

データ分布の関数パラメータとしてのカーネル中央値

Kernel median embedding as a functional parameter of the data distribution

マシュー ホーランド^{*1}

Matthew J. Holland

^{*1}大阪大学 データビリティフロンティア機構

Osaka University, Institute for Datability Science

In both supervised and unsupervised learning tasks, embedding the underlying data into a function space using a “kernel mean” has been well-studied, and is known to be an efficient means of characterizing even complex distributions. Here we consider a broad generalization of this notion to countless “functional parameters” of the underlying distribution, and as a concrete example explore what may naturally be called the “kernel median” of the data. In this short paper, we formulate the new parameter class, provide a procedure for obtaining an important special case, with basic convergence guarantees and expressions useful for practical implementation.

1. はじめに

機械学習の研究者の多くがニューラルネットワークに着目するなか、強力な理論的保証と長い歴史を持つカーネル法も忘れてはならない。本稿では、kernel mean をデータ分布から再生核ヒルベルト空間への写像と見なして、データ分布の特徴を簡潔に求める新しい手法を提案する。

古典的な統計学では、所与とした標本 $z_1, \dots, z_n \in \mathcal{Z}$ の確率分布が $z \sim \mu$ であるとして、その分布 μ をある程度（場合によっては完全に）特徴づけるような有限個のパラメータを推定することが代表的な問題である。しかし、現代の機械学習の応用問題を見据えて、有限の長さを持つ実数値パラメータベクトルだけで、コンパクトに複雑な分布を符号化することは難しい。そこで、計算コストと表現力なるパフォーマンスの釣り合った方法として、入力空間上のカーネル関数 $k: \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ を介して、その分布 μ を再生核ヒルベルト空間別の関数空間 $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ へと写像していく “kernel mean” にまつわる手法が注目されている [Muandet et al., 2016]。その定義をまず述べておくと、 k と $\mathcal{H}$ を固定し、

$$\mathbf{E}_{\mu} k(z, \cdot) := \int_{\mathcal{Z}} k(z, \cdot) d\mu(z) \in \mathcal{H} \quad (1)$$

を確率分布 μ のカーネル平均 (kernel mean) と呼ぶ。カーネル平均さえ知っていれば十分とい問題・手法が多々ある [Muandet et al., 2014]。そのため、この理論上の関数 $\mathbf{E}_{\mu} k(z, \cdot)$ を効率的に近似できるのであれば、密度推定を介さずとも多様な問題が解けるようになるという大きな潜在的なメリットを持つ。従来の近似方法

として、 μ を経験分布に置き換えて、

$$\bar{\pi} := \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n d(\Phi_{z_i}, f) \quad (2)$$

を使うことが多い。文献の上では、この推定量は調べ尽くされている [Tsybakov, 2009]。また、カーネル関数を $\Phi_z(\cdot) := k(z, \cdot)$ 、ノルムを $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ で表記すると、

$$\mathbf{E}_{\mu} k(z, \cdot) \in \arg \min_{f \in \mathcal{H}} \mathbf{E}_{\mu} \|\Phi_z - f\|_{\mathcal{H}}^2 \quad (3)$$

が成り立つことは容易に確認できる。学習理論から見ると、これは空間 $\mathcal{H}$ 上の自乗誤差を使つての経験期待損失最小化 (empirical risk minimization; ERM) の解にほかならず、その脆弱性がここ数年では大いに注目されている [Minsker, 2016, Lugosi and Mendelson, 2016]。

これらの事実を踏まえて、次の疑問点を挙げる。

1. 学習効率の観点から、 $\bar{\pi} \approx \mathbf{E}_{\mu} k(z, \cdot)$ で良いか。
2. 自乗誤差以外の距離を使うメリットがあるか。

次節以降では後者に焦点を当て、カーネル平均の概念を拡張、特に「カーネル中央値」とでもいうべき新たな符号化方法の効率的なアルゴリズムを提案する。

2. 提案手法

本研究の基となる発想は至って単純である。実数線上の確率変数 x を考えると、その期待値 $\mathbf{E} x$ が、 $\mathbf{E}(x-u)^2$ を u の関数として最小にするのである。一方、絶対誤差にすると、 $\mathbf{E}|x-u|$ を u について最小にするのは、 x の中央値である。先ほどの式 (3) を念頭において類推すると、

$$\text{med}_k(\mu) := \arg \min_{f \in \mathcal{H}} \mathbf{E}_{\mu} \|\Phi_z - f\|_{\mathcal{H}}, \quad (4)$$

