IBM research

Tokyo Research Laboratory

目的変数が範囲で与えられる回帰問題に対するEM法
Regression with Intervals

鹿島 久嗣、山崎 一孝 （IBM）
西郷 浩人（Max Planck Institute）
猪口 明博（大阪大学）

要旨：
少し変わった回帰問題に対する、少し新しい解法を与えました

- 少し変わった回帰問題： 出力が範囲で与えられるような回帰問題
- 新しい解法： EMアルゴリズムに基づくモデル推定法

問題の説明

手法の提案

正当化

実験

出力が範囲で与えられるような回帰問題を考えます

- 回帰問題とは、入力ベクトルと実数値出力の関係を推定する問題
- 従来の回帰問題は、訓練データの出力が「点（y）」で与えられる
  - xからyを予測するモデルをつくりたい

入力 **x**				出力 y
x_1	x_2	x_3	x_4	
2	5	12	6	13
...	...	...	...	...
5	3	9	10	7

点で与えられる

- 今回考える回帰では、**訓練データの出力が「範囲（l, r）」で与えられる**
  - （目的は普通の回帰と同じ）**x**からyを予測するモデルをつくりたい

入力 **x**				出力	
x_1	x_2	x_3	x_4	l	r
2	5	12	6	11	14
...	...	...	...		
5	3	9	10	6	∞

範囲で与えられる

いくつかの実問題の例

- 機会損失がある場合の売り上げ予測
  - 卸した数以上には売れないので、観測された売り上げよりも、本来の売り上げは多いかもしれない
- プロジェクトの品質分析におけるトラブル数予測
  - まだ終了していないプロジェクトや、途中で終了してしまったプロジェクトにおける本来のトラブル件数は、実際に観測された数より多いかもしれない
- 不活性の化合物が与えられたときの、化合物の活性値予測（QSAR）
  - いくつかの化合物は、専門家からみて、実際に活性値を計測しなくても活性がないとわかる

従来の回帰手法における問題点： 範囲出力を適切に扱えない

- 通常の回帰を解くアルゴリズムは存在する
  - 条件付分布 $p(y|\mathbf{x})$ を、下表のデータから求めることができる

入力 **x**				出力 y
x_1	x_2	x_3	x_4	
2	5	12	6	13
...	...	...	...	...
5	3	9	10	7

推定 ⇒ $p(y|\mathbf{x})$

- 通常の回帰アルゴリズムでは、範囲出力は扱えない
  - ナイーブな解法として「範囲出力をもつデータは無視する」

入力 **x**				出力 **y**	
x_1	x_2	x_3	x_4	l	r
2	5	12	6	11	14
...	...	...	...		
5	3	9	10	6	6

← 無視
← 使う

推定 ⇒ $p(y|\mathbf{x})$

問題の説明

手法の提案

正当化

実験

提案するアプローチ：
範囲出力の「代表値」を用いることによる繰り返し推定法

- 範囲出力に対する代表値を割り当てることによって、通常の回帰手法を適用可能にする
- 具体的には、「モデルの繰り返し推定手法」をもちいる
  - 現在のモデルを用いた代表値の割り当て
  - その代表値を用いた、（通常の回帰手法による）モデル推定

  を交互に繰り返す

代表値を与えることによって通常の回帰手法が適用可能

入力 **x**				出力 **y**		代表値
x_1	x_2	x_3	x_4	l	r	
2	5	12	6	11	14	13
...	...	...	...		...	...
5	3	9	10	6	∞	7

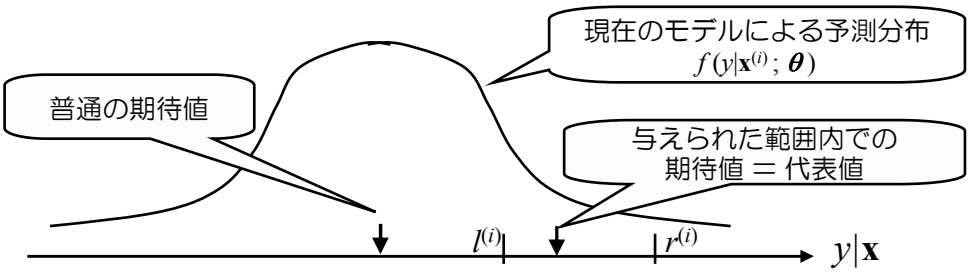
提案アプローチおける代表値の割り当て方：現在のモデルによる、条件付期待値を割り当てる

- 現在のモデル $f(y|\mathbf{x};\boldsymbol{\theta})$ をもちいて、i 番目のデータの出力範囲 $[l^{(i)},r^{(i)}]$ における条件付き期待値を求め、これを代表値として用いる

