

LDA

(latent Dirichlet allocation)

正田 備也

masada@rikkyo.ac.jp

Contents

VB の回の課題のヒント

PLSA の復習

PLSA の問題点

LDA の変分ベイズ法

VB の回の課題のヒント

- ▶ $q(\lambda_1)$ の shape パラメータは α_1 、rate パラメータは β_1
- ▶ このとき、 $\int q(\lambda_1; \alpha_1, \beta_1) \ln p(\lambda_1; a, b) d\lambda_1$ を計算せよ。
- ▶ ヒント 1 : $q(\lambda_1)$ のパラメータ α_1 と β_1 を使って答えを表す。
- ▶ ヒント 2 : $p(\lambda_1; a, b)$ がガンマ分布で、shape パラメータは a 、rate パラメータは b であることも当然使う。
- ▶ ヒント 3 : $\ln p(\lambda_1; a, b)$ 、つまり、shape パラメータが a で rate パラメータが b であるガンマ分布の密度関数の式の対数を使う。

$$\begin{aligned}
\int q(\lambda_1; \alpha_1, \beta_1) \ln p(\lambda_1; a, b) d\lambda_1 &= \int q(\lambda_1; \alpha_1, \beta_1) \ln \left[\frac{b^a}{\Gamma(a)} \lambda_1^{a-1} e^{-b\lambda_1} \right] d\lambda_1 \\
&= \int q(\lambda_1; \alpha_1, \beta_1) (a \ln b) d\lambda_1 - \int q(\lambda_1; \alpha_1, \beta_1) \ln \Gamma(a) d\lambda_1 \\
&\quad + \int q(\lambda_1; \alpha_1, \beta_1) (a-1) \ln \lambda_1 d\lambda_1 + \int q(\lambda_1; \alpha_1, \beta_1) \ln \exp(-b\lambda_1) d\lambda_1 \\
&= a \ln b - \ln \Gamma(a) + (a-1) \int q(\lambda_1; \alpha_1, \beta_1) \ln \lambda_1 d\lambda_1 - b \int q(\lambda_1; \alpha_1, \beta_1) \lambda_1 d\lambda_1 \\
&= a \ln b - \ln \Gamma(a) + (a-1) \mathbb{E}_{q(\lambda_1; \alpha_1, \beta_1)}[\ln \lambda_1] - b \mathbb{E}_{q(\lambda_1; \alpha_1, \beta_1)}[\lambda_1]
\end{aligned} \tag{1}$$

$\mathbb{E}_{q(\lambda_1; \alpha_1, \beta_1)}[\ln \lambda_1]$ は、Wikipedia に以下のように書いてあることを利用する。

$$\mathbb{E}[\ln(X)] = \psi(\alpha) - \ln(\beta)$$

$\mathbb{E}_{q(\lambda_1; \alpha_1, \beta_1)}[\lambda_1]$ は、ガンマ分布の mean。

Contents

VB の回の課題のヒント

PLSA の復習

PLSA の問題点

LDA の変分ベイズ法

PLSA (probabilistic latent semantic analysis)

- ▶ 同じ文書内でも、異なる単語トークンは異なる単語多項分布から生成されうる (=異なるトピックを表現しうる)
- ▶ 文書によって、各トピックの出現確率が異なる
- ▶ PLSA では、単語多項分布をトピック (topic) と呼ぶ
- ▶ PLSA は最もシンプルなトピックモデル
 - ▶ トピックモデルは、単語トークンの “クラスタリング”
 - ▶ 同一文書内の同一単語の異なるトークンは区別されない (bag-of-words)

Notations

- ▶ 語彙集合 $\{1, \dots, W\}$
- ▶ トピック集合 $\{1, \dots, K\}$
 - ▶ 語彙やトピックをその添字と同一視している。
- ▶ 文書集合 $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
- ▶ 文書 \mathbf{x}_i の j 番目のトークンとして現れる単語を、 $x_{i,j}$ という確率変数で表す
- ▶ 文書 \mathbf{x}_i の j 番目の単語 $x_{i,j}$ が表現するトピックを、 $z_{i,j}$ という確率変数で表す
- ▶ $x_{i,j}$ の値は観測されているが、 $z_{i,j}$ の値は観測されていない
 - ▶ つまり、 $z_{i,j}$ は潜在変数。

