathcal{L}(q)+C\sum_{d=1}^{D}(\xi_d,\xi_d^\star)-\sum_{d=1}^{D}\mu_d(\epsilon+\xi_d-y_d+E[\eta^\top\overline{Z}_d])\\&-\sum_{d=1}^{D}(\mu_d^\star(\epsilon+\xi_d^\star+y_d-E[\eta^\top\overline{Z}_d])+v_d\xi_d+v_D^\star\xi_d^\star)\\&-\sum_{d=1}^{D}\sum_{n=1}^{N}c_{d,n}(\sum_{t=1}^{K}\phi_{d,n,t}-1)\end{aligned} \quad (3)$$

この L^r を各パラメータに関して最適化することにより更新式を得る．

E-Step：

- γ に関して L^r を最適化：γ は α と ϕ から決定する．

$$\gamma_d \leftarrow \alpha+\sum_{n=1}^{N}\phi_{d,n} \quad (4)$$

- ϕ に関して L^r を最適化：$\partial L^r/\partial\phi_{d,n}=0$ とし，次式が得られる．

$$\begin{aligned}\phi_{d,i}\propto&\exp(\mathrm{E}[\log\theta\mid\gamma]+\mathrm{E}[\log p(w_{d,n}\mid\beta)]\\&+\frac{y_d}{N_d\delta^2}E[\eta]\\&-\frac{2E[\eta^\top\phi_{d,-i}\eta]+E[\eta\circ\eta]}{2N^2\delta^2}\\&+\frac{E[\eta]}{N}(\mu_d-\mu_d^\star))\end{aligned} \quad (5)$$

なお，$\phi_{d,-i}=\sum_{n\neq i}\phi_{d,n}$ であり，単語 $\phi_{d,i}$ 以外の ϕ の総和を表す．$\eta\circ\eta$ はアダマール積であり，η 同士の各要素の積からなるベクトルである．

- $q(\eta)$ に関して L^r を最適化：A を，各行がベクトル $\overline{Z}_d^\top$ からなる $D\times K$ 行列と定義する．$\partial L^r/\partial q(\eta)=0$ として，次式を得る

$$\begin{aligned}q(\eta)=\frac{p_0(\eta)}{S}\exp\big(&\eta^\top\sum_{\mathrm{d}=1}^{\mathrm{D}}(\mu_\mathrm{d}-\mu_d^\star+\frac{y_d}{\delta^2})E[\overline{Z}_d]\\&-\eta^\top\frac{E[A^\top A]}{2\delta^2}\big)\end{aligned} \quad (6)$$

また，$E[A^\top A]=\sum_{d=1}^{D}E[\overline{Z}_d\overline{Z}_d^\top]$，$E[\overline{Z}_d\overline{Z}_d^\top]=1/N^2(\sum_{n=1}^{N}\sum_{m\neq n}\phi_{d,n}\phi_{d,m}^\top+\sum_{n=1}^{N}\mathrm{diag}\{\phi_{d,n}\})$，$S$ は定数である．得られた $q(\eta)$ を L^r に代入することによって，以下の双対問題が得られる．

$$\max_{\mu,\mu^star}\ -\frac{1}{2}a^\top\Sigma a-\epsilon\sum_{d=1}^{D}(\mu_d+\mu_d^\star)+\sum_{d=1}^{D}y_d(\mu_d-\mu_d^\star) \quad (7)$$

なお，$a=\sum_{d=1}^{D}(\mu_d-\mu_d^\star+y_d/\delta^2)E[\overline{Z}_d]$ であり，この双対問題を SVM-light[1] などのソルバーによって解くことで $q(\eta),\mu,\mu^\star$ を得る．

M-Step：β と δ^2 の更新式は以下の通りである．

- β に関して L^r を最適化：

$$\beta_{k,w}\propto\sum_{d=1}^{D}\sum_{n=1}^{N_d}I(w_{d,n}=w)\phi_{d,n,t} \quad (8)$$

