NGBoost: Natural Gradient Boosting for Probabilistic Prediction

Kobe University
201T264T
Fuki Yamamoto

発表の流れ

1. 論文の背景, 概要
2. 準備
3. 手法の説明
4. 実験結果
5. まとめ

論文の概要

タイトル

NGBoost: Natural Gradient Boosting for Probabilistic Prediction

著者

Duan, Tony, et al.

出典

International Conference on Machine Learning. PMLR, 2020.

- 回帰問題において, 予測の不確かさを扱うことができる Gradient Boosting Decision Tree (GBDT)モデルを提案

- 勾配にnatural gradient(自然勾配)(Amari, 1998)を用いている

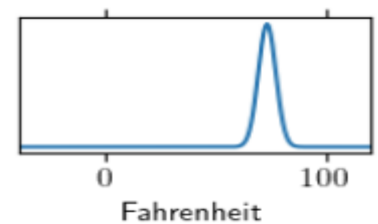
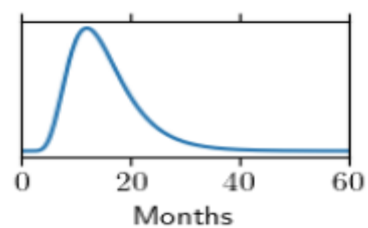
背景 (予測の不確かさとは)

回帰問題において, 予測値だけでなく,

確率分布そのものを出力できれば活用の幅が広がる

→ x ℃以上になる確率は?, x ヶ月以内に亡くなる確率は?

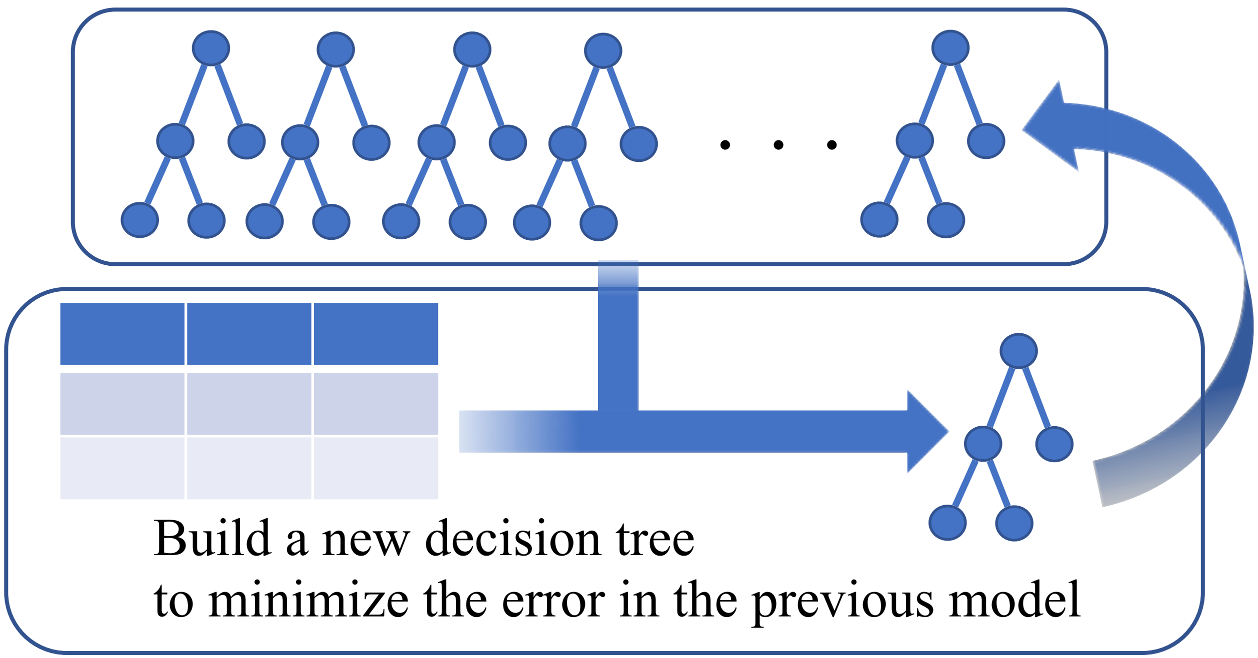
予測の信頼度はどの程度?　(分類問題では普通)

