CS 591, Lecture 12
Data Analytics: Theory and Applications
Boston University

Charalampos E. Tsourakakis

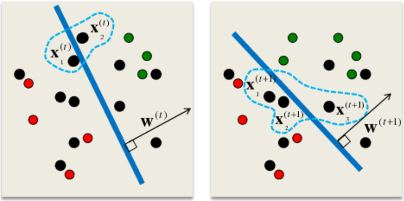
March 12th, 2017

Today's Outline

- I will follow Edith Cohen's survey on min-wise sketches [Cohen, 2016]
- Then, I will show you an example of bottom-k sketch (KMV sketch) [Bar-Yossef et al., 2002]
- Introduction to LSH (Jeffrey D. Ullman's slides, to be cont.)
- Very useful algorithms in analyzing massive datasets

LSH and Learning

LSH has many applications including machine learning applications!



Margin-based selection criterion for SVMs selects points nearest to current decision boundary: $\mathbf{x}^* = \arg\min_{\mathbf{x}_i \in \mathcal{U}} |\mathbf{w}^T \mathbf{x}_i|$ [Jain et al., 2010]

Today's problem: Distinct Elements

- Min-Hash sketches (aka Min-wise sketches) are **randomized summary structures of subsets**
- Mergeable/composable
- We have seen in detail (Lecture 5) an application of MinHash sketches (Flajolet-Martin distinct counting)
  Idea: $\mathbb{E}\left[\min(X_1, \ldots, X_n)\right] = \frac{1}{n+1}$

$\mathbb{E}\left[\min(X_1,\ldots,X_n)\right] = \frac{1}{n+1}$

Recall:

- $X_i \in U[0,1]$ for $i \in [n]$
- $Z = \min(X_1,\ldots,X_n)$

$$\mathbb{E}\left[Z\right] = \int_0^1 \mathbf{Pr}\left[Z > t\right]dt = \int_0^1 \mathbf{Pr}\left[X_1 > t\right]^n$$
$$= \int_0^1 (1-t)^n dt = \frac{1}{n+1}.$$

Set Operations in Sketch Space

Notation: Universe U, $|U| = n$, and $S(X)$ is the sketch of $X \subseteq lU|$.

- Insertion: Given a set X, and an element $y \in U$, the sketch $S(X \cup y)$ can be computed from $S(X)$ and y
- Merging sets: $S(X \cup Y)$ can be computed from $S(X), S(Y)$.

Question: Why are these two properties important for big data analytics?

Queries in Sketch Space

- Interested in $f(X)$ where X is a set
- We estimate $f(X)$ as $\tilde{f}(S(X))$
  Some queries supported by MinHash sketches:
- **Cardinality**: The number of elements in the $f(X) = |X|$.
- **Similarity**
  Jaccard coefficient: $f(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$
  Cosine similarity $f(X, Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}}$
- **Complex queries**: E.g., number of elements occuring in at least 2 sets

Constructions

Typical setting:

- Elements $x \in U$ are assigned random rank values
- Sketch $S(X)$ contains order statistics of the set of ranks
- Coordination of the sketches: when we sketch multiple sets, the same random rank assignment is common to all sketches

Constructions

1. **k-mins sketch**: k different rank functions $r_1, \ldots, r_k$, sketch $S(X) = (\tau_1, \ldots, \tau_k)$ where $\tau_i = \min_{y \in X} r_i(y)$
2. **k-partition sketch**: single rank function r, assignment $b : U \to [k]$ to k buckets, sketch $S(X) = (\tau_1, \ldots, \tau_k)$ where $\tau_i = \min_{y \in X, b(y)=i} r(y)$
3. **bottom k-sketch**: $\tau_1 < \ldots < \tau_k$ includes the k items with smallest rank in $\{r(y) : y \in X\}$

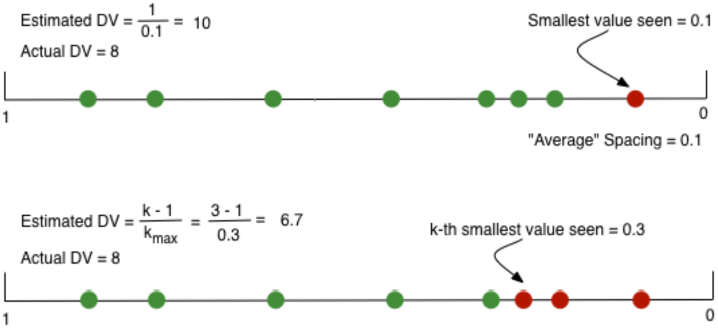
Question: What happens when $k = 1$?

Insertion and Merging

1. **k-mins sketch**: Compute rank y, compare coordinate-wise with sketch
2. **k-partition sketch**: Compute $r(y), b(y)$. Compare $r(y)$ and $\tau_{b(y)}$ to decide if we should update the value
3. **bottom k-sketch**: Compute $r(y)$, compare to τ_k

Questions: (a) How to merge? (b) Run times?

Example of a bottom k-sketch (KMV sketch)



Source: Neustar

Question: Why this estimator?

K-Minimum Values (KMV sketch)

Demo

Locality Sensitive Hashing

Ullman's slides, part I
Ullman's slides, part II

references I

Bar-Yossef, Z., Jayram, T., Kumar, R., Sivakumar, D., and Trevisan, L. (2002).

Counting distinct elements in a data stream.

In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 1–10. Springer.

Cohen, E. (2016).

Min-hash sketches.

Jain, P., Vijayanarasimhan, S., and Grauman, K. (2010).

Hashing hyperplane queries to near points with applications to large-scale active learning.

In *Advances in Neural Information Processing Systems*, pages 928–936.