連絡先: Matthew J. Holland

大阪大学データビリティフロンティア機構

大阪府吹田市山田丘 2-8

matthew-h@ids.osaka-u.ac.jp

をカーネル中央値 (kernel median) と呼ぶことが自然であろう。同様に考えることで分布 μ の多様なパラメータのクラスを新たに提案することができて、たとえば、適当な距離関数 $d: \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}_+$ を取って、

$$\pi(\mu; d) := \arg \min_{f \in \mathcal{H}} \mathbf{E}_\mu d(\Phi_{\mathbf{z}}, f) \quad (5)$$

のような関数パラメータ族が考えられる。自乗誤差と同様、 $d(\Phi_{\mathbf{z}}, f)$ の分布の如何によっては、その推定しやすさと、 μ の代替的表現としての性質が大きく変わり得る。紙幅の制約もあるため、経験分布の下、カーネル中央値を推定するアルゴリズムについて考察していく。

観測データ $S_n := \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \subset \mathcal{Z}$ を与えられ、 $\text{med}_k(\mu)$ の不偏推定量として、

$$\text{med}_k S_n := \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \|\Phi_{\mathbf{z}_i} - f\|_{\mathcal{H}} \quad (6)$$

を考える。この経験カーネル中央値 (empirical kernel median) を求める計算方法を Algorithm 1 として一覧化する。

導出方法について簡単に解説していく。

$$C(h) := \sum_{i=1}^n \|h - h_i\|$$

という目的関数を考えると、基本的には単なる $C(\cdot)$ の最急降下法に近い。もし任意の $i = 1, \dots, n$ に対して $f \neq \Phi_{\mathbf{z}_i}$ なのであれば、

$$\sum_{i=1}^n \frac{\Phi_{\mathbf{z}_i} - f}{\|\Phi_{\mathbf{z}_i} - f\|_{\mathcal{H}}} = 0$$

を満たすことが最小値の一階条件である。整理すると

$$f \leftarrow \left(\sum_{i=1}^n \|\Phi_{\mathbf{z}_i} - f\|_{\mathcal{H}}^{-1} \right)^{-1} \sum_{i=1}^n \|\Phi_{\mathbf{z}_i} - f\|_{\mathcal{H}}^{-1} \Phi_{\mathbf{z}_i} \quad (7)$$

という反復更新則がそのまま出てくる。これは Algorithm 1 にある WEISZFELD のスペシャルケースである^{*1}。ここで問題となるのは、更新則 (7) は $f \notin \{\Phi_{\mathbf{z}_i}\}_{i=1}^n$ を前提としていることである。しかし、更新後の結果が $\{\Phi_{\mathbf{z}_i}\}_{i=1}^n$ の要素と一致しない保証はない。一致しないなら $A_{(t)}$ のみ非零となって WEISZFELD を実行することになるが、一致してしまった場合は $B_{(t)}$ も非零となって、最適解でない限り $\{\Phi_{\mathbf{z}_i}\}_{i=1}^n$ から候補 $\mathbf{u}_{(t)}$ を動かすのである。 $B_{(t)}$ の具体的な形式は、 $C(\cdot)$ を局所的に近似する関数 $\hat{C}(\cdot)$ に対して、確実な改善 $\hat{C}(\mathbf{u}_{(t+1)}) < \hat{C}(\mathbf{u}_{(t)})$ を保証する条件として導出できる (詳細は省略)。

Algorithm 1 経験カーネル中央値の算出

```
function WEISZFELD( $\mathbf{u}; S = \{\mathbf{v}_1, \mathbf{v}_2, \dots\}$ )
  return  $\left( \sum_{i \in [|S|]} \|\mathbf{v}_i - \mathbf{u}\|^{-1} \right)^{-1} \sum_{i \in [|S|]} \|\mathbf{v}_i - \mathbf{u}\|^{-1} \mathbf{v}_i$ 
end function
```

inputs: $\epsilon > 0$, $\Phi = \{\Phi_{\mathbf{z}_1}, \dots, \Phi_{\mathbf{z}_n}\}$, $\tilde{T}(\mathbf{u}) := \text{WEISZFELD}(\mathbf{u}; \Phi)$

initialize: $\mathbf{u}_{(0)} = \frac{1}{n} \sum_{i=1}^n \Phi_{\mathbf{z}_i}$, $\epsilon_{(0)} > \epsilon$, $t = 0$

while $\epsilon_{(t)} > \epsilon$ **do**

$\mathcal{I}_{(t)} \leftarrow \{i : \Phi_{\mathbf{z}_i} \neq \mathbf{u}_{(t)}\}$

$\mathbf{g}_{(t)} \leftarrow \sum_{i \in \mathcal{I}_{(t)}} \frac{\Phi_{\mathbf{z}_i} - \mathbf{u}_{(t)}}{\|\Phi_{\mathbf{z}_i} - \mathbf{u}_{(t)}\|}$

