Lecture 9

Dimensionality Reduction: Johnson-Lindenstrauss Lemma

Course: Algorithms for Big Data

Instructor: Hossein Jowhari

Department of Computer Science and Statistics
Faculty of Mathematics
K. N. Toosi University of Technology

Spring 2021

Distance-Preserving Dimensionality Reduction

Given n vectors $A = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\mathbb{R}^d$,

we want a fast transformation $f : \mathbb{R}^d \to \mathbb{R}^t$ so that

- (Distances are approximately preserved):

  For all pairs $\boldsymbol{x}, \boldsymbol{y} \in A$, we have $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \approx \|\boldsymbol{x} - \boldsymbol{y}\|$

- (The dimension is reduced considerably)

  $t \ll d$

Such a transformation will be very useful in practice (when the input dimension is large)

Two applications

Faster clustering algorithms: As an example, recall that for the k-center problem we had a 2-factor approximation algorithm with running time $O(nkd)$ (when the points lie in $\mathbb{R}^d$.)

Using the transformation $f : \mathbb{R}^d \to \mathbb{R}^t$, we first compute $f(x)$ for all $x \in A$

Then we run the k-center alg. on $A' = \{f(x_1), \ldots, f(x_n)\}$

$$\text{Running time} = \underbrace{\text{time of transformation}} + \underbrace{\text{time of clustering } A'}_{O(nkt)}$$

Approximation quality ≈ 2

Approximate nearest neighbor queries: Most exact nearest neighbor data structures have time complexity $n^{O(d)}$

How to reduce the dimension?

- A bad idea: Randomly select a small subset of dimensions. Restrict every $\boldsymbol{x}$ to the selected dimensions. In other words, choose $S \subseteq \{1, \ldots, d\}$ randomly. $f(\boldsymbol{x}) = \boldsymbol{x}_S$

  Example:

  $$\boldsymbol{x}_1 = \overbrace{(1, 0, \ldots, 0)}^{d}, \boldsymbol{x}_2 = \overbrace{(0, 1, 0, \ldots, 0)}^{d}, \ \|x - y\| = 1$$

  $$S = \{3, 5, 20, 25\}$$

  $$f(\boldsymbol{x}_1) = (0, 0, 0, 0), f(\boldsymbol{x}_2) = (0, 0, 0, 0), \ \|f(x) - f(y)\| = 0$$

- Dimensionality reduction methods such as PCA do not preserve the pairwise distances.

Johnson-Lindenstrauss Lemma

JL Lemma (existential formulation): Let $\epsilon \in (0, \frac{1}{2})$. Given a set of n vectors $A = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^d$, there is a mapping $f : \mathbb{R}^d \to \mathbb{R}^t$ where $t = O(\frac{\log n}{\epsilon^2})$ where

$$\forall x, y \in A,\ (1-\epsilon)\|\boldsymbol{x}-\boldsymbol{y}\| \leq \|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq (1+\epsilon)\|\boldsymbol{x}-\boldsymbol{y}\|$$

JL lemma is essentially tight with respect to the target dimension t:

Noga Alon has shown that any such mapping requires $t = \Omega(\frac{\log n}{\epsilon^2 \log \frac{1}{\epsilon}})$

Larsen and Nelson have shown that any linear mapping requires $t = \Omega(\frac{\log n}{\epsilon^2})$

Construction of the mapping f: Let $\boldsymbol{M}$ be a $t \times d$ matrix where every entry M_{ij} is an independent random sample from the normal standard distribution $N(0,1)$. In other words, each $M_{ij} \sim N(0,1)$

$$\frac{1}{\sqrt{t}} \underbrace{\begin{bmatrix} M_{11} & \dots & M_{1d} \\ M_{21} & \dots & M_{2d} \\ \vdots & \vdots & \vdots \\ M_{t1} & \dots & M_{td} \end{bmatrix}}_{\boldsymbol{M}} \underbrace{\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{d1} & x_{d2} & \dots & x_{dn} \end{bmatrix}}_{\boldsymbol{A}} = \begin{bmatrix} \vdots & \vdots & \vdots \\ f(\boldsymbol{x}_1) & \vdots & f(\boldsymbol{x}_n) \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Let $\boldsymbol{M}^{(i)}$ be the i-th row of $\boldsymbol{M}$.