Question	Point Prediction (No uncertainty estimate)	Probabilistic Prediction (Uncertainty is implicit)
What will be the temperature at noon tomorrow?	73.4 Fahrenheit	
How long will this patient live?	11.3 months	

既存の手法は, 単純なモデルであったり, 計算コストが高い

GBDT とは

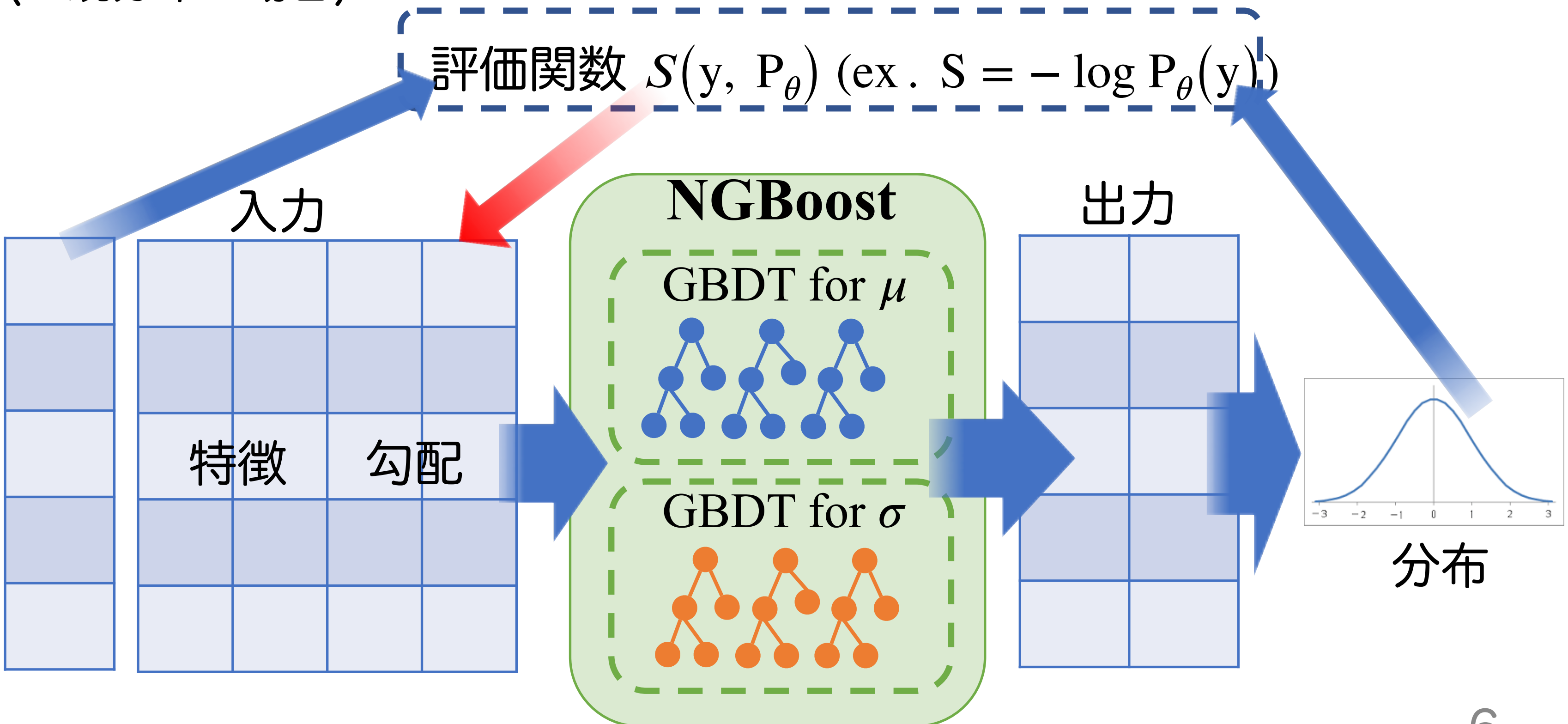
- 複数のweak learner(浅い決定木)を組み合わせたモデル
- テーブルデータに経験的に高精度
- XGBoost, Light GBM, Cat Boostなどが存在する