$I(w_{d,n}=w)$ は，単語 n の語彙が w である場合にのみ $\beta_{k,w}$ に加算することを意味する．

- δ^2 に関して L^r を最適化：

$$\delta^2\leftarrow\frac{1}{D}(y^\top y-2y^\top E[A]E[\eta]+E[\eta^\top E[A^\top A]\eta]) \quad (9)$$

なお，$E[\eta^\top E[A^\top A]\eta]=\mathrm{tr}(\mathrm{E[A^\top A]E[\eta\eta^\top]})$ であり，tr は行列の対角成分の和を表す．

2.1.2 分類タスクを想定した最大マージントピックモデル

離散ラベルを持つ文書データを扱う MedLDA Classification (MedLDA-Cla) について説明する．MedLDA-Cla では以下のような最適化問題が定義される．

P2(MedLDA-Cla):

$$\min_{q(\mathbf{Z},\boldsymbol{\Theta}),\alpha,\beta,\xi}\mathcal{L}(q(Z,\Theta))(q)+KL(q(\eta)\|p_0(\eta))+C\sum_{d=1}^{D}(\xi_d)$$

$$\mathrm{s.t.}\forall\mathrm{d}:\begin{cases}y\neq y_d\\E[\eta^\top\triangle\mathrm{f}_d(\mathrm{y})]\geq 1-\xi_\mathrm{d}\\\xi_d\geq 0\end{cases}$$

[1] http://svmlight.joachims.org/

各変数は $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_D\}$，$\mathbf{z}_d = \{z_{d,1}, \ldots, z_{d,N_d}\}$，$\mathbf{\Theta} = \{\theta_1, \ldots, \theta_D\}$，$\mathbf{Y} = \{y_1, \ldots, y_D\}$，$\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$，$\mathbf{w}_d = \{w_{d,1}, \ldots, w_{d,N_d}\}$，$\beta = \{\beta_1, \ldots, \beta_K\}$ を表す．制約式中の ξ は訓練データの誤差を吸収する程度を示すスラック変数であり，文書ごとに設定する．KL は 2 つの確率分布の差異を表すカルバック・ライブラー情報量であり，次式で表される．

$$KL(q(\eta) \parallel p_0(\eta)) = \int q(\eta) \log \frac{q(\eta)}{p_0(\eta)} \tag{10}$$

ここからは MedLDA-Cla の更新式の導出を行う．目的関数において，

$$\begin{aligned} \mathcal{L}(q(Z, \Theta)) = &- E[\log p(\theta, \mathbf{z}, \mathbf{Y}, \mathbf{W} \mid \alpha, \beta, \delta^2)] \\ &- \mathcal{H}(q(Z, \Theta)) \end{aligned} \tag{11}$$

$$\triangle \mathbf{f}_d(y) = \mathbf{f}(y_d, \overline{Z}_d) - \mathbf{f}(y, \overline{Z}_d) \tag{12}$$

である．MedLDA-Reg と同様に最適化問題 P2 に関しても変分近似を行い，q についての条件付独立性を与える．

$$q(\Theta, Z \mid \gamma, \phi) = \prod_{d=1}^{D} q(\theta_d \mid \gamma) \prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n}) \tag{13}$$

そして目的関数に制約式を含めたラグランジュ関数 L^c を次のように定義し，L^c を各パラメータに関して最適化することで更新式を得る．なお，γ, β に関しては MedLDA-Reg と更新式が同じであるため省略する．

$$\begin{aligned} L^c = &\mathcal{L}(q(Z, \Theta)) + KL(q(\eta) \parallel p_0(\eta)) \\ &+ C \sum_{d=1}^{D} \xi_d - \sum_{d=1}^{D} v_d \xi_d \\ &- \sum_{d=1}^{D} \sum_{y \neq y_d} \mu_d(y)(E[\eta^\top \triangle \mathrm{f} d(y)] + \xi_d - 1) \\ &- \sum_{d=1}^{D} \sum_{n=1}^{N_d} c_{d,n} (\sum_{t=1}^{K} \phi_{d,n,t} - 1) \end{aligned} \tag{14}$$

