Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms

Sep. 12, 2024

Contents

Data Valuation

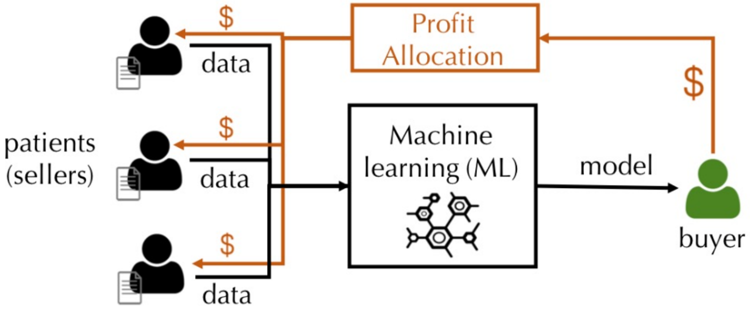

Figure 1. Motivating Example of Data Valuation.

Shapley Value

- $$s(v,i) = \frac{1}{N} \sum_{S \subseteq I \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{N-1}{|S|}}$$
- Group Rationality
- Fairness
- Additivity

KNN classifier

- Utility function:

$$\nu(S) = \frac{1}{K} \sum_{k=1}^{\min(K,|S|)} \mathbb{1}[y_{\alpha_k(S)} = y_{\text{test}}]$$

- Lemma 1:

$$s_i - s_j = \frac{1}{N-1} \sum_{S \subseteq I \setminus \{i,j\}, |S| = N-2} (\nu(S \cup \{i\}) - \nu(S \cup \{j\}))$$

SV in KNN

- Theorem 1:

$$s_{\alpha_N} = \frac{\mathbb{1}[y_{a_N} = y_{\text{test}}]}{N}$$

$$s_{\alpha_i} = s_{\alpha_{i+1}} + \frac{\left(\mathbb{1}[y_{\alpha_i} = y_{\text{test}}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{\text{test}}]\right)}{K} \frac{\min\{K, i\}}{i}$$

Exact SV Algorithm

Algorithm 1: Exact algorithm for calculating the SV for an unweighted KNN classifier.

input : Training data $D = \{(x_i, y_i)\}_{i=1}^{N}$, test data $D_{test} = \{(x_{test,i}, y_{test,i})\}_{i=1}^{N_{test}}$

output : The SV $\{s_i\}_{i=1}^{N}$

for $j \leftarrow 1$ **to** N_{test} **do**

  $(\alpha_1, ..., \alpha_N) \leftarrow$ Indices of training data in an ascending order using $d(\cdot, x_{test})$;

  $s_{j,\alpha_N} \leftarrow \frac{\mathbb{1}[y_{\alpha_N} = y_{test}]}{N}$;

  for $i \leftarrow N - 1$ **to** 1 **do**

    $s_{j,\alpha_i} \leftarrow s_{j,\alpha_{i+1}} + \frac{\mathbb{1}[y_{\alpha_i} = y_{test,j}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{test,j}]}{K} \frac{\min(K, i)}{i}$;

  end

end

for $i \leftarrow 1$ **to** N **do**

  $s_i \leftarrow \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} s_{j,i}$;

end

LSH based Approximation

- Theorem 2:

$$\widehat{s}_{\alpha_i} = 0 \qquad \textit{if } i \geqslant K^*$$

$$\widehat{s}_{\alpha_i} = \widehat{s}_{\alpha_{i+1}} + \frac{\left(\mathbb{1}[y_{\alpha_i} = y_{\text{test}}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{\text{test}}]\right)}{K} \frac{\min\{K, i\}}{i}$$

$$\textit{if } i \leqslant K^* - 1$$

- where $\mathrm{K}^* = \max\left\{K, \left\lceil \frac{1}{\epsilon} \right\rceil\right\}$ *for some* $\epsilon > 0$

Improved MC Approximation

- THEOREM 5. Given the range $[-r, r]$ of the utility difference ϕ_i, the sample size required such that
$$P\left[\|\hat{s} - s\|_\infty > \epsilon\right] \leqslant \delta$$
is $T > T^*$. T^* is the solution of
$$\sum_{i=1}^{N} \exp\left(-T^*(1-q_i^2)h\left((1-q_i^2)r\right)\right) = \frac{\delta}{2}.$$
where $h(u) = (1+u)\log(1+u) - u$ and
$$q_i = \begin{cases} 0, & i = 1, \ldots, K \\ \frac{i-K}{i}, & i = K+1, \ldots, N \end{cases}$$