NGBoostとは

- 各サンプルに対し分布のパラメータ(正規分布ならσ, μ)を出力
- 学習をうまく進めるためnatural gradient(自然勾配)を導入
- 各パラメータ毎にGBDTを学習するmultiparameter boostingを導入

(正規分布の場合)



NGBoost: 学習アルゴリズム

Algorithm 1 NGBoost for probabilistic prediction

Data: Dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$.
Input: Boosting iterations M, Learning rate η, Probability distribution with parameter θ, Proper scoring rule $\mathcal{S}$, Base learner f.
Output: Scalings and base learners $\{\rho^{(m)}, f^{(m)}\}_{m=1}^M$.

$\theta^{(0)} \leftarrow \arg\min_\theta \sum_{i=1}^n \mathcal{S}(\theta, y_i)$ {initialize to marginal}
for $m \leftarrow 1, \dots, M$ **do**
  for $i \leftarrow 1, \dots, n$ **do**
    $g_i^{(m)} \leftarrow \mathcal{I}_\mathcal{S}\left(\theta_i^{(m-1)}\right)^{-1} \nabla_\theta \mathcal{S}\left(\theta_i^{(m-1)}, y_i\right)$
  end
  $f^{(m)} \leftarrow \text{fit}\left(\left\{x_i, g_i^{(m)}\right\}_{i=1}^n\right)$
  $\rho^{(m)} \leftarrow \arg\min_\rho \sum_{i=1}^n \mathcal{S}\left(\theta_i^{(m-1)} - \rho \cdot f^{(m)}(x_i), y_i\right)$
  for $i \leftarrow 1, \dots, n$ **do**
    $\theta_i^{(m)} \leftarrow \theta_i^{(m-1)} - \eta\left(\rho^{(m)} \cdot f^{(m)}(x_i)\right)$
  end
end

M:weak learnerの数
n: サンプル数
η: 学習率
$\boldsymbol{\theta}$: 分布のパラメータ
$\boldsymbol{f}^{(m)}$: m個目のlearner
(正規分布ならば,
$\boldsymbol{\theta} = \{\mu, \sigma\}$, $\boldsymbol{f}^{(m)} = \{f_\mu^{(m)}, f_\sigma^{(m)}\}$)

自然勾配を用いて勾配を更新

木を学習,
パラメータを更新

NGBoost: 学習アルゴリズム

Algorithm 1 NGBoost for probabilistic prediction

Data: Dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$.

Input: Boosting iterations M, Learning rate η, Probability distribution with parameter θ, Proper scoring rule $\mathcal{S}$, Base learner f.

Output: Scalings and base learners $\{\rho^{(m)}, f^{(m)}\}_{m=1}^M$.

$\theta^{(0)} \leftarrow \arg\min_\theta \sum_{i=1}^n \mathcal{S}(\theta, y_i)$ {initialize to marginal}

for $m \leftarrow 1, \ldots, M$ **do**

  for $i \leftarrow 1, \ldots, n$ **do**

    $g_i^{(m)} \leftarrow \mathcal{I}_\mathcal{S}\left(\theta_i^{(m-1)}\right)^{-1} \nabla_\theta \mathcal{S}\left(\theta_i^{(m-1)}, y_i\right)$

  end

  $f^{(m)} \leftarrow \text{fit}\left(\left\{x_i, g_i^{(m)}\right\}_{i=1}^n\right)$

  $\rho^{(m)} \leftarrow \arg\min_\rho \sum_{i=1}^n \mathcal{S}\left(\theta_i^{(m-1)} - \rho \cdot f^{(m)}(x_i), y_i\right)$

  for $i \leftarrow 1, \ldots, n$ **do**

    $\theta_i^{(m)} \leftarrow \theta_i^{(m-1)} - \eta\left(\rho^{(m)} \cdot f^{(m)}(x_i)\right)$

  end

end

M:*weak learner*の数

n: サンプル数

η: 学習率

$\boldsymbol{\theta}$: 分布のパラメータ

$\boldsymbol{f}^{(\boldsymbol{m})}$: m個目の*learner*

(正規分布ならば,

$\boldsymbol{\theta} = \{\mu, \sigma\}$, $\boldsymbol{f}^{(\boldsymbol{m})} = \{f_\mu^{(m)}, f_\sigma^{(m)}\}$)