E-Step：

- ϕ に関して L^c を最適化：$\partial L^c / \partial \phi_{d,n}$ とし，次式が得られる．

$$\begin{aligned} \phi_{d,n} \propto &\exp(\mathrm{E}[\log \theta \mid \gamma] + \mathrm{E}[\log p(w_{d,n} \mid \beta)] \\ &+ \frac{1}{N} \sum_{y \neq y_d} \mu_d(y) E[\eta_{y_d} - \eta_y]) \end{aligned} \tag{15}$$

  最初の 2 項は MedLDA-Reg と同様である．

- $q(\eta)$ に関して L^c を最適化：

$$q(\eta) = \frac{1}{Z} p_0(\eta) \exp(\eta^\top \mu_\eta) \tag{16}$$

  ただし，$\mu_\eta = \sum_{d=1}^{D} \sum_{y \neq y_d} \mu_d(y) E[\triangle \mathbf{f}_d(y)]$.

2.2 双対分解

双対分解 (dual decomposition) は，複雑な目的関数を効率的に求める手法である．直接的に求めることが困難な目的関数をいくつかの関数に分割でき，それぞれの関数の最適解が効率的に求まる場合に適用可能である．効率的に解くことができない次の関数を対象として，双対分解の例を示す．

$$\arg\max_{y} \ f(y) + h(y) \tag{17}$$

$\arg\max_y f(y)$，$\arg\max_y h(y)$ は効率的に求まると仮定する．このとき，次の問題は上の問題と同じ意味を持つ．

$$\arg\max_{y,z} \ f(z) + h(y) \tag{18}$$

$$\text{s.t. } \mathrm{y} = \mathrm{z} \tag{19}$$

この問題の解を $L^\star$ とする．そして，この問題に対してラグランジュ緩和を適用する．

$$L(u, y, z) = f(z) + h(y) + u(y - z) \tag{20}$$

u はラグランジュ乗数である．次に $L(u, y, z)$ に関して最大値をとるものを考える．

$$\begin{aligned} L(u) &= \max_{y,z} L(u, y, z) \\ &= \max_{z}(f(z) - uz) + \max_{y}(h(y) + uy) \end{aligned} \tag{21}$$

この関数は $y = z$ の制約を持たないため，最初の問題より広い解空間を持ち，$L^\star \leq L(u)$ が成り立つ．これにより，元の最適化問題の上限を与えている．よって，双対定理により以下が成り立つ．

$$L^\star = \min_u L(u) \tag{22}$$

$\min_u L(u)$ は凸関数であるので，u に関する勾配を求めることができれば，勾配降下法により最適化できる．よって，劣微分の 1 つである d_u は次のように求めることができる．

$$d_u = y^\star - z^\star \tag{23}$$

$$z^\star = \arg\max_{z} f(z) - uz \tag{24}$$

$$y^\star = \arg\max_{y} f(y) - uy \tag{25}$$

そして，勾配法に基づき以下のように u を更新する．

$$u \leftarrow u - \mu(y^\star - z^\star) \tag{26}$$

μ はステップ数である．この更新を繰り返して $L(u)$ を小さくし，$y^\star = z^\star$ となる時が主問題と双対問題の値が一致したときなので，最適解を求めることができる．

3 双対分解を利用したマルチタスク最大マージントピックモデル

3.1 モデルの定義

2.2 節で述べたように，連続値または離散値の付加情報を持つ文書データの解析を行うためには MedLDA を利用すればよい．しかし，MedLDA では連続値と離散値の両方の付加情報を持つ文書データの解析を行うことができない．この問題を解決する為に，我々は双対分解を利用したマルチタスク最大マージントピックモデル (Multi-task MedLDA: MultiMedLDA) を提案する．MultiMedLDA では複数種類のラベルが付与された文書に対して適用可能なモデルであり，MedLDA を双対分解を利用して拡張している．以下に MultiMedLDA の生成過程を示す．