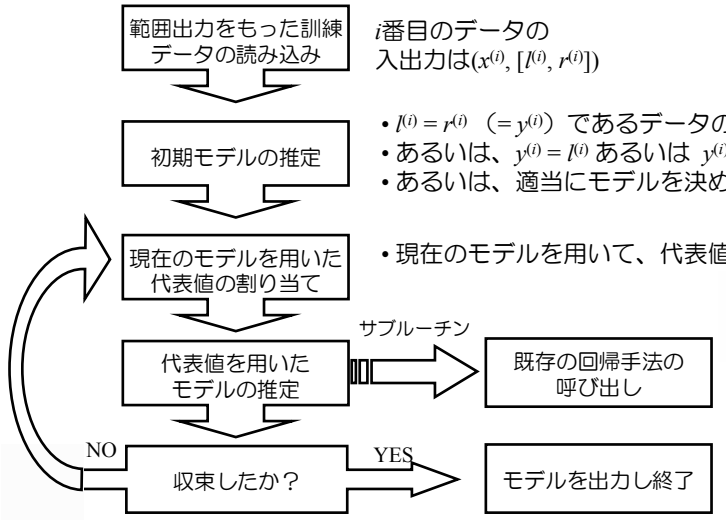
$$y^{(i)} = \frac{\int_{\ell^{(i)}}^{r^{(i)}} y f(y|\mathbf{x}^{(i)};\hat{\theta})dy}{\int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)};\hat{\theta})dy}$$

$[l^{(i)},r^{(i)}]$内での期待値

  - 含まれる積分は解析的に求めるのが困難な場合、サンプリングによって計算する



手続きをまとめると...



i番目のデータの
入出力は$(x^{(i)}, [l^{(i)}, r^{(i)}])$

- $l^{(i)} = r^{(i)}$ （$= y^{(i)}$）であるデータのみを使って推定
- あるいは、$y^{(i)} = l^{(i)}$ あるいは $y^{(i)} = r^{(i)}$ などとして推定
- あるいは、適当にモデルを決める

- 現在のモデルを用いて、代表値 $y^{(i)}$ を推定

$$y^{(i)} = \frac{\int_{\ell^{(i)}}^{r^{(i)}} y f(y|\mathbf{x}^{(i)};\hat{\theta})dy}{\int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)};\hat{\theta})dy}$$

問題の説明

手法の提案

正当化

実験

正当化： 提案手法は、指数分布族に対するEMアルゴリズムになっています

- 最大化する目的関数 = 範囲 $[l^{(i)}, r^{(i)}]$ の間に落ちる確率 とする

$$E = \sum_{i=1}^{N} \log \Pr(y^{(i)} \in [\ell^{(i)}, r^{(i)}] | \mathbf{x}^{(i)}) = \sum_{i=1}^{N} \log \int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)}) dy$$

- 陽に積分がとれないので、下界を最大化することにする
  - g は現在の事後分布の近似として、以下を最大化する

$$\sum_{i=1}^{N} \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \log f(y|\mathbf{x}^{(i)}) dy$$

- モデルが指数分布族であることを仮定する → $f(y|\mathbf{x}) = h(y)k(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \boldsymbol{\psi}(y))$

$$1 = k(\boldsymbol{\eta}) \int h(y) \exp(\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)) dy$$

  - なお、$\boldsymbol{\eta}$がパラメータ
- 結局、現在のモデルでの条件付き期待値で仮置きしたものを使って最尤推定すればよいことになる

$$\sum_{i=1}^{N} \log k(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^\top \underbrace{\int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \boldsymbol{\psi}(y) dy}_{\text{現在のモデルでの条件付き期待値}}\right)$$

具体的な導出例 （線形ガウスモデルの場合）

- 以下の線形ガウスモデルを考える

$$f(y|\mathrm{x},\boldsymbol{\theta},\sigma) = \mathcal{N}(y|\boldsymbol{\theta}\boldsymbol{\phi}^\top(\mathrm{x}),\sigma\mathrm{I}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\| y-\boldsymbol{\theta}\boldsymbol{\phi}^\top(\mathrm{x})\|^2}{2\sigma^2}\right)$$

  - $\boldsymbol{\phi}(\mathbf{x})$ は基底ベクトル、これを行列で書いたものを

$$\Phi = \begin{pmatrix} \phi_1(\mathbf{x}^{(1)}), & \phi_2(\mathbf{x}^{(1)}), & \ldots, & \phi_Z(\mathbf{x}^{(1)}) \\ \phi_1(\mathbf{x}^{(2)}), & \phi_2(\mathbf{x}^{(2)}), & \ldots, & \phi_Z(\mathbf{x}^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(\mathbf{x}^{(N)}), & \phi_2(\mathbf{x}^{(N)}), & \ldots, & \phi_Z(\mathbf{x}^{(N)}) \end{pmatrix}$$

- アルゴリズム
  - Eステップ
    - 出力値の仮置き値を求める
    - 積分はメトロポリス法で行う

$$y^{(i)} = \frac{\int_{\ell^{(i)}}^{r^{(i)}} y\exp\left(-\frac{1}{2\hat{\sigma}^2}(y-\hat{\boldsymbol{\theta}}\boldsymbol{\phi}^\top(\mathbf{x}^{(i)}))^2\right)dy}{\int_{\ell^{(i)}}^{r^{(i)}} \exp\left(-\frac{1}{2\hat{\sigma}^2}(y-\hat{\boldsymbol{\theta}}\boldsymbol{\phi}^\top(\mathbf{x}^{(i)}))^2\right)dy}$$