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{t}} \boldsymbol{M}\boldsymbol{x} = \left(\frac{1}{\sqrt{t}} \boldsymbol{M}^{(1)}.\boldsymbol{x}, \ \dots \ , \frac{1}{\sqrt{t}} \boldsymbol{M}^{(t)}.\boldsymbol{x}\right)$$

An example: $n = 5, d = 7, t = 4$

$$\frac{1}{\sqrt{4}}\underbrace{\begin{bmatrix} -0.23 & -0.02 & -0.22 & -0.68 & +0.39 & +0.24 & +0.36 \\ +0.08 & +0.46 & +0.68 & +0.47 & -0.28 & +1.90 & +1.13 \\ +0.89 & -0.24 & +0.83 & +1.92 & -0.47 & +0.10 & +0.33 \\ +0.47 & -1.42 & -1.09 & +2.27 & -0.90 & -0.99 & -0.11 \end{bmatrix}}_{\boldsymbol{M}} \underbrace{\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{A}} = \begin{bmatrix} 0 & -0.11 & -0.24 & -0.38 & -0.08 \\ 0 & +0.04 & +0.62 & +0.71 & +2.23 \\ 0 & +0.44 & +0.74 & +1.46 & +1.68 \\ 0 & +0.23 & -1.01 & -0.33 & -0.88 \end{bmatrix}$$

Lemma 1: If $t \geq \frac{c \log n}{\epsilon^2}$ for a large enough constant c then with probability at least $3/4$ for all pairs $\boldsymbol{x}, \boldsymbol{y} \in A$, we have

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \in \left[(1-\epsilon)\|\boldsymbol{x} - \boldsymbol{y}\|, (1+\epsilon)\|\boldsymbol{x} - \boldsymbol{y}\|\right]$$

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \approx_\epsilon \|\boldsymbol{x} - \boldsymbol{y}\|$$

Lets consider a special case:

$A = \{\mathbf{0}, \boldsymbol{x}\}$ where $\boldsymbol{x}$ is a unit vector in $\mathbb{R}^n$. $\|\boldsymbol{x}\| = 1$.

We want $\|f(\boldsymbol{x}) - f(\mathbf{0})\| = \|f(\boldsymbol{x})\| \approx_\epsilon 1$.

Lemma 2: Given $\boldsymbol{x} \in \mathbb{R}^d$ where $\|x\| = 1$, assuming $t \geq \frac{c\log(\frac{1}{\delta})}{\epsilon^2}$ when c is a large enough constant then we have

$$Pr\Big(\|f(\boldsymbol{x})\|^2 \in [1-\epsilon, 1+\epsilon]\Big) \geq 1-\delta$$

Before proving Lemma 2, we show Lemma 1 is a consequence of Lemma 2.

Observation 1: Since $f(\boldsymbol{x}) - f(\boldsymbol{y}) = f(\boldsymbol{x} - \boldsymbol{y})$ (the mapping f is linear) then it is enough to show that for any arbitrary vector $\boldsymbol{z} \in \mathbb{R}^d$ we have

$$\|f(\boldsymbol{z})\| \approx_\epsilon \|\boldsymbol{z}\|$$

Observation 2: Let $\boldsymbol{z}' = \frac{z}{\|z\|}$. The vector $\boldsymbol{z}'$ is a unit vector. If we have $\|f(\boldsymbol{z}')\| \approx_\epsilon 1$ then (by linearity of f) we have

$$\|f(\boldsymbol{z})\| = \|f(\|z\|z')\| = \|\|z\|f(\boldsymbol{z}')\| = \|z\|\|f(\boldsymbol{z}')\| \approx_\epsilon \|z\|$$

Observation 3: There are $\binom{n}{2}$ pair of vectors in A. From Lemma 2 and the above observations we have for a pair $\boldsymbol{x}, \boldsymbol{y} \in A$, if $t \geq \frac{c \log(\frac{1}{\delta})}{\epsilon^2}$ then $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \approx_\epsilon \|\boldsymbol{x} - \boldsymbol{y}\|$ with probability $1 - \delta$.