自然勾配を用いて勾配を更新

木を学習,

パラメータを更新

NGBoost: 自然勾配法

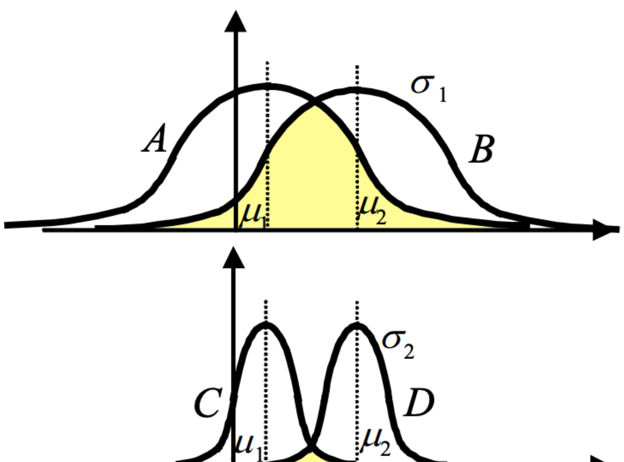
通常の勾配降下法では

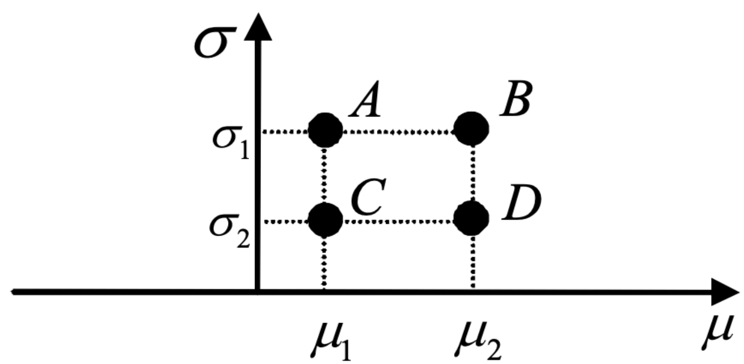
パラメータの変動のユークリッド距離をϵ (ϵは微小な定数)で固定し,

その中で最もモデルを改善する方向を示す

しかし,

- パラメータ間のユークリッド距離は分布間の差異に対応しない
- パラメータの変動をユークリッド距離で固定しても,
  パラメータ毎の感度が違えば分布の変動は固定されない





点A, B間と点C, D間のユークリッド距離は同じだが,

それらの点が示す分布(正規分布)同士の差異は大きく異なる

NGBoost: 自然勾配法

通常の勾配 ($\boldsymbol{g} = \boldsymbol{\nabla}_{\theta}\boldsymbol{S}$)

- 分布の各パラメータが分布の変動に与える影響は均一でない
- 通常の勾配ではそれらを加味できない

自然勾配 **(Amari, 1998)**

$$\boxed{g_i^{(m)} \leftarrow \mathcal{I}_S\left(\theta_i^{(m-1)}\right)^{-1} \nabla_\theta \mathcal{S}\left(\theta_i^{(m-1)}, y_i\right)}$$

- 分布間の差異とパラメータ間の距離を対応させることができる
- パラメータの変動でなくモデルの変動を固定して方向を決定する
- $I_S(\theta)$によってパラメータ毎の歪みを考慮

