Fully Understanding the Hashing Trick

Lior Kamma, Aarhus University

Joint work with Casper Freksen and Kasper Green Larsen.
Recommendation and Classification

PG-13
Comic Book
Super Hero
Sci Fi
Adventure
Action

Violent
Scary
Comedy
Drama
Horror

Fully Understanding the Hashing Trick
Recommendation and Classification

PG-13
Comic Book
Super Hero
Sci Fi
Adventure
Action

Violent
Scary
Comedy
Drama
Horror

Categorical Variables

How do we decide these are “close”?
Feature Vectors

Denote the feature dimension by $n$
Storing a corpus of $M$ items requires $\Omega(nM)$ memory.
$k$-Nearest Neighbours

**New Movie**

How do we find the $k$ closest movies?
Dimensionality Reduction

- Given $\varepsilon, \delta \in (0,1)$ find

Approximation Ratio

Error Probability
Dimensionality Reduction

- Given $\varepsilon, \delta \in (0,1)$ find random $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that for every $x, y \in \mathbb{R}^n$

For some small $m$

Think of $n$ as HUGE
Dimensionality Reduction

Given $\varepsilon, \delta \in (0,1)$ find random $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that for every $x, y \in \mathbb{R}^n$

$$\Pr[\|f(x) - f(y)\|_2^2 \in (1 \pm \varepsilon)\|x - y\|_2^2] \geq 1 - \delta$$
Dimensionality Reduction

Given $\epsilon, \delta \in (0, 1)$ find random $A \in \mathbb{R}^{m \times n}$ such that for every $x, y \in \mathbb{R}^n$

$$\Pr[\|A(x - y)\|_2^2 \in (1 \pm \epsilon)\|x - y\|_2^2] \geq 1 - \delta$$

Why linear?
- Cool Math
- Streaming (updates).
- Good in practice
Dimensionality Reduction

- Given $\varepsilon, \delta \in (0,1)$ find random $A \in \mathbb{R}^{m \times n}$ such that for every $x \in \mathbb{R}^n$

$$
\Pr[\|A(x)\|_2^2 \in (1 \pm \varepsilon)\|x\|_2^2] \geq 1 - \delta
$$

Why linear?
- Cool Math
- Streaming (updates).
- Good in practice

Focus on linear projections
Johnson Lindenstrauss Lemma [JL’84]

Given $\varepsilon, \delta \in (0,1)$ there exists a random linear $A \in \mathbb{R}^{m \times n}$ such that for every $x$

$$\Pr[\|A(x)\|_2^2 \in (1 \pm \varepsilon)\|x\|_2^2] \geq 1 - \delta$$

$m = O\left(\frac{\lg 1/\delta}{\varepsilon^2}\right)$

In most proofs matrix is as dense as possible. Embedding takes $O(mn)$ operations.
Johnson Lindenstrauss Lemma [JL’84]

Given $\varepsilon, \delta \in (0,1)$ there exists a random linear $A \in \mathbb{R}^{m \times n}$ such that for every $x$

$$\Pr[\|A(x)\|_2^2 \in (1 \pm \varepsilon)\|x\|_2^2] \geq 1 - \delta$$

If $A$ is sparse, this can be made faster.

In most proofs matrix is as dense as possible. Embedding takes $O(mn)$ operations.
Feature Hashing [Weinberger et al. 2009]

General Idea: Shuffle the entries of $x$

Add random signs

$x$

$(1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1)$
Feature Hashing [Weinberger et al. 2009]

**General Idea:** Shuffle the entries of $x$.

**Add random signs**

**$f(x)$**

1 0 1 1 0 0 1 0 1

$m = 3$

Fully Understanding the Hashing Trick
Feature Hashing [Weinberger et al. 2009]

Add random signs

General Idea: Shuffle the entries of $x$

$x$  
(1 0 1 1 0 0 1 0 1 1)

$+ -$

$f(x)$  
0 1 $-1$

$m = 3$

Fully Understanding the Hashing Trick
Feature Hashing [Weinberger et al. 2009]

General Idea: Shuffle the entries of \( x \)

Add random signs

**Observation:** This operation is linear. Moreover, every column has exactly one non-zero entry.

\[ f(x) \]

\[ m = 3 \]

Fully Understanding the Hashing Trick
The Hashing Trick – With High Prob.