PLSAにおける同時分布

- ▶ PLSA では、文書 x_i の j 番目のトークンがトピック k を表現し、かつそのトピックを表現するために単語 w が使われる同時確率、つまり $p(x_{i,j} = w, z_{i,j} = k)$ は

$$p(x_{i,j} = w, z_{i,j} = k) = p(z_{i,j} = k)p(x_{i,j} = w|z_{i,j} = k) \quad (2)$$

- ▶ $p(z_{i,j} = k)$ は、文書 x_i の j 番目のトークンが（他のトピックでなく）トピック k を表現する確率
- ▶ $p(x_{i,j} = w|z_{i,j} = k)$ は、文書 x_i の j 番目のトークンがトピック k を表現するとき（他の単語でなく）単語 w が使われる確率
- ▶ さらに、PLSA では以下のように仮定する（次スライド）

PLSAにおいて仮定すること

- ▶ どの j, j' についても $p(z_{i,j} = k) = p(z_{i,j'} = k)$ と仮定
 - ▶ 同じ文書内なら、どの単語トークンであれ、トピック k を表現する確率は、同じ（場所によってトピックの確率が違ったりしない）
 - ▶ そこで、 $p(z_{i,\cdot} = k) = \theta_{i,k}$ とおく
- ▶ どの i, i' と j, j' についても、 $p(x_{i,j} = w | z_{i,j} = k) = p(x_{i',j'} = w | z_{i',j'} = k)$ と仮定
 - ▶ 同じコーパス内なら、どの文書のどの単語トークンであれ、それがトピック k を表現するために使われるならば（条件付き確率の条件の部分）、 k を表現するためにどの単語が使われるかの確率は、同じ
 - ▶ つまり、単語確率分布とトピックが一对一に対応している
 - ▶ そこで、 $p(x_{\cdot,j} = w | z_{\cdot,j} = k) = \phi_{k,w}$ とおく

PLSAにおける観測データの尤度

個々の単語トークンにおけるトピックと単語の同時分布は

$$p(x_{i,j} = w, z_{i,j} = k) = p(z_{i,j} = k)p(x_{i,j} = w|z_{i,j} = k) = \theta_{i,k}\phi_{k,x_{i,j}} \quad (3)$$

潜在変数である $z_{i,j}$ を周辺化

$$p(x_{i,j} = w) = \sum_{z_{i,j}=1}^K p(x_{i,j} = w, z_{i,j} = k) = \sum_{k=1}^K \theta_{i,k}\phi_{k,x_{i,j}} \quad (4)$$

各トークンの独立性の仮定より

$$p(\mathbf{x}_i) = \prod_{j=1}^{n_i} p(x_{i,j}) = \prod_{j=1}^{n_i} \left(\sum_{k=1}^K \theta_{i,k}\phi_{k,x_{i,j}} \right) \quad (5)$$

各文書の独立性の仮定より

$$p(\mathcal{X}) = \prod_{i=1}^N p(\mathbf{x}_i) = \prod_{i=1}^N \prod_{j=1}^{n_i} \left(\sum_{k=1}^K \theta_{i,k}\phi_{k,x_{i,j}} \right) \quad (6)$$

Contents

VB の回の課題のヒント

PLSA の復習

PLSA の問題点

LDA の変分ベイズ法

PLSAの問題点とベイズ化による改良

- ▶ 各文書におけるトピック確率 $\theta_i = (\theta_{i,1}, \dots, \theta_{i,K})$ に関して、異なる文書の間で何の関係性も仮定されていない
 - ▶ θ_i と $\theta_{i'}$ の間に何の関係もない。
- ▶ このことが過学習をもたらすかもしれない
- ▶ そこで、コーパスに属する全文書の θ_i が、同一のディリクレ事前分布 $\text{Dir}(\alpha)$ から draw されると仮定する
- ▶ 他はPLSAのまま
 - ▶ 各トピックの単語確率 ϕ_k についても別のディリクレ分布 $\text{Dir}(\beta)$ を導入できるが、そうしなくてもよい

