潜在トピック空間上でのマルチタスク学習による企業評価テキストデータを用いた財務指標予測

Predicting a financial index by articles: Multi-task learning on latent topic space

茂庭 綾香[1*] 中川 雄太[1] 江口 浩二[1†]
Ayaka Moniwa[1], Yuta Nakagawa[1], Koji Eguchi[1]

[1] 神戸大学 大学院 システム情報学研究科
[1] Graduate School of System Informatics, Kobe University

Abstract: This paper aims to predict a company's financial index by analyzing articles about the company. The authors propose MultiMedLDA, which is one of supervised topic models. MultiMedLDA assumes that each document has two types of labels, discrete value label and continuous one. It models relation between each document and these labels, and predicts an unknown label based on known labels and the documents. Making use of not only documents but also the known labels, it improves prediction accuracy. We evaluated our model with data from the "Japan Company Handbook". Using comments for each company as a document, the type of industry as a discrete value label and the company's ROE (Return On Equity) as a continuous value label, we predicted the ROE in the evaluation.

1 はじめに

企業の今後の業績を予想する際の手掛かりには，過去の業績に関する数値情報の他に，ニュース記事などの文書データが挙げられる．文書データには，事業展開や市場の動向など数値では表しきれない抽象的な情報が含まれている．本稿ではこのことに着目し，文書データをもとに企業の財務指標を予測する課題に取り組む．

文書データから数値を予測する既存のモデルには，教師ありトピックモデル (Supervised topic models)[1] や最大マージントピックモデル（Maximum Entropy Discrimination LDA : MedLDA）[2] が挙げられる．これらはトピックモデルの一種であり，文書を単語の多重集合（Bag-of-words : BoW）として捉える．Bag-of-words は文書中の各語彙の出現頻度にのみ着目した表現形式であり，語順は考慮しない．ここで挙げた既存モデルはいずれも，文書と予測対象数値ラベルの組を一対のデータとし，文書情報のみからラベルを予測している．しかし，企業の財務指標を予測する際には，文書情報以外にも業種や国籍といった付加情報の活用が期待できる．そこで本稿では，それらの付加情報もラベルとして文書に付与し，文書と複数のラベルが組になったデータを扱うマルチタスク最大マージントピックモデル（MultiMedLDA）を提案する．

*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: moniaya@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

2 関連研究

2.1 最大マージントピックモデル

最大マージントピックモデル（Maximum Entropy Discriminated LDA : MedLDA）[2] はトピックモデルの一種である．トピックモデルは文書データを表現するモデルの一種であり，文書一つをそこに含まれる単語の多重集合（Bag-of-Words : BoW）と捉える．それぞれの単語の背後には「トピック」と呼ばれる潜在変数を仮定し，これを推定することで文書の潜在特徴を表現する．すなわち，文書一つはそこに含まれる単語数と同じ個数のトピックの多重集合として表すことができる．

その上で MedLDA では，各文書に数値ラベルが一つずつ付与されていると想定する．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，$\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ を選択する．
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 $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ を選択する．
   (b) 単語 $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{\mathrm{t}})$ $(t = z_{d,n})$ を選択する．
3. D 個の文書に対してラベル $y_d \sim F(\boldsymbol{\eta}, \boldsymbol{z}_d)$ を選択する．

MedLDA における変数同士の関係を表した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，$\boldsymbol{\eta}$ はラベル評価時の各トピックに対する重み係数である．