- Observation: If $m$ is large enough, and the “mass” of $x$ is not concentrated in few entries, then the trick works with high probability.

- $\varepsilon = 0.1$

- $\begin{pmatrix}1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

- $\frac{\|x\|_\infty}{\|x\|_2} = \frac{1}{\sqrt{2}}$.

- $\Pr_{h:\{1,2,...,n\}\rightarrow\{1,2,...,m\}}[h(1) = h(2)] = \frac{1}{m}$
The Hashing Trick – With High Prob.

Success iff no collision occurs

\[ \varepsilon = 0.1 \]

\[
\begin{pmatrix}
1 \\
1 \\
0 \\
\vdots \\
0
\end{pmatrix}
\]

\[
\frac{\|x\|_\infty}{\|x\|_2} = \frac{1}{\sqrt{n}}
\]

\[
\Pr_{h:\{1,2,\ldots,n\} \rightarrow \{1,2,\ldots,m\}} [h(1) = h(2)] = \frac{1}{m}
\]

To succeed we need \( m \geq \frac{1}{\delta} \)
Tight Bounds – Formal Problem

- Fix $m, \varepsilon, \delta$.
- Define $\nu(m, \varepsilon, \delta)$ to be the maximum $\nu$ such that whenever $\|x\|_\infty \leq \nu \|x\|_2$ then feature hashing works.
Tight Bounds – Formal Problem

- Fix $m, \varepsilon, \delta$.
- Define $\nu(m, \varepsilon, \delta)$ to be the maximum $\nu$ such that whenever $\|x\|_\infty \leq \nu \|x\|_2$ then feature hashing works.

We have a fixed budget, and a fixed room for error.

Evaluating $\nu$ has been an open question for almost a decade.
Tight Bounds – Our Result

- Fix $m, \varepsilon, \delta$.

Theorem.

1. If $m < \frac{c \log \frac{1}{\delta}}{\varepsilon^2}$ then $\nu = 0$.

Essentially, this means our budget is too small to do anything meaningful.
Tight Bounds – Our Result

- Fix $m, \varepsilon, \delta$.

**Theorem.**

1. If $m < \frac{c \log \frac{1}{\delta}}{\varepsilon^2}$ then $\nu = 0$.
2. If $m \geq \frac{2}{\delta \varepsilon^2}$ then $\nu = 1$.

Essentially, this means our budget is rich enough to do anything.
Tight Bounds – Our Result

- Fix $m, \varepsilon, \delta$.

Theorem.

1. If $m < C \log \frac{1}{\varepsilon^2 \delta}$ then $\nu = 0$.

2. If $m \geq 2 \varepsilon^2 \delta \nu^2$ then $\nu = 1$.

3. If $C \log \frac{1}{\varepsilon^2 \delta} \leq m < \frac{1}{\delta \varepsilon^2}$ then

\[ \nu = \Theta \left( \sqrt{\varepsilon} \cdot \min \left\{ \frac{\log \frac{\varepsilon m}{\log \frac{1}{\delta}}}{\log \frac{1}{\delta}}, \sqrt{\frac{\log \frac{\varepsilon^2 m}{\log \frac{1}{\delta}}}{\log \frac{1}{\delta}}} \right\} \right) \]

This is tight, which means this is the right expression.
Empirical Analysis

Results show that the $\Theta$-constant is close to $1$.

This implies that Feature Hashing’s performance can be very well predicted in practice using our formula.

$$
\nu = \Theta \left( \sqrt{\varepsilon} \cdot \min \left\{ \frac{\log \frac{\varepsilon m}{1/\delta}}{\log \frac{1}{\delta}}, \sqrt{\frac{\log \frac{\varepsilon^2 m}{1/\delta}}{\log \frac{1}{\delta}}} \right\} \right)
$$

$\sqrt{\varepsilon} \min \left\{ \frac{\log \frac{\varepsilon m}{1/\delta}}{\log \frac{1}{\delta}}, \sqrt{\frac{\log \frac{\varepsilon^2 m}{1/\delta}}{\log \frac{1}{\delta}}} \right\}$

0.725
Questions?

Come see poster

Read the paper

Talk offline

All of the above

Tight Cell-Probe Bounds for Succinct Boolean Matrix-Vector Multiplication
Questions?

Come see poster

Read the paper

Talk offline

All of the above

Thank you