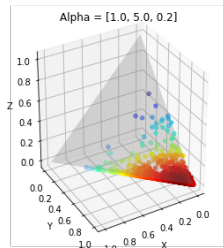
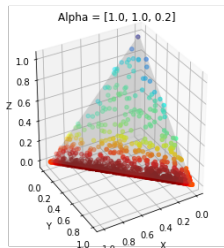
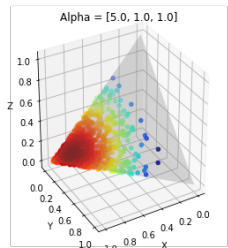
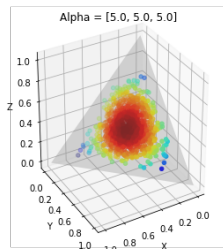
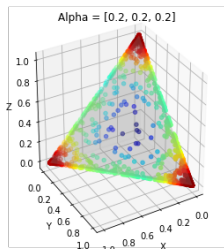
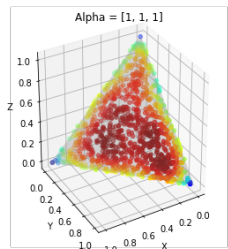


Figure: トピック数が3の場合のディリクレ分布 (ここから引用)

PLSA と LDA の比較

PLSA における x_i の尤度

$$\begin{aligned} p(x_i; \theta_i, \Phi) &= \sum_{z_i} p(x_i, z_i; \theta_i, \Phi) = \sum_{z_i} p(z_i; \theta_i) p(x_i | z_i; \Phi) \\ &= \prod_{j=1}^{n_i} \left(\sum_{z_{i,j}=1}^K p(z_{i,j}; \theta_i) p(x_{i,j} | z_{i,j}; \Phi) \right) = \prod_{j=1}^{n_i} \left(\sum_{k=1}^K \theta_{i,k} \phi_{k,x_{i,j}} \right) \end{aligned}$$

LDA における x_i の尤度 ($p(x_i; \theta_i, \Phi)$ は $p(x_i | \theta_i; \Phi)$ に変わる)

$$\begin{aligned} p(x_i; \Phi, \alpha) &= \int p(\theta_i; \alpha) p(x_i | \theta_i; \Phi) d\theta_i \\ &= \int \sum_{z_i} p(\theta_i; \alpha) p(z_i | \theta_i) p(x_i | z_i; \Phi) d\theta_i \\ &= \int \left(\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1} \right) \prod_{j=1}^{n_i} \left(\sum_{k=1}^K \theta_{i,k} \phi_{k,x_{i,j}} \right) d\theta_i \quad (7) \end{aligned}$$

Contents

VB の回の課題のヒント

PLSA の復習

PLSA の問題点

LDA の変分ベイズ法

LDA の変分ベイズ法

Jensen の不等式を適用して ELBO を求める

$$\begin{aligned}\ln p(\mathbf{x}_i; \Phi, \alpha) &= \ln \int \sum_{\mathbf{z}_i} p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) p(\mathbf{x}_i | \mathbf{z}_i; \Phi) d\boldsymbol{\theta}_i \\ &= \ln \int \sum_{\mathbf{z}_i} q(\mathbf{z}_i, \boldsymbol{\theta}_i) \frac{p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) p(\mathbf{x}_i | \mathbf{z}_i; \Phi)}{q(\mathbf{z}_i, \boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\ &\geq \int \sum_{\mathbf{z}_i} q(\mathbf{z}_i, \boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) p(\mathbf{x}_i | \mathbf{z}_i; \Phi)}{q(\mathbf{z}_i, \boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i\end{aligned}\tag{8}$$