1. 文書 $d(d \in 1, \ldots, D)$ に対して，$\theta \sim \mathrm{Dirichlet}(\alpha)$ を選択．
2. 文書 d 内の N_d 個の単語 $w_{d,n}(n \in 1, \ldots, N_d)$ に対して，
   (a) トピック $z_{d,n} \sim \mathrm{Multinomial}(\theta_d)$ を選択．
   (b) 単語 $w_{d,n} \sim \mathrm{Multinomial}(\beta_k)$ を選択．
3. D 個の文書に対して，連続ラベル $y_d^r \sim F(\eta^r, z_d)$，離散ラベル $y_d^c \sim F(\eta^c, z_d)$ を選択．

なお，η^r, η^c は重み係数である．
MultiMedLDA のグラフィカルモデルを図 2 に示す．

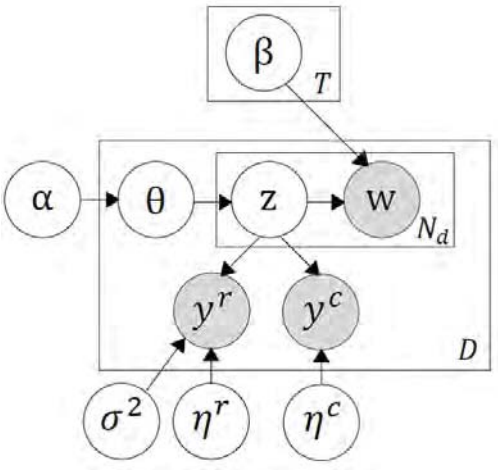

図 2: MultiMedLDA のグラフィカルモデル

3.2 モデルの推定

連続ラベル $y^r \in R$ と離散ラベル $y^c \in \{1, \ldots, M\}$ が各文書に付加されている，データセットについて考える．このとき，MedLDA-Reg の最適化問題と MedLDA-Cla の最適化問題を統合することによって，以下のような最適化問題を定義することができる．なお，目的関数第 1，2 項が回帰タスクの目的関数，目的関数第 3，4 項が分類タスクの目的関数である．同様に制約式第 1，2，3 行が回帰タスクの制約式，制約式第 4，5 行が分類タスクの制約式である．第 6 行は双対分解のための制約である．

$$\mathrm{P3(MultiMedLDA)}: \min_{q(Z^r,\Theta^r,\eta^r),q(Z^c,\Theta^c,\eta^c),\alpha,\beta,\delta^2,\xi^r,\xi^{r\star},\xi^c} \mathcal{L}^r(q(Z^r,\Theta^r,\eta^r)) + C^r \sum_{d=1}^{D}(\xi_d^r + \xi_d^{r\star}) + \mathcal{L}^c(q(Z^c,\Theta^c,\eta^c)) + C^c \sum_{d=1}^{D} \xi_d^c$$

$$\text{subject to } \forall d \begin{cases} y_d^r - E[{\eta^r}^\top \overline{z}_d] \le \epsilon + \xi_d^r \\ -y_d^r + E[{\eta^r}^\top \overline{z}_d] \le \epsilon + \xi_d^{r\star} \\ \xi_d \ge 0, \xi_d^\star \ge 0 \\ y^c \ne y_d^c : E[{\eta^c}^\top \triangle \mathrm{f}_d(\mathbf{y^c})] \ge 1 - \xi_d^c \\ \xi_d^c \ge 0 \\ \phi^r = \phi^c \end{cases}$$

各変数は $Z^r = \{\mathbf{z}_1^r, \ldots, \mathbf{z}_D^r\}$，$\mathbf{z}^r = \{z_{d,1}^r \ldots, z_{d,N_d}^r\}$，$\mathbf{Z}^c = \{\mathbf{z}_1^c, \ldots, \mathbf{z}_D^c\}$，$\mathbf{z}^c = \{z_{d,1}^c \ldots, z_{d,N_d}^c\}$，$E[z_{d,n}^r] = \phi_{d,n}^r$，$E[z_{d,n}^c] = \phi_{d,n}^c$，$\mathbf{\Theta} = \{\theta_1, \ldots, \theta_D\}$ である．そして，$\xi^r, \xi^{r\star}, \xi^c$ はそれぞれ訓練データの誤差を吸収する程度を示すスラック変数，ϵ は許容誤差である．また MedLDA-Cla と同様に $\triangle \mathbf{f}_d(\mathbf{y^c}) = \mathbf{f}(y_d^c, \overline{z}_d) - \mathbf{f}(y, \overline{z}_d)$ である．$\mathbf{f}(y^c, \overline{z}_d)$ は $(y^c - 1)K + 1$ から $y^c K$ の要素がベクトル $\overline{z}$ であり，他の要素が 0 であるような特徴ベクトルであり，$\overline{z}_d \leftarrow (1/N_d) \sum_{n=1}^{N} z_{d,n}$ である．