Improved MC Approximation

Algorithm 2: Improved MC Approach

input : Training set - $D = \{(x_i, y_i)\}_{i=1}^{N}$, utility function $\nu(\cdot)$, the number of measurements - M, the number of permutations - T

output : The SV of each training point - $\hat{s} \in \mathbb{R}^N$

for $t \leftarrow 1$ **to** T **do**
    $\pi_t \leftarrow \mathrm{GenerateUniformRandomPermutation}(D)$;
    Initialize a length-K max-heap H to maintain the KNN;
    for $i \leftarrow 1$ **to** N **do**
        Insert $\pi_{t,i}$ to H;
        if H *changes* **then**
            $\phi^t_{\pi_{t,i}} \leftarrow \nu(\pi_{t,1:i}) - \nu(\pi_{t,1:i-1})$;
        else
            $\phi^t_{\pi_{t,i}} \leftarrow \phi^t_{\pi_{t,i-1}}$;
        end
    end
end
$\hat{s}_i = \frac{1}{T} \sum_{t=1}^{T} \phi^t_i$ for $i = 1, \ldots, N$;

Runtime for unweighted KNN

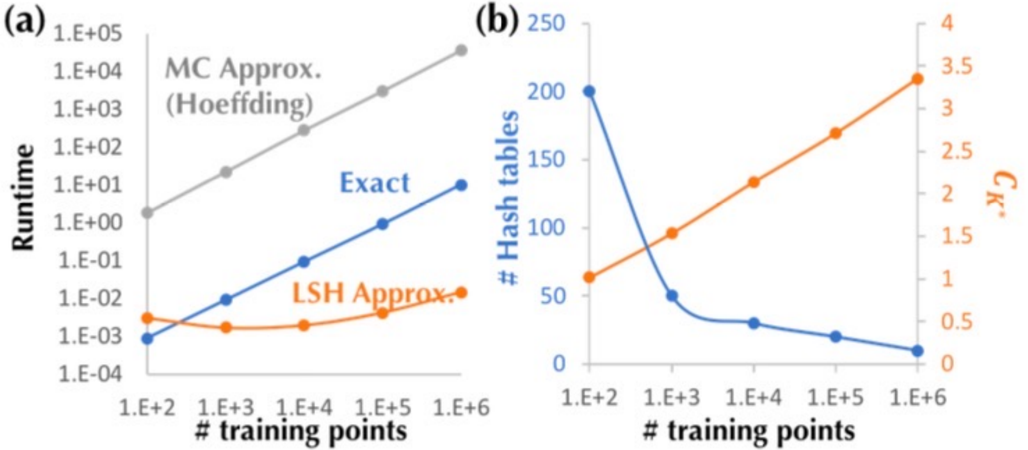

Figure 6. Performance of unweighted KNN classification in the single-data-per-seller case.

LSH on different datasets

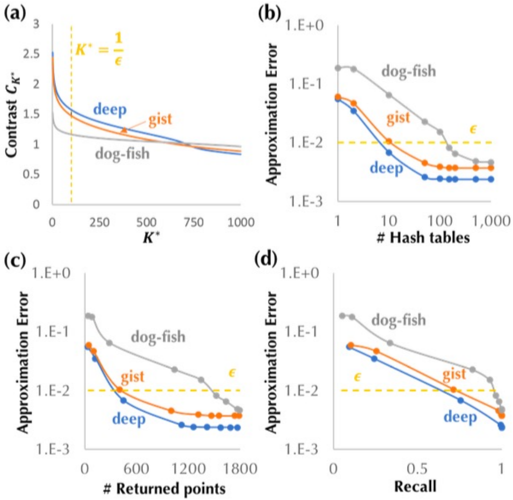

Figure 9. Performance of LSH on three datasets: `deep`, `gist`, `dog-fish`. (a) Relative contrast C_{K^*} vs. K^*. (b), (c) and (d) illustrate the trend of the SV approximation error for different number of hash tables, returned points and recalls.

Experiment on MC approximation

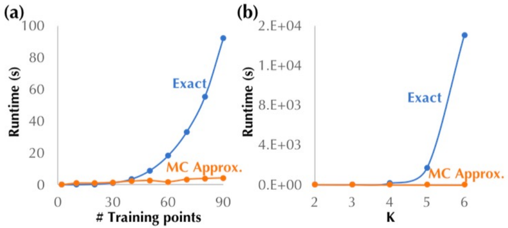

Figure 12. Performance of the weighted KNN classification.

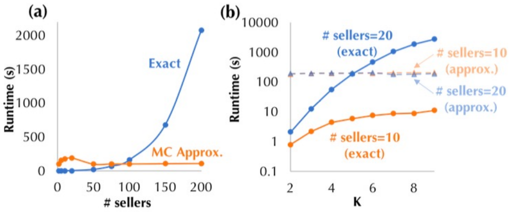

Figure 13. Performance of the KNN classification in the multi-data-per-seller case.

References

- R. Jia, et al., "Efficient task-specific data valuation for nearest neighbor algorithms," *Proceedings of the VLDB Endowment*, vol. 12, no. 11, pp. 1610–1623, 2019.