  - Mステップ
    - 仮置き値を使った最尤推定を解く

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \sum_{i=1}^{N} \log f(y^{(i)}|\mathbf{x}^{(i)},\boldsymbol{\theta})$$

    - 解は

$$\hat{\boldsymbol{\theta}} = (\lambda\mathrm{I}+\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{y}^\top$$

$$\hat{\sigma} = \sqrt{\frac{1}{N}(\hat{\boldsymbol{\theta}}\Phi^\top-\mathbf{y})(\hat{\boldsymbol{\theta}}\Phi^\top-\mathbf{y})^\top}$$

提案手法の他に考えられる割り当て方法と、その問題点

- 単純な平均を代表値として割り当てる
  - [5, ∞] の平均値は∞となってしまうのでダメ

- 単純なEMアルゴリズムの適用：「範囲出力」を用いずに、現在のモデルによる期待値を割り当てる

$$y^{(i)} = \int_{-\infty}^{\infty} y f(y|\mathbf{x}^{(i)};\hat{\boldsymbol{\theta}})dy$$

  - 範囲情報を適切に用いていないので、不十分

- 範囲制約を満たすような出力を出す解を求める （Knowledge-based Kernel Approximation）
  - 解が制約を満たすような計画問題を解く

$$y^{(i)} < f(y|\mathbf{x};\theta) < y^{(i)}$$

  - 確率的な解釈がない

問題の説明

手法の提案

正当化

実験

2つのデータセットを用いた実験で、提案手法の有効性が確認できました

- 2つのデータセットを用いました
  - 価格予測の問題（Boston housing）
  - 化合物の活性予測の問題（EDKB）
- 提案手法を、2つの既存手法と比べ、提案手法の優位性を確認した
  - 既存手法 1： 範囲出力をもったデータを無視する方法
  - 既存手法 2： 範囲出力を使わずに、現在のモデルによる期待値を割り当てる方法
- ベースの回帰手法としては、線形ガウスモデルを用いた

$$f(y|\mathrm{x},\boldsymbol{\theta},\sigma) = \mathcal{N}(y|\boldsymbol{\theta}\boldsymbol{\phi}^{\mathsf{T}}(\mathrm{x}),\sigma\mathrm{I})$$

実験結果1: 価格推定問題

- Boston housing data： 地域の犯罪発生率や、部屋数など、13 の指標（**x**）から、家の価格（y）を推定する問題
  - データ数は506
- 本来は、点出力の問題だが、半分のデータを点出力として、残り半分のデータを範囲出力を持つデータに変換して用いる
  - 変換の仕方は、yの10%までの一様乱数 Δ_l と Δ_r を用いて
    - 提案手法 L： $[y - \Delta_l, \infty]$
    - 提案手法LR： $[y - \Delta_l, y + \Delta_r]$
- 基底はガウスカーネルを用いる

$$\phi_i(\mathbf{x}) = \exp(-\alpha \parallel \mathbf{x} - \mathbf{x}^{(i)} \parallel^2)$$

手法	既存手法 1	既存手法 2	提案手法 L	提案手法 LR
平均 2 乗誤差	14.18	13.84	12.03	11.37

比較結果

提案手法が良い

実験2: 化合物の活性値予測問題

- Endocrine Disruptors Knowledge Databaseに含まれるE-SCREENアッセイデータの化合物の活性値（logRPP値）を予測する
  - データ数は59個

- 「専門家からみて明らかに活性のないデータ」を仮想的に作成して用いる
  - 活性値の小さいほうから1/3のデータのうち半分を活性のない化合物とする
  - これらの活性値を [-∞, -0.8421（小さいほうから1/3の活性値）] の範囲出力を持ったデータとする
- 部分グラフを用いた特徴ベクトル表現
  - 全体の5%以上に含まれる部分グラフ（最大サイズ20）で特徴ベクトルを作成
  - 13,600次元の2値ベクトル

手法	既存手法 1	既存手法 2	提案手法
平均 2 乗誤差	0.198	0.208	0.190

比較結果

提案手法が良い

まとめ： 実問題としての需要がありうるが、これまであまり扱われていなかった、範囲出力をもった回帰問題をEMアルゴリズムで解きました

- 範囲出力を用いるような回帰問題は、実問題において、現れうる
  - 従来法の単純な適用では、与えられた情報を十分に用いることができない
- 範囲出力を用いるような回帰問題を、陽に扱い、これに解法を与えた
  - 従来法の単純な適用では、適切に解くことができない

今後の研究

- Knowledge-based kernel approximationとの比較
- テスト時にも範囲が与えられている場合の実験
- ガウス過程の範囲出力化