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 2.1.1 節と 2.1.2 節にて後述する．

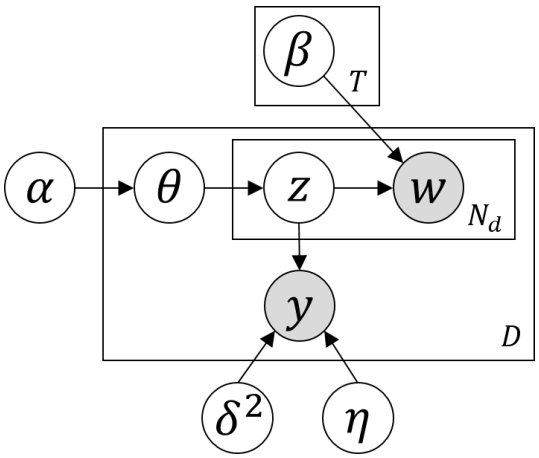

図 1: MedLDA のグラフィカルモデル

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

連続値ラベル $y \in \mathbb{R}$ 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 $y_d|\boldsymbol{z}_d, \boldsymbol{\eta}, \delta^2 \sim \mathcal{N}\left(\boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d, \delta^2\right)$ とし，マージン最大化 [3][4] 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) :

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

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

制約式中の $\xi_d, \xi_d^\star$ 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．定数 $C(>0)$ は正則化パラメータ，$\mu_d, \mu_d^\star, v_d, v_d^\star$ はラグランジュ乗数である．上式右端の [ ] 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．$\mathbb{E}[\cdot]$ は期待値を表す．また，各変数は $\boldsymbol{Z} := \{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_D\}$,$\boldsymbol{z}_d := \{\boldsymbol{z}_{d,1}, \ldots, \boldsymbol{z}_{d,N_d}\}$, $\boldsymbol{\Theta} := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$, $\boldsymbol{y} := \{y_1,\ldots,y_D\}$, $\boldsymbol{W} := \{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_D\}$, $\boldsymbol{w}_d := \{w_{d,1},\ldots,w_{d,N_d}\}$, $\boldsymbol{\beta} := \{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_T\}$, $\boldsymbol{\xi} := \{\xi_1,\ldots,\xi_D\}$, $\boldsymbol{\xi}^\star := \{\xi_1^\star,\ldots,\xi_D^\star\}$ を表す．$\boldsymbol{z}_{d,n}$ は $t = z_{d,n}$ 番目の要素のみ 1，それ以外の要素は 0 となる T 次元の指標ベクトルであり，確率変数 $Z_{d,n}$ のインスタンスである．$\overline{Z}_d := \frac{1}{N_d}\sum_{n=1}^{N_d} Z_{d,n}$，$\bar{\boldsymbol{z}}_d := (1/N_d)\sum_{n=1}^{N_d} \boldsymbol{z}_{d,n}$ 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

$$\mathcal{L}(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})) = -\mathbb{E}_q\left[\log p(\boldsymbol{\Theta},\boldsymbol{Z},\ \boldsymbol{\eta},\boldsymbol{y},\boldsymbol{W}|\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2)\right] - \mathcal{H}\left(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})\right) \quad (1)$$

である．$\mathcal{H}$ は事後分布 $q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})$ 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，変分近似を行いて q についての独立性を仮定する．

$$q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}) = q(\boldsymbol{\eta})\prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)\prod_{n=1}^{N_d} q(\boldsymbol{z}_{d,n}|\boldsymbol{\phi}_{d,n}) \quad (2)$$

ここで，$\boldsymbol{\gamma}_d$ および $\boldsymbol{\phi}_{d,n}$ は変分パラメータであり，$\boldsymbol{\gamma}_d$ はディリクレ分布パラメータの T 次元のベクトル，$\boldsymbol{\phi}_{d,n}$ は T トピックの多項分布パラメータである．$\mathbb{E}[Z_{d,n}] = \boldsymbol{\phi}_{d,n}$，$\mathbb{E}\left[\boldsymbol{\eta}^\top \overline{Z}_d\right] = \mathbb{E}[\boldsymbol{\eta}]^\top (1/N_d)\sum_{n=1}^{N_d}\boldsymbol{\phi}_{d,n}$ が成り立つ．

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．

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

ここで，$c_{d,n}$ は制約 $\sum_{t=1}^{T}\phi_{d,n,t}=1$ に対するラグランジュ乗数である．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．

$$\boldsymbol{\gamma}_d=\boldsymbol{\alpha}+\sum_{n=1}^{N_d}\boldsymbol{\phi}_{d,n}\tag{4}$$