以下では回帰タスクに関する目的関数第 1，2 項を $\mathcal{L}(R)$，分類タスクに関する目的関数第 3，4 項を $\mathcal{L}(C)$ とする．この最適化問題に対してラグランジュ緩和を行い，次の最適化問題を得る．なお，簡単のため制約式は省略している．

$$L(U, \phi^r, \phi^c) = \mathcal{L}(R) + \mathcal{L}(C) + U(\phi^c - \phi^r)$$

U はラグランジュ乗数である．次に $L(U, \phi^r, \phi^c)$ を最小化する ϕ^r, ϕ^c を考える．

$$\begin{aligned} L(U) &= \min_{\phi^r,\phi^c} L(U, \phi^r, \phi^c) \\ &= \min_{\phi^r}(\mathcal{L}(R) - U\phi^r) + \min_{\phi^c}(\mathcal{L}(C) + U\phi^c) \end{aligned} \tag{27}$$

この関数には最適化問題 P3 の制約式第 6 行にある $\phi^r = \phi^c$ が考慮されていないので，より広い解空間を持つ．これにより，$L^\star \geq L(U)$ が成り立つので，最適化問題 P3 の下限を与えている．また，双対定理より $L^\star = \max L(U)$ が成り立つ．よって，$L(U)$ の劣微分の 1 つである d_U，および $\phi^{r\star}$，$\phi^{c\star}$，ラグランジュ乗数 U は以下のようになる．

$$d_U = \phi^{c\star} - \phi^{r\star} \tag{28}$$

$$\phi^{r\star} = \arg\min_{\phi^r} \mathcal{L}(R) - U\phi^r \tag{29}$$

$$\phi^{c\star} = \arg\min_{\phi^c} \mathcal{L}(C) - U\phi^c \tag{30}$$

$$U \leftarrow U - \mu(\phi^{c\star} - \phi^{r\star}) \tag{31}$$

なお，μ はステップ幅であり，本研究では反復回 S の逆数を用いている．回帰タスクと分類タスクで潜在トピック ϕ^r，ϕ^c の推定を行った後，この更新を繰り返すことで下限 $L(U)$ の最大化を行う．そして $\phi^{r\star} = \phi^{c\star}$ なった時が主問題と双対問題の値が一致したときなので最適解に到達したことが保証される．

これにより得られた $\phi^{r\star}$，$\phi^{c\star}$ をそれぞれの最適化問題に与えなおすことによって，もう一方の影響を考慮した潜在トピックの推定が可能となる．なお，Multi-MedLDA の最適化問題は，MedLDA とは異なり制約式に $\phi^r = \phi^c$ が追加される．これにより，$-U\phi^r$，$U\phi^c$ の項が偏微分をした後にも残る．よって ϕ^r，ϕ^c の更新式は以下のようになる．

$$\begin{aligned}\phi^r_{d,n} \propto \exp\Big(&\mathrm{E}[\log\theta \mid \gamma] + \mathrm{E}[\log p(w_{d,n} \mid \beta)] \\ &- \frac{2E[\eta^\top \phi_{d,-i}\eta] + E[\eta \circ \eta]}{2N^2\delta^2} \\ &+ \frac{E[\eta]}{N}(\mu_d - \mu_d^\star)) + \frac{y_d}{N_d\delta^2}E[\eta] - U\Big)\end{aligned} \tag{32}$$

$$\begin{aligned}\phi^c_{d,n} \propto \exp\Big(&(\mathrm{E}[\log\theta \mid \gamma] + \mathrm{E}[\log p(w_{d,n} \mid \beta)] \\ &+ \frac{1}{N}\sum_{y \neq y_d} \mu_d(y)E[\eta_{y_d} - \eta_y]) + U\Big)\end{aligned} \tag{33}$$