以下、 $q(\mathbf{z}_i, \boldsymbol{\theta}_i) = q(\mathbf{z}_i)q(\boldsymbol{\theta}_i)$ と factorize すると仮定する。

$q(\boldsymbol{\theta}_i)$ を求める

$q(\mathbf{z}_i)$ を固定する。

$$\begin{aligned}\ln p(\mathbf{x}_i; \Phi, \alpha) &\geq \int \sum_{\mathbf{z}_i} q(\mathbf{z}_i) q(\boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) p(\mathbf{x}_i | \mathbf{z}_i; \Phi)}{q(\mathbf{z}_i) q(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\ &= \int q(\boldsymbol{\theta}_i) \left[\sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) \right] d\boldsymbol{\theta}_i - \int q(\boldsymbol{\theta}_i) \ln q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + \text{const.} \\ &= -D_{\text{KL}}(q(\boldsymbol{\theta}_i) \parallel \frac{1}{Z} \exp \left[\sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) \right]) + \text{const.}\end{aligned}\tag{9}$$

以上より、 $q(\boldsymbol{\theta}_i) \propto \exp \left[\sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) \right]$ のとき、ELBO は最大。つまり、 $q(\boldsymbol{\theta}_i) \propto p(\boldsymbol{\theta}_i; \alpha) \exp \left[\sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\mathbf{z}_i | \boldsymbol{\theta}_i) \right]$ のとき、ELBO は最大。

$$\begin{aligned}
\sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\mathbf{z}_i | \boldsymbol{\theta}_i) &= \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln \prod_{j=1}^{n_i} \theta_{i, z_{i,j}} = \sum_{j=1}^{n_i} \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln \theta_{i, z_{i,j}} \\
&= \sum_{j=1}^{n_i} \sum_{z_{i,j}=1}^K q(z_{i,j}) \ln \theta_{i, z_{i,j}} = \sum_{k=1}^K \left(\sum_{j=1}^{n_i} q(z_{i,j} = k) \right) \ln \theta_{i,k} = \sum_{k=1}^K n_{i,k} \ln \theta_{i,k} \quad (10)
\end{aligned}$$

ただし、 $n_{i,k} \equiv \sum_{j=1}^{n_i} q(z_{i,j} = k)$ と定義した。よって

$$\begin{aligned}
q(\boldsymbol{\theta}_i) &\propto \prod_{k=1}^K \theta_{i,k}^{\alpha_k - 1} \times \exp \left[\sum_{k=1}^K n_{i,k} \ln \theta_{i,k} \right] \\
&= \prod_{k=1}^K \theta_{i,k}^{\alpha_k + n_{i,k} - 1} \quad (11)
\end{aligned}$$

これは、変分事後分布 $q(\boldsymbol{\theta}_i)$ がディリクレ分布であることを意味する。

変分ディリクレ事後分布 $q(\boldsymbol{\theta}_i)$ のパラメータを ζ_i とすると、 $\zeta_{i,k} = \alpha_k + n_{i,k}$ が成り立つ。

$q(z_i)$ を求める

今度は $q(\theta_i)$ を固定する。

$$\begin{aligned}\ln p(\mathbf{x}_i; \Phi, \alpha) &\geq \int \sum_{z_i} q(z_i) q(\theta_i) \ln \frac{p(\theta_i; \alpha) p(z_i | \theta_i) p(\mathbf{x}_i | z_i; \Phi)}{q(z_i) q(\theta_i)} d\theta_i \\ &= \sum_{z_i} q(z_i) \left[\ln p(\mathbf{x}_i | z_i; \Phi) + \int q(\theta_i) \ln p(z_i | \theta_i) d\theta_i \right] - \sum_{z_i} q(z_i) \ln q(z_i) + \text{const.} \\ &= -D_{\text{KL}}(q(z_i) \parallel \frac{1}{Z} \exp \left[\ln p(\mathbf{x}_i | z_i; \Phi) + \int q(\theta_i) \ln p(z_i | \theta_i) d\theta_i \right]) + \text{const.}\end{aligned}\tag{12}$$

