CIS 700: "algorithms for Big Data"

Lecture 5: Dimension Reduction

Slides at http://grigory.us/big-data-class.html

Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

Today

- Dimensionality reduction
  - AMS as dimensionality reduction
  - Johnson-Lindenstrauss transform

L_p-norm Estimation

- Stream: $\boldsymbol{m}$ updates $(x_i, \Delta_i) \in [n] \times \mathbb{R}$ that define vector f where $f_j = \sum_{i:x_i=j} \Delta_i$.
- Example: For $n = 4$

$$\langle (1,3), (3, 0.5), (1,2), (2, -2), (2,1), (1, -1), (4,1) \rangle$$
$$f = (4, -1, 0.5, 1)$$

- L_p-norm: $||f||_p = \left(\sum_i |f|^p\right)^{\frac{1}{p}}$

L_p-norm Estimation

- L_p-norm: $||f||_p = (\sum_i |f|^p)^{\frac{1}{p}}$
- Two lectures ago:
  - $||f||_0 = F_0$-moment
  - $||f||_2^2 = F_2$-moment (via AMS sketching)
- Space: $O\left(\frac{\log n}{\epsilon^2} \log \frac{1}{\delta}\right)$
- Technique: linear sketches
  - $||f||_0$: $\sum_{i \in S} f_i$ for random sets S
  - $||f||_2^2$: $\sum_i \sigma_i f_i$ for random signs σ_i

AMS as dimensionality reduction

- Maintain a "linear sketch" vector

$$\boldsymbol{Z} = (Z_1, \dots, Z_k) = Rf$$

$$Z_i = \sum_{j \in [n]} \sigma_{ij} f_j, \quad \text{where } \sigma_{ij} \in_R \{-1,1\}$$

- Estimator $\boldsymbol{Y}$ for $||f||_2^2$:

$$\frac{1}{k} \sum_{i=1}^{k} Z_i^2 = \frac{||Rf||_2^2}{k}$$

- "Dimensionality reduction": $x \to Rx$, "heavy" tail

$$\Pr\left[\left|\boldsymbol{Y} - ||f||_2^2\right| \geq c \left(\frac{2}{k}\right)^{\frac{1}{2}} ||f||_2^2\right] \leq \frac{1}{c^2}$$

Normal Distribution

- Normal distribution $N(0,1)$
  - Range: $(-\infty, +\infty)$
  - Density: $\mu(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$
  - Mean = 0, Variance = 1
- Basic facts:
  - If X and Y are independent r.v. with normal distribution then $X + Y$ has normal distribution
  - $Var[cX] = c^2 Var[X]$
  - If X, Y are independent, then $Var[X + Y] = Var[X] + Var[Y]$

Johnson-Lindenstrauss Transform

- Instead of ± 1 let σ_i be i.i.d. random variables from normal distribution $N(0,1)$

$$Z = \sum_i \sigma_i f_i$$

- We still have $\mathbb{E}[Z^2] = \sum_i f_i^2 = ||f||_2^2$ because:
  - $\mathbb{E}[\sigma_i]\mathbb{E}[\sigma_j] = 0$; $\mathbb{E}[\sigma_i^2] =$ "variance of σ_i " $= 1$
- Define $\boldsymbol{Z} = (Z_1, \dots, Z_k)$ and define:

$$||\boldsymbol{Z}||_2^2 = \sum_j Z_j^2 \quad \left(\mathbb{E}\left[||\boldsymbol{Z}||_2^2\right] = k||f||_2^2\right)$$

- JL Lemma: There exists $C > 0$ s.t. for small enough $\epsilon > 0$:

$$\Pr\left[\left|||\boldsymbol{Z}||_2^2 - k\,||f||_2^2\right| > \epsilon k||f||_2^2\right] \le \exp(-C\epsilon^2 k)$$

Proof of JL Lemma

- JL Lemma: $\exists C > 0$ s.t. for small enough $\epsilon > 0$:

$$\Pr\left[\left|\|\boldsymbol{Z}\|_2^2 - k\,\|f\|_2^2\right| > \epsilon k\|f\|_2^2\right] \le \exp(-C\epsilon^2 k)$$