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

$$\begin{aligned}\boldsymbol{\phi}_{d,n}\propto&\exp(\mathbb{E}\left[\log\boldsymbol{\theta}_{\mathrm{d}}\mid\boldsymbol{\gamma}_{\mathrm{d}}\right]+\log p(w_{d,n}\mid\boldsymbol{\beta})\\&+\frac{y_d}{N_d\delta^2}\mathbb{E}\left[\boldsymbol{\eta}\right]\\&-\frac{2\mathbb{E}\left[\boldsymbol{\eta}^\top\boldsymbol{\phi}_{d,-n}\boldsymbol{\eta}\right]+\mathbb{E}\left[\boldsymbol{\eta}\circ\boldsymbol{\eta}\right]}{2N_d^2\delta^2}\\&+\frac{\mathbb{E}\left[\boldsymbol{\eta}\right]}{N_d}(\mu_d-\mu_d^\star))\end{aligned}\tag{5}$$

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

- $q(\boldsymbol{\eta})$ 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 $D\times T$ 行列と定義する．$\partial L^r/\partial q(\boldsymbol{\eta})=0$ として，次式を得る

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

また，$\mathbb{E}\left[A^\top A\right]=\sum_{d=1}^{D}\mathbb{E}\left[\overline{Z}_d\overline{Z}_d^\top\right]$，$\mathbb{E}\left[\overline{Z}_d\overline{Z}_d^\top\right]=1/N_d^2(\sum_{n=1}^{N_d}\sum_{m\neq n}\boldsymbol{\phi}_{d,n}\boldsymbol{\phi}_{d,m}^\top+\sum_{n=1}^{N_d}\mathrm{diag}\{\boldsymbol{\phi}_{d,n}\})$，$X$ は定数である．得られた $q(\boldsymbol{\eta})$ 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．

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

ここで，$\boldsymbol{\mu}:=\{\mu_1,\ldots,\mu_D\}$，$\boldsymbol{\mu}^\star:=\{\mu_1^\star,\ldots,\mu_D^\star\}$ である．E を単位行列とすると，$q(\boldsymbol{\eta})$ の $K\times K$ 共分散行列 $\Sigma=\left(E+1/\delta^2\mathbb{E}\left[A^\intercal A\right]\right)^{-1}$，$\boldsymbol{a}=\sum_{d=1}^{D}(\mu_d-\mu_d^\star+y_d/\delta^2)\mathbb{E}\left[\overline{Z}_d\right]$ 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 $q(\boldsymbol{\eta}),\mu,\mu^\star$ 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：

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

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

- δ^2 に関して L^r を最適化：

$$\delta^2=\frac{1}{D}(\boldsymbol{y}^\top\boldsymbol{y}-2\boldsymbol{y}^\top\mathbb{E}\left[A\right]\mathbb{E}\left[\boldsymbol{\eta}\right]+\mathbb{E}\left[\boldsymbol{\eta}^\top\mathbb{E}\left[A^\top A\right]\boldsymbol{\eta}\right])\tag{9}$$

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

2.1.2 分類問題を想定した最大マージントピックモデル

離散値ラベル $y\in\{1,\ldots,M\}$ 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 $y_d|\boldsymbol{z}_d,\boldsymbol{\eta}=\underset{y\in\{1,\ldots,M\}}{\arg\max}\,\mathbb{E}\left[\boldsymbol{\eta}^\top\mathbf{f}(y,\overline{Z}_d)|\boldsymbol{w}_d,\boldsymbol{\alpha},\boldsymbol{\beta}\right]$ とし，以下のような最適化問題が定義される．

P2(MedLDA-Cla):

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

$$\text{s.t. }\forall d:\begin{cases}\widehat{y}\neq y_d\\\mathbb{E}\left[\boldsymbol{\eta}^\top\Delta\mathbf{f}_d(\widehat{y})\right]\geq 1-\xi_d\\\xi_d\geq 0\end{cases}$$