4 実験

4.1 データセット

本研究では，データセットとして東洋経済新報社が発行する会社四季報を使用した．これは，四半期ごとに発表される上場企業 3675 社 (2017 年度新春版) の，企業名をはじめとした上場コード，業種，営業利益，株価，短評などを載せている記事である．2014 年度新春版から 2017 年度新春版までの 13 四半期分のデータを使用した．文書データには短評，離散値として業種，連続値として前年度からの営業利益の変化率を使用した．各企業には上場する際に登録された 32 種類の業種のうち 1 つが選ばれており，その中で登録企業数の多い上位 10 種類 (サービス業，情報・通信業，小売業，卸売業，電気機器，機械，化学，建設業，食料品，輸送用機器) の業種の企業データを使用した．各業種の企業数には偏りが存在するため，1 業種につき 88 企業を無作為に選択している．また，全文書中で 3 文書未満しか出現しない低頻度語を除外している．なお，文書データは MeCab[1] を用いて形態素解析を行い，助詞や接続詞といった機能語を除外している．以上の処理を行ったデータセットの情報を表 1 に示す．

表 1: 四季報データセット

	Shikiho
Number of documents	2657
Number of words	437539
Size of vocabulary	6945
Number of labels	10

また，次節以降の実験を行う為に，データセットの分割を行った．まず各年度 (2014 年，2015 年，2016 年) で文書を 3 分割し，2015 年と 2016 年のデータに関しては企業単位で 20%をテストセットとして確保した．残りは最適パラメータを導出する実験を行う為に，60%の学習セットと 20%の検証セットに分けた．

4.2 実験設定

提案手法である MultiMedLDA，データセットの時系列性を考慮した MultiMedLDA-Seq，既存手法である MedLDA-Reg の 3 つのモデルで最適パラメータを導出する実験を行った．実験で使用するデータセットは，全ての企業に対して 3 年分の時系列データとなっており，MultiMedLDA-Seq では前年の学習で得られた β と η を初期状態として与えている．ただし，初年度である 2014 年のデータは初期状態を乱数によって定めている．ハイパーパラメータは $\alpha = 0.1$，損失パラメータは $l = 1$ に設定した．また，提案した 2 手法の学習において，MedLDA-Reg と MedLDA-Cla の特徴を活かすため，双方の学習結果を影響させない burn-in period を 5 回目の反復までに設定している．これにより，回帰タスクと分類タスクの特徴を活かした状態で潜在トピックの統合が図られている．学習を停止する条件として，テストセット対数尤度 [8] の変化率が負であることを設けている．

正則化パラメータ C^r，C^c を以下の手順で決定する．$C^r \in \{0.25, 1, 4, 16, 64\}$，$C^c \in \{0.25, 1, 4, 16, 64\}$ の範

[1] http://taku910.github.io/mecab/

囲で値を変化させ，2015 年と 2016 年の検証セットから RMSE を計算する．そして 2015 年と 2016 年の RMSE の平均が最も優れているものを最適な正則化パラメータ C^r，C^c とする．RMSE は 4.3.2 節で説明する．

次に回帰問題における許容誤差 ϵ の決定について述べる．2014 年，2015 年，2016 年の各年度に関して $\epsilon \in \{0.01, 0.1, 1, 2, 4\}$ の範囲で値を変化させ，2015 年と 2016 年の検証セットから RMSE を計算する．そして 2015 年と 2016 年の RMSE の平均が最も優れているものを最適な許容誤差パラメータ ϵ として決定した．

最後に得られた最適パラメータ C^r，C^c，ϵ を設定して，2015 年と 2016 年のテストセットを使用した連続値ラベルの予測を行う．以上の実験をトピック数 $T \in \{15, 20, 25\}$ で行い，提案手法である MultiMedLDA-Seq，MultiMedLDA と既存手法である MedLDA-Reg を比較することによって，提案手法の評価を行う．