- Assume $\|f\|_2^2 = 1$.
- We have $\boldsymbol{Z}_i = \sum_j \sigma_{ij} f_i$ and $\boldsymbol{Z} = (\boldsymbol{Z_1}, \dots, \boldsymbol{Z_k})$

$$\mathbb{E}\left[\|\boldsymbol{Z}\|_2^2\right] = k\|f\|_2^2 = k$$

- Alternative form of JL Lemma:

$$\Pr\left[\|\boldsymbol{Z}\|_2^2 > k(1+\epsilon)^2\right] \le \exp(-\epsilon^2 k + O(k\,\epsilon^3))$$

Proof of JL Lemma

- Alternative form of JL Lemma:

$$\Pr\left[\left|\left|\boldsymbol{Z}\right|\right|_2^2 > k(1+\epsilon)^2\right] \le \exp(-\epsilon^2 k + O(k\,\epsilon^3))$$

- Let $\boldsymbol{Y} = \left|\left|\boldsymbol{Z}\right|\right|_2^2$ and $\alpha = k(1+\epsilon)^2$
- For every $\boldsymbol{s} > 0$ we have:

$$\Pr[\boldsymbol{Y} > \alpha] = \Pr[e^{\boldsymbol{sY}} > e^{\boldsymbol{s}\alpha}]$$

- By Markov and independence of $\boldsymbol{Z}_i's$:

$$\Pr[e^{\boldsymbol{sY}} > e^{\boldsymbol{s}\alpha}] \le \frac{\mathbb{E}[e^{\boldsymbol{sY}}]}{e^{\boldsymbol{s}\alpha}} = e^{-\boldsymbol{s}\alpha}\mathbb{E}\left[e^{\boldsymbol{s}\sum_i \boldsymbol{Z}_i^2}\right] = e^{-\boldsymbol{s}\alpha}\prod_{i=1}^{k}\mathbb{E}\left[e^{\boldsymbol{s}\boldsymbol{Z}_i^2}\right]$$

- We have $Z_i \in N(0,1)$, hence:

$$\mathbb{E}\left[e^{\boldsymbol{s}\boldsymbol{Z}_i^2}\right] = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{\infty} e^{\boldsymbol{s}t^2} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{1-2\boldsymbol{s}}}$$

Proof of JL Lemma

- Alternative form of JL Lemma:

$$\Pr\left[\left||\boldsymbol{Z}|\right|_2^2 > k(1+\epsilon)^2\right] \le \exp(-\epsilon^2 k + O(k\,\epsilon^3))$$

- For every $\boldsymbol{s} > 0$ we have:

$$\Pr[\boldsymbol{Y} > \alpha] \le e^{-\boldsymbol{s}\alpha} \prod_{i=1}^{k} \mathbb{E}\left[e^{\boldsymbol{s}\boldsymbol{Z}_i^2}\right] = e^{-\boldsymbol{s}\alpha}(1 - 2\boldsymbol{s})^{-\frac{k}{2}}$$

- Let $\boldsymbol{s} = \frac{1}{2}\left(1 - \frac{k}{\alpha}\right)$ and recall that $\alpha = k(1+\epsilon)^2$
- A calculation finishes the proof:

$$\Pr[\boldsymbol{Y} > \alpha] \le \exp(-\epsilon^2 k + O(k\,\epsilon^3))$$

Johnson-Lindenstrauss Transform

- Single vector: $k = O\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$
  - Tight: $k = \Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ [Woodruff'10]
- $\boldsymbol{n}$ vectors simultaneously: $k = O\left(\frac{\log \boldsymbol{n} \log\frac{1}{\delta}}{\epsilon^2}\right)$
  - Tight: $k = \Omega\left(\frac{\log \boldsymbol{n} \log\frac{1}{\delta}}{\epsilon^2}\right)$ [Molinaro, Woodruff, Y. '13]
- Distances between $\boldsymbol{n}$ vectors = $O(\boldsymbol{n}^2)$ vectors:

$$k = O\left(\frac{\log \boldsymbol{n} \log\frac{1}{\delta}}{\epsilon^2}\right)$$