以上より、 $q(z_i) \propto p(\mathbf{x}_i | z_i; \Phi) \exp \left[\int q(\theta_i) \ln p(z_i | \theta_i) d\theta_i \right]$ のとき、ELBO は最大。

変分ディリクレ事後分布 $q(\boldsymbol{\theta}_i)$ のパラメータが ζ_i であることを使うと、

$$\begin{aligned} \int q(\boldsymbol{\theta}_i; \zeta_i) \ln p(\mathbf{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i &= \int q(\boldsymbol{\theta}_i; \zeta_i) \ln \prod_{j=1}^{n_i} \theta_{i,z_{i,j}} d\boldsymbol{\theta}_i = \sum_{j=1}^{n_i} \int q(\boldsymbol{\theta}_i; \zeta_i) \ln \theta_{i,z_{i,j}} d\boldsymbol{\theta}_i \\ &= \sum_{j=1}^{n_i} \left\{ \psi(\zeta_{i,z_{i,j}}) - \psi\left(\sum_k \zeta_{i,k}\right) \right\} = \sum_{j=1}^{n_i} \psi(\zeta_{i,z_{i,j}}) + \text{const.} \end{aligned} \quad (13)$$

よって、

$$\begin{aligned} q(\mathbf{z}_i) &\propto p(\mathbf{x}_i | \mathbf{z}_i; \Phi) \exp \left[\int q(\boldsymbol{\theta}_i) \ln p(\mathbf{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \right] = \prod_{j=1}^{n_i} \phi_{z_{i,j}, x_{i,j}} \times \exp \left(\sum_{j=1}^{n_i} \psi(\zeta_{i,z_{i,j}}) \right) \\ &= \prod_{j=1}^{n_i} \phi_{z_{i,j}, x_{i,j}} \times \prod_{j=1}^{n_i} \exp(\psi(\zeta_{i,z_{i,j}})) = \prod_{j=1}^{n_i} \phi_{z_{i,j}, x_{i,j}} \exp(\psi(\zeta_{i,z_{i,j}})) \end{aligned} \quad (14)$$

つまり、

$$q(z_{i,j} = k) = \frac{\phi_{k, x_{i,j}} \exp(\psi(\zeta_{i,k}))}{\sum_{l=1}^K \phi_{l, x_{i,j}} \exp(\psi(\zeta_{i,l}))} \quad (15)$$

変分事後分布を使ってELBOを書き下す

$$\begin{aligned}\ln p(\mathbf{x}_i; \Phi, \alpha) &\geq \int \sum_{\mathbf{z}_i} q(\mathbf{z}_i) q(\boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \alpha) p(\mathbf{z}_i | \boldsymbol{\theta}_i) p(\mathbf{x}_i | \mathbf{z}_i; \Phi)}{q(\mathbf{z}_i) q(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\ &= \int \sum_{\mathbf{z}_i} q(\mathbf{z}_i) q(\boldsymbol{\theta}_i) \ln p(\mathbf{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\mathbf{x}_i | \mathbf{z}_i; \Phi) \\ &\quad - D_{\text{KL}}(q(\boldsymbol{\theta}_i) \parallel p(\boldsymbol{\theta}_i; \alpha)) - \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln q(\mathbf{z}_i)\end{aligned}\tag{16}$$

式 (16) の右辺の最初の項を計算してみる。

$$\begin{aligned}\int \sum_{\mathbf{z}_i} q(\mathbf{z}_i) q(\boldsymbol{\theta}_i) \ln p(\mathbf{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i &= \sum_{j=1}^{n_i} \sum_{\mathbf{z}_{i,j}=1}^K q(\mathbf{z}_{i,j}) \int q(\boldsymbol{\theta}_i) \ln \theta_{i,\mathbf{z}_{i,j}} d\boldsymbol{\theta}_i \\ &= \sum_{k=1}^K \left(\sum_{j=1}^{n_i} q(\mathbf{z}_{i,j} = k) \right) \left(\psi(\zeta_{i,k}) - \psi\left(\sum_l \zeta_{i,l}\right) \right)\end{aligned}\tag{17}$$