4.3 評価尺度

4.3.1 Root Mean Squared Error：RMSE

Root Mean Squared Error(以下 RMSE) はモデルの予測能力を表す指標のひとつである．モデルの予測値と真値から算出される相対的な評価指標である．RMSE は以下の式で表される．

$$RMSE = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(y_d - \hat{y}_d)^2} \tag{34}$$

y_d はモデルの予測値であり，$\hat{y}_d$ は真値である．予測値が真値から離れているほど大きい値をとるため，0 に近いほど優れている．

4.4 予備実験

本実験の前に，提案手法である MultiMedLDA-Seq と MultiMedLDA，既存手法である MedLDA-Reg において正則化パラメータ C^r および C^c，許容誤差 ϵ の最適値を導出する予備実験を行った．本実験用データを除いたデータセットを 4 つのデータセットに分割し，3 セットを学習用，1 セットを検証用とした．そして上述の RMSE で評価することによって，最適なパラメータを決定した．各手法のトピック数 $T \in \{15, 20, 25\}$ での最適値を表 2，表 3，表 4 に示す

表 2: トピック数 T=15 での最適パラメータ

	C^c	C^r	ϵ
MultiMedLDA-Seq	64	4	1
MultiMedLDA	0.25	16	1
MedLDA-Reg	-	0.25	0.1

表 3: トピック数 T=20 での最適パラメータ

	C^c	C^r	ϵ
MultiMedLDA-Seq	16	4	0.01
MultiMedLDA	64	1	0.01
MedLDA-Reg	-	4	0.01

表 4: トピック数 T=25 での最適パラメータ

	C^c	C^r	ϵ
MultiMedLDA-Seq	16	64	1
MultiMedLDA	4	0.25	0.01
MedLDA-Reg	-	0.25	0.01

4.5 連続値ラベルの予測実験および考察

4.4 節で導出した C^r，C^c，ϵ を用いて各モデルの連続ラベル予測性能比較実験を行った．評価指標は RMSE を用いている．結果は図 3 のようになった．図 3 から，全てのトピック数に関して提案手法である MultiMedLDA-Seq および MultiMedLDA が MedLDA-Reg よりも RMSE で改善された結果を示していることがわかる．これは，提案手法は連続値ラベルと離散値ラベルの両方を考慮しているので，より精度の高い連続値ラベルの予測が行えていると考えられる．また，本研究における実験では MedLDA の原著論文で使用されていたデータセットよりも小規模なデータセットを使用している．このため，単一のラベルのみを考慮した MedLDA-Reg では文書の潜在トピック構造をうまく解析できていないと考えられる．それに対して MultiMedLDA では複数種類のラベルを付加情報として用いているので，文書データによる潜在トピック構造の推定を，付加情報がうまく補完していると推測する．更に，MultiMedLDA-Seq は時系列性を考慮しない MultiMedLDA よりも RMSE で改善された結果を示している．これは，MultiMedLDA-Seq がデータセットの時系列性を捉えることによって，連続値ラベルの推定に適したトピックの構築が可能になっていると考えられる．最後に，RMSE が最も低くなるトピック数は，既存手法では $T = 25$ であるのに対して，提案手法では $T = 15$ である．このことから，既存の単一の回帰問題を想定した MedLDA では回帰問題を一部のトピックでのみ対応し，残りのトピックは単語の事後確率最大化にのみ寄与するような分担が起こっていると考えられる．それに対して，提案する MultiMedLDA では双対分解によって回帰と分類と事後確率最大化を共通のトピックで実現していると考え