$I_S(\theta)$, Sは分布間の差異として何を用いるかによって変化する

1. **KL-divergence**

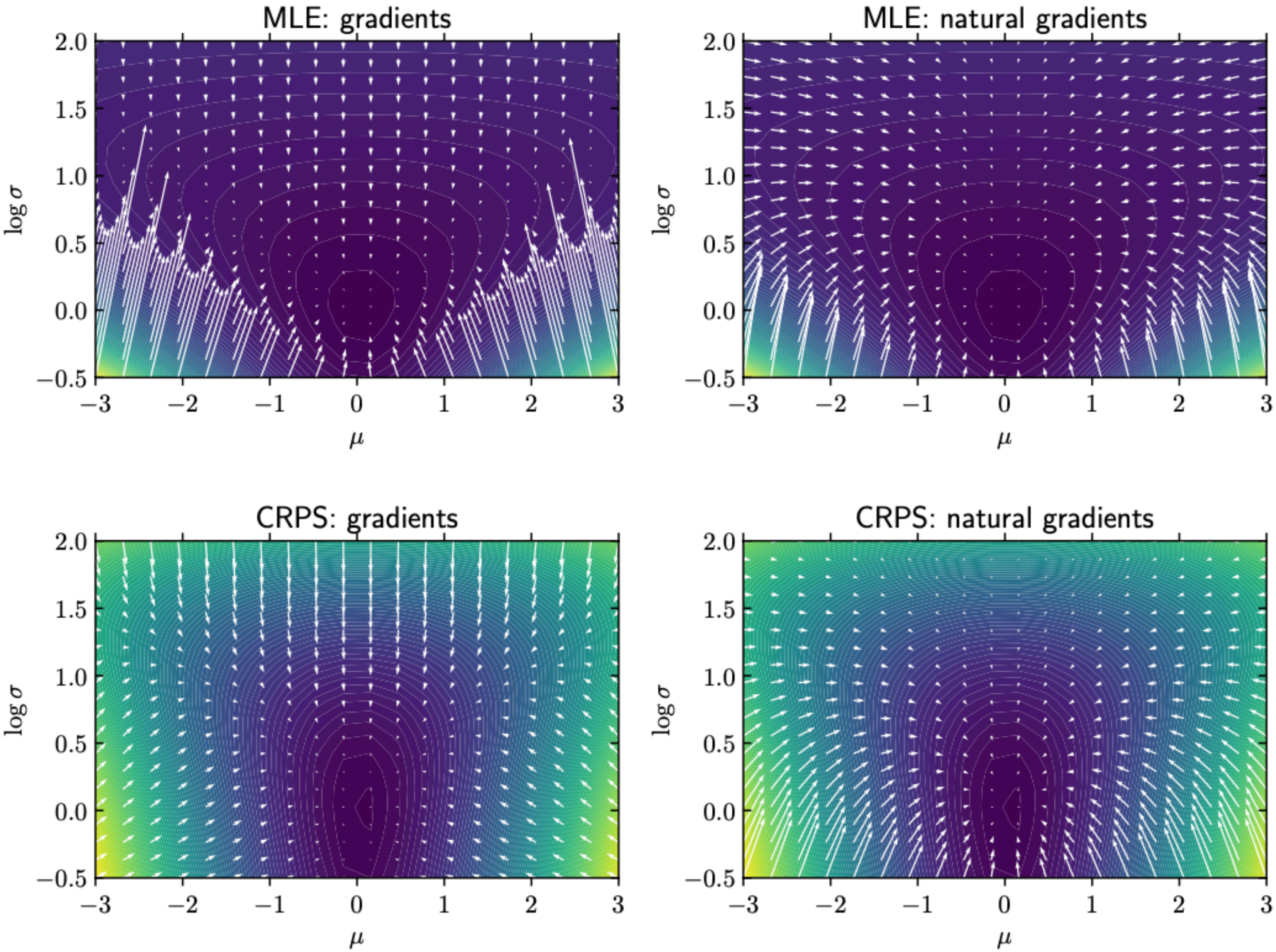
   $S = -\log\left(\mathrm{P}_\theta\left(y\right)\right)$, $I_S(\theta) = F(\theta)$ (フィッシャー情報量行列)

2. **L2-divergense**

   $S = \mathrm{CRPS}$, $I_S(\theta) = \dots$ (式は論文に記載されています)