制約式中の $\widehat{y}$ はラベルの予測値，y_d は真値である．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．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，

$$\begin{aligned}\mathcal{L}(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}))=&-\mathbb{E}_q\left[\log p(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{W}\mid\boldsymbol{\alpha},\boldsymbol{\beta})\right]\\&-\mathcal{H}(q(\boldsymbol{Z},\boldsymbol{\Theta}))+KL(q(\boldsymbol{\eta})\parallel p_0(\boldsymbol{\eta}))\end{aligned}\tag{10}$$

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

$$\Delta \mathbf{f}_d(\widehat{y}) := \mathbf{f}(y_d, \overline{Z}_d) - \mathbf{f}(\widehat{y}, \overline{Z}_d) \tag{11}$$

である．$\mathbf{f}(y, \overline{Z}_d)$ は，$(y-1)T+1$ から yT の要素がベクトル $\overline{Z}_d$ であり他の要素が 0 であるような特徴ベクトルである．

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．

$$KL(q(\boldsymbol{\eta}) \parallel p_0(\boldsymbol{\eta})) = \int q(\boldsymbol{\eta}) \log \frac{q(\boldsymbol{\eta})}{p_0(\boldsymbol{\eta})} \mathrm{d}\boldsymbol{\eta} \tag{12}$$

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．

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

また，$\mathbb{E}[Z_{d,n}] = \boldsymbol{\phi}_{d,n}$，$\mathbb{E}\left[\boldsymbol{\eta}^\top \mathbf{f}(y, \overline{Z}_d)\right] = \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(y, 1/N_d \sum_{n=1}^{N_d} \boldsymbol{\phi}_{d,n})$ である．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．

$$\begin{aligned} L^c = {} & \mathcal{L}(q(\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\eta})) + C \sum_{d=1}^{D} \xi_d \\ & - \sum_{d=1}^{D} v_d \xi_d - \sum_{d=1}^{D} \sum_{\widehat{y} \neq y_d} \mu_d(\widehat{y}) (\mathbb{E}\left[\boldsymbol{\eta}^\top \Delta \mathbf{f} d(\widehat{y})\right] + \xi_d - 1) \\ & - \sum_{d=1}^{D} \sum_{n=1}^{N_d} c_{d,n} (\sum_{t=1}^{T} \phi_{d,n,t} - 1) \end{aligned} \tag{14}$$

E-Step：

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

$$\begin{aligned} \boldsymbol{\phi}_{d,n} \propto & \exp(\mathbb{E}[\log \boldsymbol{\theta}_{\mathrm{d}} \mid \boldsymbol{\gamma}_{\mathrm{d}}] + \log p(w_{d,n} \mid \boldsymbol{\beta}) \\ & + \frac{1}{N_d} \sum_{\widehat{y} \neq y_d} \mu_d(\widehat{y}) \mathbb{E}\left[\boldsymbol{\eta}_{y_d} - \boldsymbol{\eta}_{\widehat{y}}\right]) \end{aligned} \tag{15}$$

  最初の 2 項は MedLDA-Reg と同様である．

- $q(\boldsymbol{\eta})$ に関して L^c を最適化：

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

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

2.2 双対分解

双対分解 (dual decomposition)[5] 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．

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

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

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．

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

ν はステップ幅である．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

2.1 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 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，双対分解を利用して MedLDA を拡張している．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，$\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ 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 $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ を選択する．
   (b) 単語 $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_t)$ $(t = z_{d,n})$ を選択する．
3. D 個の文書に対して，連続値ラベル $y_d^r \sim F(\boldsymbol{\eta}^r, \boldsymbol{z}_d)$，離散値ラベル $y_d^c \sim F(\boldsymbol{\eta}^c, \boldsymbol{z}_d)$ を選択する．

なお，$\boldsymbol{\eta}^r, \boldsymbol{\eta}^c$ 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．