$A_{(t)} \leftarrow \left(1 - \frac{|\mathcal{I}_{(t)}^c|}{\|\mathbf{g}_{(t)}\|} \right) \tilde{T}(\mathbf{u}_{(t)})$

$B_{(t)} \leftarrow \min \left(1, \frac{|\mathcal{I}_{(t)}^c|}{\|\mathbf{g}_{(t)}\|} \right) \mathbf{u}_{(t)}$

$\mathbf{u}_{(t+1)} \leftarrow A_{(t)} + B_{(t)}$

$\epsilon_{(t+1)} \leftarrow \|\mathbf{u}_{(t+1)} - \mathbf{u}_{(t)}\|$

$t \leftarrow t + 1$

end while

return $\mathbf{u}_{(t)}$

3. 学習則の基本的な性質

まず、証明できる重要な理論上の性質を端的に紹介する。軽微な正則条件の下、解は必ず存在する。その上で $\{\Phi_{\mathbf{z}_i}\}_{i=1}^n$ がすべて同一線上である場合とそうでない場合に分ける。前者は、それらの重みづけの中央値を取ることで解がすぐに求まる。後者は反復していくしかないが、Algorithm 1 の更新則で $\mathbf{u}_{(t)}$ が最適解でない限り、

$$C(\mathbf{u}_{(t+1)}) < C(\mathbf{u}_{(t)})$$

と単調的な改善は保証できる。そのため、有限回の更新で最適解に到達することも証明できる。

上記の性質のほか、計算機でこのアルゴリズムを実装するにあたって、 $\mathbf{u}_{(t)}$ が関数であることを踏まえて、効率的にどのように表現すべきかは一見して、自明ではない。幸い、初期値 $\mathbf{u}_{(0)}$ が $\{\Phi_{\mathbf{z}_i}\}_{i=1}^n$ の凸包に含まれているなら、更新則の性質上、 $\mathbf{u}_{(t)}$ も同様に凸包に含まれる。つまり、下記のとおりの特徴量関数の凸結合と

^{*1} ユークリッド空間の場合、この更新則は古く、遅くても [Weiszfeld, 1937] が提案している。

して表せるのである。

$$\mathbf{u}_{(t)} = \sum_{i=1}^n w_i^{(t)} \Phi_{\mathbf{z}_i}$$

$$0 \leq w_i^{(t)} \leq 1, i \in [n], \quad \sum_{i=1}^n w_i^{(t)} = 1$$

よって, n 個の係数とデータ $\{\mathbf{z}_i\}_{i=1}^n$ さえ格納しておけば $\mathbf{u}_{(t)}$ は求められる。

実装上, 計算する係数がいくつもあるので, ここで整理して表わすことにする。まず, 更新 $\tilde{T}(\cdot)$ は定義上, 特徴量関数の凸結合であり, $\tilde{T}(\mathbf{u}_{(t)}) = \sum_{i=1}^n \tilde{w}_i^{(t)} \Phi_{\mathbf{z}_i}$ と表せる。これらの重み $\{\tilde{w}_i^{(t)}\}_{i=1}^n$ は

$$\tilde{w}_i^{(t)} = \left(\sum_{j=1}^n \|\Phi_{\mathbf{z}_j} - \mathbf{u}_{(t)}\|^{-1} \right)^{-1} \|\Phi_{\mathbf{z}_i} - \mathbf{u}_{(t)}\|^{-1}$$

という形を取るが, 再生核ヒルベルト空間の基本的な性質を利用して, 距離を次のように計算できる。内積は

$$\begin{aligned} \|\Phi_{\mathbf{z}_i} - \mathbf{u}_{(t)}\|^2 &= \|\Phi_{\mathbf{z}_i}\|^2 - 2\langle \Phi_{\mathbf{z}_i}, \mathbf{u}_{(t)} \rangle + \|\mathbf{u}_{(t)}\|^2 \\ &:= A_i - 2B_i^{(t)} + C^{(t)} \end{aligned}$$

と分解できる ($A_i := \|\Phi_{\mathbf{z}_i}\|^2$, $B_i^{(t)} := \langle \Phi_{\mathbf{z}_i}, \mathbf{u}_{(t)} \rangle$, $C^{(t)} := \|\mathbf{u}_{(t)}\|^2$)。これらの要素は

$$\begin{aligned} A_i &= k(\mathbf{z}_i, \mathbf{z}_i) \\ B_i^{(t)} &= \sum_{j=1}^n w_j^{(t)} k(\mathbf{z}_i, \mathbf{z}_j) \\ C^{(t)} &= \sum_{i=1}^n \left(w_i^{(t)} \right)^2 k(\mathbf{z}_i, \mathbf{z}_i) + 2 \sum_{i < j} w_i^{(t)} w_j^{(t)} k(\mathbf{z}_i, \mathbf{z}_j) \end{aligned}$$