られる．以上より，提案手法は，複数種類のラベルを持つ文書データに対して，既存手法である MedLDA-Reg よりも優れた手法であるといえる．

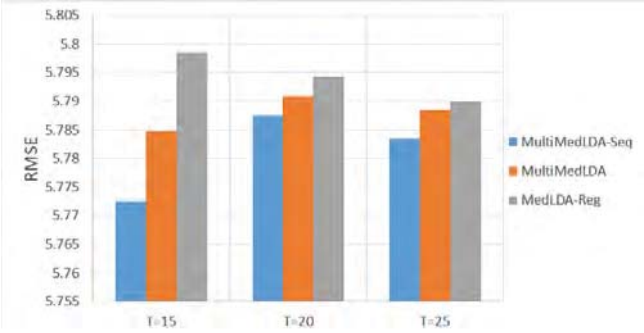

図 3: 四季報データセットでの各手法の RMSE

5 おわりに

本論文では，複数種類のラベルが付加されている文書に対して適用することができるモデルである MultiMedLDA を提案した．MultiMedLDA は既存手法である MedLDA とは異なり，複数の分類問題や回帰問題を伴うマルチタスク問題を最適化することが可能である．そこで，多くの付加情報を持つ会社四季報の企業データにたいしてこのモデルを適用した．実験では連続値ラベルの推定を行い，各年度の RMSE の平均値において MedLDA よりも優れた結果を示したことから MultiMedLDA の優位性を示すことができた．また，データセットの時系列性を考慮することによって更なるモデルの改善が可能であることも示せた．

モデルの改善点としては，分類問題と回帰問題でそれぞれの性能が発揮されるトピック数が異なる場合に，どちらかのトピックを選択しなければならない問題が挙げられる．この問題を解決し，それぞれに適切なトピック数を設定することによって，MultiMedLDA の予測性能を更に高めることが期待できる．

今後の展望として，より多くの付加情報を考慮することが挙げられる．MultiMedLDA は連続値と離散値の複数種類の付加情報を持つ文書データに対応したモデルであり，本論文ではそれぞれ 1 つずつ持つ四季報データセットを使用した．より多くの連続値ラベルと離散値ラベルを考慮することで，更なる予測精度の向上が期待される．また，単語が属す話題を表す潜在トピックの推定に加え，文章の修辞的機能を解析する研究 [9] も存在する．このような異なるモデルを組み合わせることで，モデルの精度を向上させることも期待できる．最後に，経済・金融業界において大規模データの解析に対する関心は高まっており，今回扱った会社四季報の企業活動データだけでなく，株価，為替，債権などより複雑な要因を持ち，多くの付加情報を持ちうるデータセットに対して MultiMedLDA を適用することで，モデルの性能を示すことができるかもしれない．

謝辞 本研究を行うにあたり，有益な助言を頂いた神戸大学大学院経済学研究科の羽森茂之教授と金京拓司教授に感謝する．本研究の一部は科学研究費補助金基盤研究 (B)(15H02703) の援助による．

参考文献

[1] David M Blei, Andrew Y Ng, and Michael I Jordan.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, (2003)

[2] David M Blei and Jon D. McAuliffe.: Supervised topic models, *Advances in neural information processing systems*, pp. 121–128, (2008)

[3] Jun Zhu, Amr Ahmed, and Eric P Xing.: MedLDA: Maximum Margin Supervised Topic Models, *Journal of Machine Learning Reasearch*, Vol. 13, pp. 2237–2278, (2012)

[4] Vladimir Vapnik.: *Statistical Learning Theory*, John Wiley and Sons, New York, (1998)

[5] Alex J Smola and Bernhard Schölkopf.: A tutorial on support vector regression, *Statistics and Computing*, Vol. 14, pp. 199–222, (2004)

[6] Tommi Jaakkola, Marina Meila, and Tony Jebara.: Maximum entropy discrimination, *Neural Information Processing Systems*, Vol. 12, pp. 470–476, (1999)

[7] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright.: Optimization for Machine Learning, *Neural Information Processing Systems*, Mit Press, (2012)

[8] Yee Whye Teh, David Newman, and Max Welling.: A Collapsed Variational Baysian Inference Algorithm for Latent Dirichlet Allocation, *Neural Information Processing Systems*, Vol. 6, pp. 1378–1385, (2006)

[9] Bei Chen, Jun Zhum Nan Yang, Tian Tian, Ming Zhou, and Bo Zhang.: Jointly Modeling Topics and Intents with Global Order Structure, *arXiv preprint arXiv*, 1512, 02009, (2015).