NGBoost: 自然勾配法

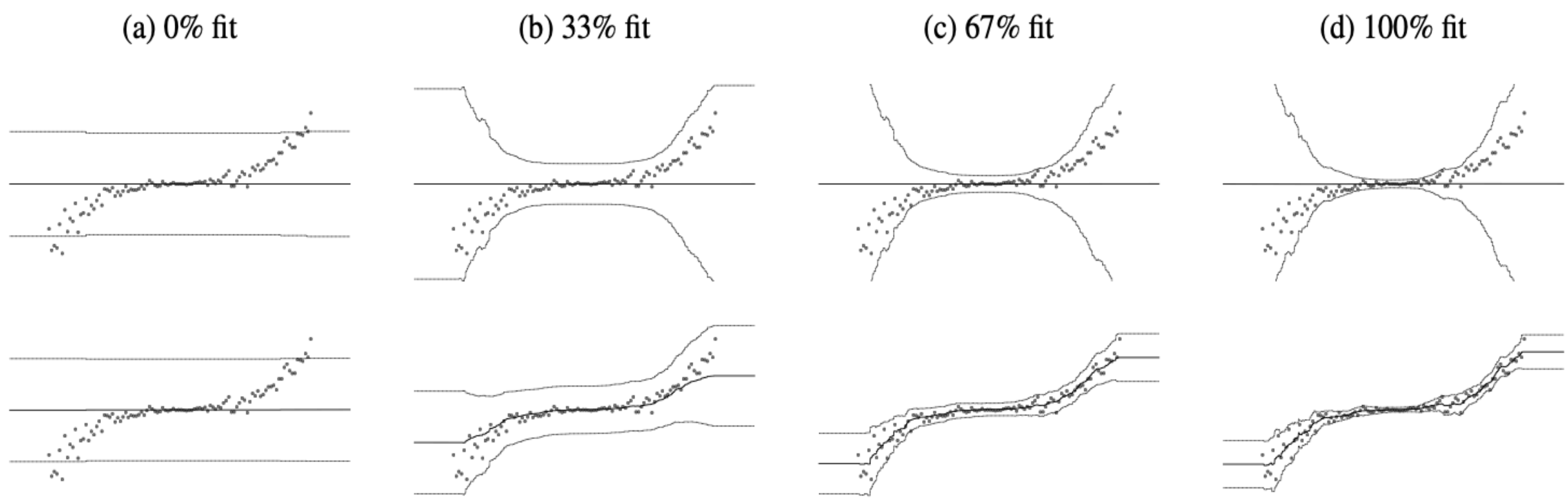
$N(0, 1)$からの標本で勾配を計算 *(左: 通常の勾配　右: 自然勾配)*



NGBoost: 自然勾配法

下図のような1次元データで学習を行なった図

(上: 通常の勾配, 下: 自然勾配)



通常勾配(上側)

平均はほとんど更新されず, 中央の分散ばかり学習

自然勾配(下側)

全体的にバランスが取れている

NGBoost: 学習アルゴリズム

Algorithm 1 NGBoost for probabilistic prediction

Data: Dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$.
Input: Boosting iterations M, Learning rate η, Probability distribution with parameter θ, Proper scoring rule $\mathcal{S}$, Base learner f.
Output: Scalings and base learners $\{\rho^{(m)}, f^{(m)}\}_{m=1}^M$.

$\theta^{(0)} \leftarrow \arg\min_\theta \sum_{i=1}^n \mathcal{S}(\theta, y_i)$ {initialize to marginal}
for $m \leftarrow 1, \ldots, M$ **do**
  for $i \leftarrow 1, \ldots, n$ **do**
    $g_i^{(m)} \leftarrow \mathcal{I}_{\mathcal{S}}\left(\theta_i^{(m-1)}\right)^{-1} \nabla_\theta \mathcal{S}\left(\theta_i^{(m-1)}, y_i\right)$
  end
  $f^{(m)} \leftarrow \text{fit}\left(\left\{x_i, g_i^{(m)}\right\}_{i=1}^n\right)$
  $\rho^{(m)} \leftarrow \arg\min_\rho \sum_{i=1}^n \mathcal{S}\left(\theta_i^{(m-1)} - \rho \cdot f^{(m)}(x_i), y_i\right)$
  for $i \leftarrow 1, \ldots, n$ **do**
    $\theta_i^{(m)} \leftarrow \theta_i^{(m-1)} - \eta\left(\rho^{(m)} \cdot f^{(m)}(x_i)\right)$
  end
end

M:weak learnerの数
n: サンプル数
η: 学習率
θ: 分布のパラメータ
$f^{(m)}$: m個目のlearner
(正規分布ならば,
$\theta = \{\mu, \sigma\}$, $f^{(m)} = \{f_\mu^{(m)}, f_\sigma^{(m)}\}$)

自然勾配を用いて勾配を更新

木を学習,
パラメータを更新

NGBoost: パラメータの更新

$f^{(m)}$の学習

gを応答変数として決定木を学習 (不純度を最小化するように)