と容易に求めることができる。 A_i は事前に計算しておくことができるが, $B_i^{(t)}$ と $C^{(t)}$ は更新するごとに値が変わる。 $|\mathcal{I}_{(t)}^c|$ は各 $i \in [n]$ に対して $\|\Phi_{\mathbf{z}_i} - \mathbf{u}_{(t)}\|$ がわかれば十分なので, A_i , $B_i^{(t)}$, $C^{(t)}$ を入手すれば計算できる。最後は $\|\mathbf{g}_{(t)}\|$ だが, その自乗を取ると, 次のように求まる:

$$\begin{aligned} \|\mathbf{g}_{(t)}\|^2 &= \sum_{i=1}^n \left(a_i^{(t)} \right)^2 \left(A_i - 2B_i^{(t)} + C^{(t)} \right) \\ &\quad + 2 \sum_{i < j} a_i^{(t)} a_j^{(t)} \left(k(\mathbf{z}_i, \mathbf{z}_j) - B_i^{(t)} - B_j^{(t)} + C^{(t)} \right). \end{aligned}$$

可読性を高めるため, $a_i^{(t)} := \|\Phi_{\mathbf{z}_i} - \mathbf{u}_{(t)}\|^{-1}$ と表記している。

4. 関連研究との比較

本稿の第一節で先述の通り, 分布を関数空間へと写像していくカーネル平均を生かした学習アルゴリズム

が多数あり, 実験・理論ともに解析の対象となっている [Muandet et al., 2016, Tolstikhin et al., 2016]。本研究の最大の特徴として, このカーネル平均の概念を一般的な関数パラメータへと拡張した上で, 解釈しやすい計算可能な具体例としてカーネル中央値を提案している点が挙げられる。より技術的な貢献として, 本稿では詳細を省いているが, WEISZFELD の解析を手がけた [Kuhn, 1973] の結果をユークリッド空間から任意のヒルベルト空間 (=完備内積空間) へ拡張している点, また [Vardi and Zhang, 2000] の手法を, $\{\Phi_{\mathbf{z}_i}\}_{i=1}^n$ 内の複数点が重複しても $C(\cdot)$ の単調減少が保証できるように改造している点が新しい。距離の和を最小にするアルゴリズムは一般に geometric median と呼ばれることが多く, 学習アルゴリズムに導入したときの統計的なロバスト性は [Minsker et al., 2014, Minsker, 2015] などで調べられている。

5. おわりに

本稿では, データの分布をコンパクトに符号化する新しい手法として分布の関数パラメータに着目し, そのスペシャルケースとしてカーネル中央値を求める算法を取り上げた。このアルゴリズムの単調性と収束条件および計算機上の表現など, 基本的な性質を明らかにした。この土台を踏まえて, 次はカーネル平均が導入している統計的学習アルゴリズムに代替的に適用することによって, 学習効率をめぐる定量的な性能保証がどう変わるか, それぞれの挙動が定性的にどう異なるかという点に着目しながら, 学習課題に応じての優劣を明らかにして新たな方法論を目指していく。

参考文献

- [Kuhn, 1973] Kuhn, H. W. (1973). A note on Fermat's problem. *Mathematical Programming*, 4(1):98–107.
- [Lugosi and Mendelson, 2016] Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.
- [Minsker, 2015] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335.
- [Minsker, 2016] Minsker, S. (2016). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *arXiv preprint arXiv:1605.07129*.
- [Minsker et al., 2014] Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (2014). Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.
- [Muandet et al., 2014] Muandet, K., Fukumizu, K., Sriperumbudur, B., Gretton, A., and Schölkopf,

- B. (2014). Kernel mean estimation and Stein effect. In *JMLR W&CP: Proceedings of The International Conference on Machine Learning*, volume 32, pages 10–18.
- [Muandet et al., 2016] Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. (2016). Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17(48):1–41.
- [Tolstikhin et al., 2016] Tolstikhin, I., Sriperumbudur, B., and Muandet, K. (2016). Minimax estimation of kernel mean embeddings. *arXiv preprint arXiv:1602.04361*.
- [Tsybakov, 2009] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.
- [Vardi and Zhang, 2000] Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1 -median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.
- [Weiszfeld, 1937] Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386.