Setting $\delta = \frac{1}{4n^2}$, from the union bound, the statement is true for all pairs in A with probability at least $1 - \binom{n}{2}\frac{1}{4n^2} > 3/4$.

Therefore we get the statement of Lemma 1.

Proof of Lemma 2

Basic facts regarding Gaussian distribution:

- The Gaussian distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2 has the following probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{x-\mu^2}{2\sigma^2}}$$

- If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ then

$$cX \sim N(c\mu_1, c^2\sigma_1^2)$$

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Consider $\boldsymbol{x} \in \mathbb{R}^d$ where $\|\boldsymbol{x}\| = 1$.

Let Y_i be the i-th oordinate of $\boldsymbol{Mx}$. Note that $f(\boldsymbol{x}) = \frac{1}{\sqrt{t}}\boldsymbol{Mx}$.

Observation: $Y_i = (G_1, \ldots, G_d).\boldsymbol{x}$ where each G_i is an independent sample from $N(0,1)$. In other words,

$$Y = G_1 x_1 + \ldots + G_d x_d$$

Y_i is a linear combination of independent Gaussians. Therefore

$$Y_i \sim N(0, x_1^2 + \ldots + x_d^2) = N(0,1)$$

We need to analyze Y_i^2 since $Y = \|f(x)\|^2 = \frac{1}{t}(Y_1^2 + \ldots + Y_t^2)$

$$E[Y_i^2] = Var[Y_i] + E^2[Y_i] = 1 \;\; \Rightarrow \;\; E[Y] = 1$$

So, in expectation, $Y = \|f(\boldsymbol{x})\|$ is exactly 1. Very good but not enough.

We need to bound the probability $Pr(Y > 1 + \epsilon)$.

$Y = \frac{1}{t}\sum_i^t Y_i^2$ is the sum of independent random variables. We could use Chernoff but unfortunately Y_i is not bounded. Still similar ideas that were used in the proof of Chernoff are helpful here. For any $r > 0$

$$Pr(Y > 1+\epsilon) = Pr(e^{trY} > e^{tr(1+\epsilon)}) \leq \overbrace{\frac{E[e^{trY}]}{e^{tr(1+\epsilon)}}}^{\text{Markov Inequality}} = \prod_{i=1}^{t} \frac{E[e^{rY_i^2}]}{e^{r(1+\epsilon)}}$$

$$E[e^{rY_i^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{ry^2} e^{-\frac{y^2}{2}} dy = \frac{1}{\sqrt{1-2r}} \quad \text{when } r \leq \frac{1}{2}$$

Therefore for $0 < r < \frac{1}{2}$, we have

$$Pr(Y > 1 + \epsilon) \leq \left(\frac{1}{e^{r(1+\epsilon)}\sqrt{1-2r}}\right)^t$$

Also one can show that $\frac{1}{e^{r(1+\epsilon)}\sqrt{1-2r}} \leq e^{\frac{r^2}{1-2r}}$

Therefore we have

$$Pr(Y > 1 + \epsilon) \leq e^{\frac{tr^2}{1-2r}}$$

We set $r = \frac{\epsilon}{4}$. Using $1 - 2r \geq \frac{1}{2}$ when $\epsilon \leq \frac{1}{2}$, we get

$$Pr(Y > 1 + \epsilon) \leq e^{-\frac{t\epsilon^2}{8}} \leq \frac{\delta}{2} \;\; \Rightarrow \;\; t \geq \frac{8}{\epsilon^2}\ln(\frac{2}{\delta})$$

Similarly we can show

$$Pr(Y < 1 - \epsilon) \leq e^{-\frac{t\epsilon^2}{8}} \leq \frac{\delta}{2}$$

Therefore having $t \geq \frac{8}{\epsilon^2}\ln(\frac{2}{\delta})$ we get

$$Pr(1 - \epsilon \leq Y \leq 1 + \epsilon) \geq 1 - \delta$$

Good news: JL lemma still holds when the Gaussian distribution $N(0,1)$ is replaced with random $-1,+1$ coefficients.