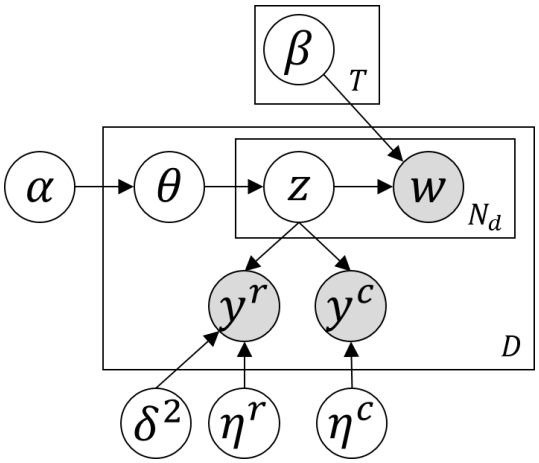

図 2: MultiMedLDA のグラフィカルモデル

3.2 モデルの推定

連続値ラベル $y^r \in \mathbb{R}$ と離散値ラベル $y^c \in \{1,\ldots,M\}$ 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．以下，回帰タスクに関する変数は上付き文字の r，分類タスクに関する変数は上付き文字の c で表す．

$$\mathrm{P3(MultiMedLDA)}: \min_{q(\boldsymbol{Z}^r,\boldsymbol{\Theta}^r,\boldsymbol{\eta}^r),q(\boldsymbol{Z}^c,\boldsymbol{\Theta}^c,\boldsymbol{\eta}^c),\alpha,\beta,\delta^2,\boldsymbol{\xi}^r,\boldsymbol{\xi}^{r\star},\boldsymbol{\xi}^c}$$

$$\mathcal{L}^r(q(\boldsymbol{Z}^r,\boldsymbol{\Theta}^r,\boldsymbol{\eta}^r)) + C^r\sum_{d=1}^{D}(\xi_d^r + \xi_d^{r\star}) + \mathcal{L}^c(q(\boldsymbol{Z}^c,\boldsymbol{\Theta}^c,\boldsymbol{\eta}^c)) + C^c\sum_{d=1}^{D}\xi_d^c$$

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

$\xi^r, \xi^{r\star}, \xi^c$ 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．$\boldsymbol{\phi}_d^r := \{\boldsymbol{\phi}_{d,1}^r,\ldots,\boldsymbol{\phi}_{d,N_d}^r\}$，$\boldsymbol{\Phi}^r := \{\boldsymbol{\phi}_1^r,\ldots,\boldsymbol{\phi}_D^r\}$，$\boldsymbol{\phi}_d^c := \{\boldsymbol{\phi}_{d,1}^c,\ldots,\boldsymbol{\phi}_{d,N_d}^c\}$，$\boldsymbol{\Phi}^c := \{\boldsymbol{\phi}_1^c,\ldots,\boldsymbol{\phi}_D^c\}$ 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．

$$L(\boldsymbol{U},\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c) = \mathcal{L}(R)\boldsymbol{I} + \mathcal{L}(C)\boldsymbol{I} + \boldsymbol{U}\circ(\boldsymbol{\Phi}^c - \boldsymbol{\Phi}^r)$$

$\boldsymbol{I}$ は全ての要素が 1 であるベクトルである．$\boldsymbol{U} := \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_D\}$，$\boldsymbol{u}_d := \{\boldsymbol{u}_{d,1},\ldots,\boldsymbol{u}_{d,N_d}\}$ であり，$\boldsymbol{u}_{d,n}$ は $\boldsymbol{\phi}_{d,n}^r$，$\boldsymbol{\phi}_{d,n}^c$ に対応するラグランジュ乗数を表す．次に $L(\boldsymbol{U},\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c)$ を最小化する $\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c$ を考える．

