Finding Heavily-Weighted Features with the Weight-Median Sketch

Kai Sheng Tai, Vatsal Sharan, Peter Bailis, Gregory Valiant

Stanford University

Background

Sketches

- Sketches are useful for approximate query processing under memory constraints
- *Examples:* top-k frequent items, quantiles, # of distinct items

Memory-Constrained Online Learning

- ML on mobile devices, wearables, appliances: update models online on locally-observed data
- *Examples:* language modeling on phones, human activity classifiers on wearables

Problems

High-Dimensional Feature Spaces

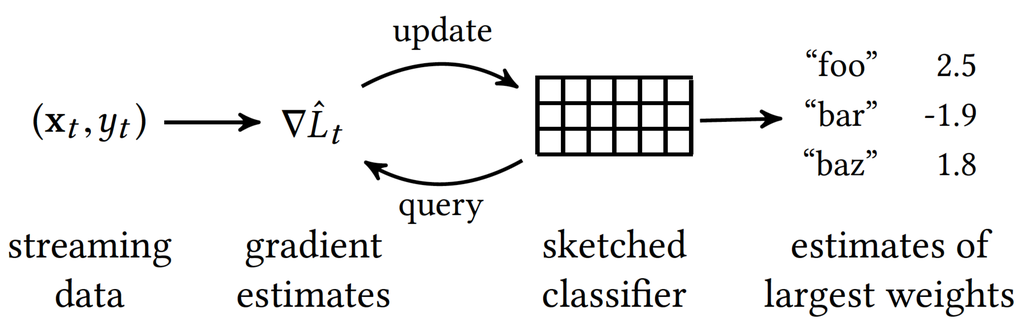
- Memory usage can increase over time as new features are observed
- *Example:* Spam classifier on text stream continually observes new n-grams
- Over time, models can grow to exceed memory constraints

Loss of Interpretability

- *Feature hashing* fixes the feature dimension, but at cost of interpretability
- Difficult to debug and perform analyses on learned weights

Paper: arXiv:1711.02305

The Weight-Median Sketch (WM-Sketch)

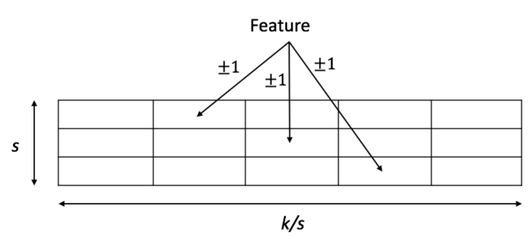


Sketched Linear Classifier

- Learns a compressed classifier within a fixed space budget
- Input feature vectors $\mathbf{x}_t$ compressed via *random projection*

Efficient Weight Recovery

- Supports efficient estimation of *high-magnitude weights*
- Identify most influential features in model



Algorithm

Random Projection

- Sparse Count-Sketch projection R
- Efficiently implemented via hashing

Updates

- Project example $\mathbf{x}_t \mapsto R\mathbf{x}_t$
- Update sketch state using gradient descent

Estimating weight of feature i

- Take median of buckets that feature i hashes to (+ random sign flips)
- Maintain heap of top-K weights

Theoretical Guarantees

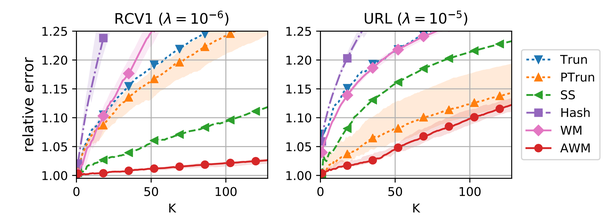
Theorem (informal) Given feature dimension d, failure probability δ, $\epsilon > 0$, set sketch size $k = O\left(\epsilon^{-4}\log^3(d/\delta)\right)$ and sketch depth $s = O\left(\epsilon^{-2}\log^2(d/\delta)\right)$. Then:

$$\|\mathbf{w}_* - \mathbf{w}_{\text{est}}\|_\infty \leq \epsilon \|\mathbf{w}_*\|_1,$$

where $\mathbf{w}_*$ is the optimal weight vector and $\mathbf{w}_{\text{est}}$ are the recovered weights.

Remark. Sketch size has only polylogarithmic dependence on d!

Evaluation



- Lower error in weight recovery than baselines
- Comparable classification accuracy with best baseline methods
- Consistent improvement in classification accuracy vs. feature hashing

Applications

Enables streaming analyses that can be formulated as classification tasks

- Online feature selection
- Streaming data explanation
- Detecting large relative differences between data streams
- Streaming pointwise mutual information for finding highly-correlated pairs

STANFORD DAWN