出力は葉に存在するデータの応答変数の平均とする

Scaling $\rho^{(m)}$を探索

m個目のfの出力に対するスケールパラメータ

大きな 探索

$$f^{(m)} \leftarrow \mathrm{fit}\left(\left\{x_i, g_i^{(m)}\right\}_{i=1}^{n}\right)$$

$$\rho^{(m)} \leftarrow \arg\min_\rho \sum_{i=1}^{n} \mathcal{S}\left(\theta_i^{(m-1)} - \rho \cdot f^{(m)}(x_i), y_i\right)$$

for $i \leftarrow 1, \ldots, n$ **do**

$$\theta_i^{(m)} \leftarrow \theta_i^{(m-1)} - \eta\left(\rho^{(m)} \cdot f^{(m)}(x_i)\right)$$

end

実験結果

以下の三つの実験でNGBoostの有用性を検証した

1. 他の確率回帰モデルとの精度比較 **(NLL)**

2. **NGBoost**において,
   自然勾配以外の勾配を用いた場合と比較 **(NLL)**

3. 他のモデル(確率回帰モデル以外)との精度比較 **(RMSE)**

全ての手法において, NGBoostの以下のパラメータは固定

木の深さ:3, η:{0.1(MSD), 0.01(others)}, S: $-\log(P_\theta(y))$, 分布:正規分布

更新回数Mのみhold-outでチューニング

実験結果

1. 他の確率回帰モデルとの精度比較 (NLL)

Dataset	N	NGBoost	MC dropout	Deep Ensembles	Concrete Dropout	Gaussian Process	GAMLSS	DistForest
Boston	506	**2.43 ± 0.15**	**2.46 ± 0.25**	**2.41 ± 0.25**	2.72 ± 0.01	**2.37 ± 0.24**	2.73 ± 0.56	2.67 ± 0.08
Concrete	1030	**3.04 ± 0.17**	**3.04 ± 0.09**	**3.06 ± 0.18**	3.51 ± 0.00	**3.03 ± 0.11**	3.24 ± 0.08	3.38 ± 0.05
Energy	768	**0.60 ± 0.45**	1.99 ± 0.09	1.38 ± 0.22	2.30 ± 0.00	**0.66 ± 0.17**	1.24 ± 0.86	1.53 ± 0.14
Kin8nm	8192	-0.49 ± 0.02	-0.95 ± 0.03	**-1.20 ± 0.02**	-0.65 ± 0.00	-1.11 ± 0.03	-0.26 ± 0.02	-0.40 ± 0.01
Naval	11934	-5.34± 0.04	-3.80 ± 0.05	-5.63 ± 0.05	**-5.87 ± 0.05**	-4.98 ± 0.02	-5.56 ± 0.07	-4.84 ± 0.01
Power	9568	2.79 ± 0.11	2.80 ± 0.05	2.79 ± 0.04	2.75 ± 0.01	2.81 ± 0.05	2.86 ± 0.04	**2.68 ± 0.05**
Protein	45730	2.81 ± 0.03	2.89 ± 0.01	2.83 ± 0.02	2.81 ± 0.00	2.89 ± 0.02	3.00 ± 0.01	**2.59 ± 0.04**
Wine	1588	**0.91 ± 0.06**	**0.93 ± 0.06**	**0.94 ± 0.12**	1.70 ± 0.00	**0.95 ± 0.06**	**0.97 ± 0.09**	1.05 ± 0.15
Yacht	308	**0.20 ± 0.26**	1.55 ± 0.12	1.18 ± 0.21	1.75 ± 0.00	**0.10 ± 0.26**	0.80 ± 0.56	2.94 ± 0.09
Year MSD	515345	3.43 ± NA	3.59 ± NA	**3.35 ± NA**	NA± NA	NA ± NA	NA ± NA	NA ± NA

データセット10個中5つが太字(最良+標準偏差で重なる)