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

この関数には最適化問題 P3 の制約式第 6 行にある $\boldsymbol{\phi}_d^r = \boldsymbol{\phi}_d^c$ 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，$L^\star \geq L(\boldsymbol{U})$ 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 $L^\star = \max_{\boldsymbol{U}} L(\boldsymbol{U})$ が成り立つ．よって，$L(\boldsymbol{U})$ の劣微分の 1 つである $\boldsymbol{d}_U$，および $\boldsymbol{\Phi}^{r\star}$，$\boldsymbol{\Phi}^{c\star}$，ラグランジュ乗数 $\boldsymbol{U}$ は以下のようになる．

$$\boldsymbol{d}_U = \boldsymbol{\Phi}^{c\star} - \boldsymbol{\Phi}^{r\star} \tag{28}$$

$$\boldsymbol{\Phi}^{r\star} = \arg\min_{\boldsymbol{\Phi}^r} \quad \mathcal{L}(R)\boldsymbol{I} - \boldsymbol{U} \circ \boldsymbol{\Phi}^r \tag{29}$$

$$\boldsymbol{\Phi}^{c\star} = \arg\min_{\boldsymbol{\Phi}^c} \quad \mathcal{L}(C)\boldsymbol{I} + \boldsymbol{U} \circ \boldsymbol{\Phi}^c \tag{30}$$

$$\boldsymbol{U} \leftarrow \boldsymbol{U} - \nu(\boldsymbol{\Phi}^{c\star} - \boldsymbol{\Phi}^{r\star}) \tag{31}$$

なお，ν 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 $\boldsymbol{\Phi}^r$，$\boldsymbol{\Phi}^c$ の推定を行った後，この更新を繰り返すことで下限 $L(\boldsymbol{U})$ の最大化を行う．そして $\boldsymbol{\Phi}^{r\star} = \boldsymbol{\Phi}^{c\star}$ と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 $\boldsymbol{\Phi}^{r\star}$，$\boldsymbol{\Phi}^{c\star}$ 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，MultiMedLDA の最適化問題は，MedLDA とは異なり制約式に $\boldsymbol{\phi}_d^r = \boldsymbol{\phi}_d^c$ が追加される．これにより，$-\boldsymbol{U} \circ \boldsymbol{\Phi}^r$，$\boldsymbol{U} \circ \boldsymbol{\Phi}^c$ の項が偏微分をした後にも残る．よって $\boldsymbol{\phi}$ の更新式は以下のようになる．

$$\begin{aligned}\boldsymbol{\phi}_{d,n}^r \propto \exp\Big(&\mathbb{E}\left[\log \boldsymbol{\theta}_{\mathrm{d}}^{\mathrm{r}} \mid \boldsymbol{\gamma}_{\mathrm{d}}^{\mathrm{r}}\right] + \log p(w_{d,n} \mid \boldsymbol{\beta}^r) \\ &- \frac{2\mathbb{E}\left[\boldsymbol{\eta}^{r\top} \boldsymbol{\phi}_{d,-n}^r \boldsymbol{\eta}^r\right] + \mathbb{E}\left[\boldsymbol{\eta}^r \circ \boldsymbol{\eta}^r\right]}{2N_d^2 \delta^2} \\ &+ \frac{\mathbb{E}\left[\boldsymbol{\eta}^r\right]}{N_d}(\mu_d^r - \mu_d^{r\star})) + \frac{y_d}{N_d \delta^2}\mathbb{E}\left[\boldsymbol{\eta}^r\right] - \boldsymbol{u}_{d,n}\Big)\end{aligned} \tag{32}$$

$$\begin{aligned}\boldsymbol{\phi}_{d,n}^c \propto \exp\Big(&(\mathbb{E}\left[\log \boldsymbol{\theta}_{\mathrm{d}}^{\mathrm{c}} \mid \boldsymbol{\gamma}_{\mathrm{d}}^{\mathrm{c}}\right] + \log p(w_{d,n} \mid \boldsymbol{\beta}^c) \\ &+ \frac{1}{N_d} \sum_{\widehat{y} \neq y_d^c} \mu_d^c(\widehat{y}) \mathbb{E}\left[\boldsymbol{\eta}_{y_d^c}^c - \boldsymbol{\eta}_{\widehat{y}}^c\right]) + \boldsymbol{u}_{d,n}\Big)\end{aligned} \tag{33}$$