式 (16) の右辺の 2 番目の項を計算してみる。

$$\sum_{\mathbf{z}_i} q(\mathbf{z}_i) \ln p(\mathbf{x}_i | \mathbf{z}_i; \Phi) = \sum_{j=1}^{n_i} \sum_{\mathbf{z}_{i,j}=1}^K q(z_{i,j}) \ln \phi_{z_{i,j}, x_{i,j}} = \sum_{j=1}^{n_i} \sum_{k=1}^K q(z_{i,j} = k) \ln \phi_{k, x_{i,j}} \quad (18)$$

トピック単語確率 Φ は、この項の全文書についての和 $\sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^K q(z_{i,j} = k) \ln \phi_{k, x_{i,j}}$ を最大化することで求めることができる。(ELBO の中で Φ を含むのはこの項だけだから。) 全文書の ELBO の和を \mathcal{L} と書くことにする。

$\sum_{w=1}^W \phi_{k,w} = 1$ が満たされなければならないので、ラグランジュの未定乗数法を使えば、

$$\frac{\partial \mathcal{L}}{\partial \phi_{k,w}} + \frac{\partial}{\partial \phi_{k,w}} \lambda_k \left(1 - \sum_{w=1}^W \phi_{k,w} \right) = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} q(z_{i,j} = k) \delta(x_{i,j} = w)}{\phi_{k,w}} - \lambda_k \quad (19)$$

$$\therefore \phi_{k,w} = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} q(z_{i,j} = k) \delta(x_{i,j} = w)}{\sum_{i=1}^N \sum_{j=1}^{n_i} q(z_{i,j} = k)} \quad (20)$$

LDAの変分ベイズ法のまとめ

以下の更新を繰り返し実行する。

$$q(z_{i,j} = k) \leftarrow \frac{\phi_{k,x_{i,j}} \exp(\psi(\zeta_{i,k}))}{\sum_{l=1}^K \phi_{l,x_{i,j}} \exp(\psi(\zeta_{i,l}))} \quad (21)$$

$$\zeta_{i,k} \leftarrow \alpha_k + \sum_{j=1}^{n_i} q(z_{i,j} = k) \quad (22)$$

$$\phi_{k,w} \leftarrow \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} q(z_{i,j} = k) \delta(x_{i,j} = w)}{\sum_{i=1}^N \sum_{j=1}^{n_i} q(z_{i,j} = k)} \quad (23)$$

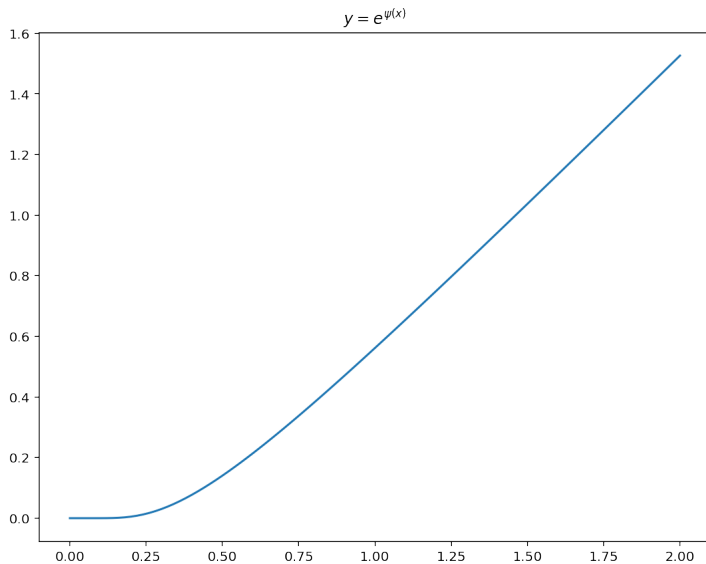


Figure: $y = e^{\psi(x)}$ のグラフ