→ 他の手法と比較しても, 上位の性能を誇ってる

実験結果

2. NGBoostにおいて,

自然勾配以外の勾配を用いた場合と比較 **(NLL)**

Dataset	N	NGBoost	2nd-Order	Multiparameter	Homoscedastic
Boston	506	**2.43** ± **0.15**	3.57 ± 0.20	3.17 ± 0.13	**2.79** ± **0.42**
Concrete	1030	**3.04** ± **0.17**	4.21 ± 0.05	3.94 ± 0.09	**3.22** ± **0.29**
Energy	768	**0.60** ± **0.45**	3.64 ± 0.06	3.24 ± 0.09	**0.68** ± **0.25**
Kin8nm	8192	-0.49 ± 0.02	0.10 ± 0.07	**-0.52** ± **0.03**	-0.37 ± 0.05
Naval	11934	**-5.34**± **0.04**	-2.80 ± 0.01	-3.46 ± 0.00	-4.35 ± 0.07
Power	9568	2.79 ± 0.11	4.11 ± 0.03	3.79 ± 0.13	**2.66** ± **0.11**
Protein	45730	**2.81** ± **0.03**	3.23 ± 0.00	3.04 ± 0.02	2.86 ± 0.01
Wine	1588	**0.91** ± **0.06**	1.21 ± 0.09	0.93 ± 0.07	**1.34** ± **0.67**
Yacht	308	**0.20** ± **0.26**	4.11 ± 0.17	3.29 ± 0.20	2.02 ± 0.21
Year MSD	515345	**3.43** ± **NA**	3.80 ± 0.00	3.60 ± NA	3.63 ± NA

データセット10個中8つが太字

→ 自然勾配を用いた場合が最も高精度

実験結果

3. 他のモデル(確率回帰モデル以外)との精度比較 (RMSE)

Dataset	N	NGBoost	Elastic Net	Random Forest	Gradient Boosting	GAMLSS	Distributional Forest
Boston	506	**2.94 ± 0.53**	4.08 ± 0.16	2.97 ± 0.30	**2.46 ± 0.32**	4.32 ± 1.40	3.99 ± 1.13
Concrete	1030	**5.06 ± 0.61**	12.1 ± 0.05	5.29 ± 0.16	**4.46 ± 0.29**	6.72 ± 0.59	6.61 ± 0.83
Energy	768	0.46 ± 0.06	2.75 ± 0.03	0.52 ± 0.09	**0.39 ± 0.02**	1.43 ± 0.32	1.11 ± 0.27
Kin8nm	8192	0.16 ± 0.00	0.20 ± 0.00	0.15 ± 0.00	**0.14 ± 0.00**	0.20 ± 0.01	0.16 ± 0.00
Naval	11934	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
Power	9568	3.79 ± 0.18	4.42 ± 0.00	3.26 ± 0.03	**3.01 ± 0.10**	4.25 ± 0.19	3.64 ± 0.24
Protein	45730	4.33 ± 0.03	5.20 ± 0.00	**3.60 ± 0.00**	3.95 ± 0.00	5.04 ± 0.04	3.89 ± 0.04
Wine	1588	0.63 ± 0.04	0.58 ± 0.00	**0.50 ± 0.01**	0.53 ± 0.02	0.64 ± 0.04	0.67 ± 0.05
Yacht	308	**0.50 ± 0.20**	7.65 ± 0.21	0.61 ± 0.08	**0.42 ± 0.09**	8.29 ± 2.56	4.19± 0.92
Year MSD	515345	8.94 ± NA	9.49 ± NA	9.05 ± NA	**8.73 ± NA**	NA ± NA	NA ± NA

データセット10個中4つが太字

→ 他手法と比較しても上位の性能

関連論文

Uncertainty in Gradient Boosting via Ensembles
Ustimenko, Aleksei, Liudmila Prokhorenkova, and Andrey Malinin
arXiv preprint arXiv:2006.10562 (2020).

- Yandexが発表
- NGBoostより後発
- NGBoostと同じく自然勾配を利用
- CatBoost(Yandexが開発)に実装されている

詳しい内容, 差異は読んでないためわかりません.

まとめ

内容

回帰問題で,

予測の不確かさ(分布)を出力可能なGBDTモデルNGBoostを提案

貢献

・勾配計算に自然勾配を導入

・パラメータ毎に異なるモデルを用いるmultiparameter boostingを導入

メリット

・チューニングが容易 (NNベースのモデルと比較して)

・様々な分布を適用可能 (パラメータで表現できるもの)

・他の手法に劣らない精度 (実験 1, 3)