4 評価実験

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 [1] を使用した．これは四半期ごとに発表される経済記事であり，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，短評などが載っている．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．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 つが選ばれているが，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 89 企業を無作為に選択している．

表 1: 四季報データセットの概要

年	2014	2015	2016
企業数	890	890	890
総単語数	164931	165272	164801
一企業あたりの単語数	185.3	185.7	185.2
語彙数	2862		
業種（離散値ラベル）の種類	10		

ある年の新春版から秋版までの短評を一つにまとめたものを文書データ，業種を離散値ラベル，文書データの翌年新春版の ROE（Return On Equity，自己資本利益率）を連続値ラベルとして使用した．また，2014 年と 2015 年のいずれかで 3 文書未満に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[2] 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．ROE の値は，年毎に平均が 0，分散が 1 になるよう正規化した．以上の処理を行ったデータセットの情報を表 1 に示す．

4.2 実験設定

連続値ラベルが未知であるという設定の下で，その値を予測する実験を行う．離散値ラベルを考慮しないモデルである MedLDA-Reg をベースラインとし，提案手法の MultiMedLDA と比較する．

学習用データでモデル構築を行い，モデルパラメータ β，η を得る．その後，学習で得られた β，η を用いてテストデータの潜在トピックを推定し連続値ラベルを予測する．テストデータの潜在トピック推定は連続値ラベルを隠した状態で行わなければならないため，MedLDA の実験では教師なしトピックモデルの一種である Latent Dirhlet Allocation（LDA）[6]，MultiMedLDA の実験では MedLDA-Cla を用いて行った．

2016 年のデータをテストデータとし，基本的に 2015 年のデータを学習用データとして用いる．モデル学習時の各パラメータの初期値は乱数で決定する．しかし，この初期値の与え方にも何らかの知見を活用したい．

[1] https://store.toyokeizai.net/cddvd/shikiho_cd/

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

そこで，2014 年のデータで学習して得られた β と η を，2015 年のデータでの学習時の初期値とする条件での実験も行った．この条件での MedLDA と MultiMedLDA をそれぞれ MedLDA-Seq，MultiMedLDA-Seq と呼び，初期値を乱数で決定する条件設定をそれぞれ MedLDA-Rand，MultiMedLDA-Rand と呼ぶことにする．MedLDA-Seq と MultiMedLDA-Seq において，2014 年のデータでの学習時の各パラメータの初期値は乱数で決定した．

ハイパーパラメータは $\alpha_t = 0.1 \quad \forall t$，損失パラメータは $l = 1$，正則化パラメータ $C^r = 1$，許容誤差 $\epsilon = 0.1$，ラグランジュ乗数 $\boldsymbol{U}$ を更新する際の反復回数 $S = 20$ に設定した．また，MultiMedLDA の学習において，MedLDA-Reg と MedLDA-Cla の特徴を活かすため，双方の計算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．学習の反復回数は 100 回とした．トピック数は $T \in \{20, 40, 60, 80, 100\}$ の 5 通り，MultiMedLDA における分類タスクの正則化パラメータは $C^c \in \{0.0625, 0.25, 1, 4, 16\}$ の 5 通りの条件で実験した．

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．

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

$\widehat{y}_d$ はモデルの予測値であり，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．

4.3.2 Mean Absolute Error : MAE

Mean Absolute Error（以下 MAE）も，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．MAE は以下の式で表される．

$$MAE = \frac{1}{D}\sum_{d=1}^{D}|\widehat{y}_d - y_d| \tag{35}$$

$\widehat{y}_d$ はモデルの予測値であり，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．

RMSE は MAE に比べて大きな誤差を重視する性質があり，MAE はその点で安定的な指標である．

4.4 実験結果

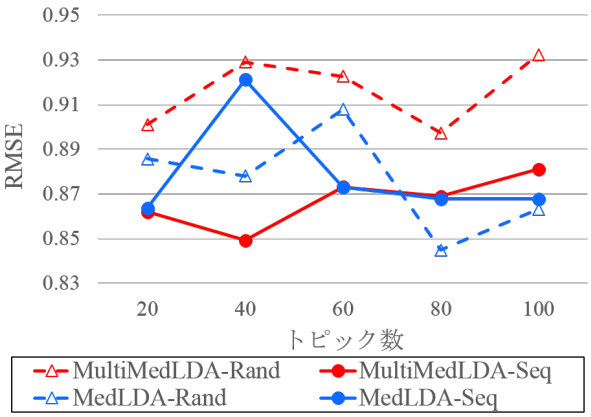

図 3: 各手法のトピックごとの RMSE

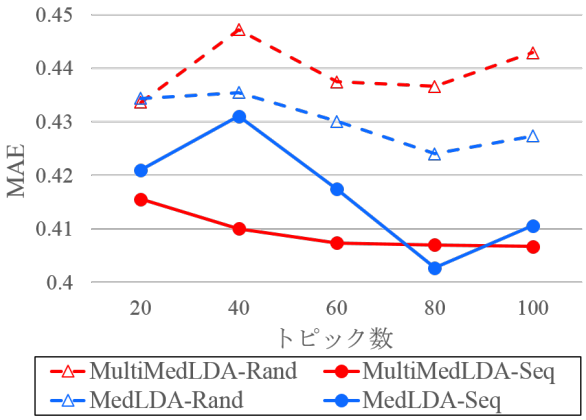

図 4: 各手法のトピックごとの MAE

結果のグラフを図 3，図 4 に示す．MultiMedLDA に関しては各トピック数において C^c の値が異なる 5 通りの実験を行ったが，その中で最も結果が良かった場合のものをプロットしている．この図から分かるように，MedLDA においても MultiMedLDA においても，初期値をランダムに与える-Rand 版よりも，過去の β，η で初期化する-Seq 版の方が高い精度を示した．各トピック数における最良の C^c を選択できれば，実験した 4 手法の内では提案手法の MultiMedLDA-Seq が安定的に最良の性能を示している．一方で MultiMedLDA-Rand は最も悪い結果となっており，既存手法の MedLDA に

比べて提案手法の MultiMedLDA は，初期値の与え方に性能が大きく左右されることが分かった．

5 むすび

本稿では，文書データをもとに企業の財務指標を予測する課題に取り組み，トピックモデルの一種である MultiMedLDA を提案した．このモデルは，双対分解を利用して既存手法の MedLDA を拡張したものである．MultiMedLDA の特徴は，離散値と連続値という 2 種類の数値ラベルを同時に持つ文書を扱うことができる点であり，文書に加えて既知ラベルの情報も考慮しつつ未知ラベルを推定することができる．『会社四季報』のデータを用いて評価実験を行った結果，既知ラベルを想定しない既存手法である MedLDA よりも良い性能を示した．また，特に提案手法に対しては，モデルパラメータを学習する際の初期値の与え方が精度向上に重要であることが分かった．

本稿ではいわゆる closed test を行ったが，今後は交差検証により最適な超パラメータを決定した上でテストを行い，より厳密にモデルの汎化性能を評価したい．また，現時点では一つの文書に離散値ラベルと連続値ラベルが一つずつ付随すると仮定しているが，これを複数個ずつに拡張することでさらなる性能の向上を図る予定である．

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

[2] 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)

[3] Edgar Osuna, Robert Freund, and Federico Girosi.: Support Vector Machines: Training and Applications, *Proceedings SVPR'97*, (1997)

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

[5] S. Sra, S. Nowozin and S. Wright, “Optimization for Machine Learning”, Neural information processing series, MIT Press (2012).

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