CPSC 665 : An Algorithmist's Toolkit Lecture 3 : 16 Jan 2015

Dimension Reduction

Lecturer: Sushant Sachdeva *Scribe: Xiao Shi*

1. Introduction

The "curse of dimensionality" refers to various phenomena that arise when analyzing and organizing data in high dimensions. For example, the solution to nearest neighbor problem grows exponentially with the dimension. Therefore dimension reduction, the process of representing data in lower dimensions while preserving the essential properties, is very useful. Common techniques include Singular Value Decomposition (SVD). This lecture covers the Johnson-Lindenstrauss Lemma and how to preserve distance information in data.

2. Problem Statement

We first consider whether we could preserve the distance under dimension reduction.

HW 1. *Prove: given $\mathcal{X} = \{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_n\} \in \mathbb{R}^n$ where $\mathbf{e}_i$ is the standard basis, $\exists f : \mathcal{X} \mapsto \mathbb{R}^d$ such that $\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\| = \|\mathbf{x}_i - \mathbf{x}_j\| \Rightarrow d \geq n$.*

The above proposition states that to maintain the exact distance for $\mathbb{R}^n$ vectors, we need a space with no fewer dimensions. How about approximation?

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^n$, does there exist a function $f : \mathcal{X} \mapsto \mathbb{R}^d$ such that $d \ll n$ and $\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\| \approx \|\mathbf{x}_i - \mathbf{x}_j\|$?

The distance here strictly refers to the L^2 norm.

3. The Johnson-Lindenstrauss Lemma

Lemma 3.1 (Johnson-Lindenstrauss Lemma)**.** *For any $\varepsilon \in (0, 1/2]$ and $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^n$, there exists a linear mapping $L : \mathbb{R}^n \mapsto \mathbb{R}^d$ for $d = O(\varepsilon^{-2} \log n)$ such that*

$$(*) \qquad \forall i, j \quad (1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\| \leq \|L\mathbf{x}_i - L\mathbf{x}_j\| \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|$$

Lemma 3.2. *There is a poly-time samplable distribution on L such that L satisfies (*) with probability at least $1 - 1/n$.*

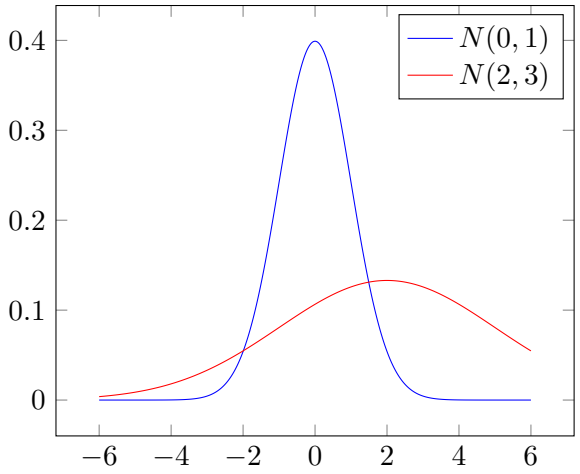
4. Proof of the Johnson-Lindenstrauss Lemma

4.1. **One-dimensional Gaussians.** $g \sim N(\mu, \sigma^2)$ is the Gaussian (normal) distribution, where $\mu \in \mathbb{R}$ is the mean of the distribution and $\sigma^2 \in \mathbb{R}^+$ is the variance. The probability distribution function of g is

$$p(x) = 1/\sigma\sqrt{2\pi}\exp(-(x-\mu)^2/2\sigma^2).$$

$g \sim N(0,1)$ is the unit Gaussian with $p(x) = 1/\sqrt{2\pi}exp(-x^2/2)$. In general, we have $g \sim N(0,1) \Rightarrow \mu + \sigma g \sim N(\mu, \sigma^2)$.

If $g_1 \sim N(0, \sigma_1^2)$ and $g_2 \sim N(0, \sigma_2^2)$ are independent Gaussians, then $g_1 + g_2 \sim N(0, \sigma_1^2 + \sigma_2^2)$. (Can be proved by using the characteristic functions.)



4.2. **Main lemma.**

Lemma 4.1. *Let R be a $d \times n$ matrix such that R_{ij} is an independent unit Gaussian $N(0,1)$, then for a fixed unit vector $\mathbf{v} \in \mathbb{R}^n$ (i.e. $\|\mathbf{v}\| = 1$), for $d = O\left(1/\varepsilon^2 \log 1/\delta\right)$,*

$$(**) \qquad \mathbf{Pr}\left[1 - \varepsilon \le \frac{\|R\mathbf{v}\|}{\sqrt{d}} \le 1 + \varepsilon\right] \ge 1 - \delta$$

4.2.1. *Implication.* This lemma implies the Johnson-Lindenstrauss Lemma.

First we notice that $\mathbf{v}$ in this lemma is a unit vector. We want to drop the "unit" constraint by simply adding $\|\mathbf{v}\|$:

$$\mathbf{Pr}\left[(1-\varepsilon)\|\mathbf{v}\| \le \frac{\|R\mathbf{v}\|}{\sqrt{d}} \le (1+\varepsilon)\|\mathbf{v}\|\right] \ge 1 - \delta.$$

Let $L = 1/\sqrt{d}R$, and $\mathbf{v} \in \{\mathbf{x}_i - \mathbf{x}_j, 1 \le i < j \le n\}$ ($\binom{n}{2}$ vectors),

$$(**) \Rightarrow \forall \text{ fixed } i < j, \quad \mathbf{Pr}\left[(1-\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\| \le \left\|L\mathbf{x}_i - L\mathbf{x}_j\right\| \le (1+\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right] \ge 1 - \delta.$$

Now we want to generalize this result for all vectors instead of fixed ones. By union bound on the $\binom{n}{2}$ vectors, we have

$$\mathbf{Pr}\left[\forall i < j, (1-\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\| \le \left\|L\mathbf{x}_i - L\mathbf{x}_j\right\| \le (1+\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right] \ge 1 - \binom{n}{2}\delta.$$

Pick $\delta = 1/\left(n\binom{n}{2}\right) = \Omega(n^{-3})$,

$$d = O\left(1/\varepsilon^2 \log 1/\delta\right) = O\left(1/\varepsilon^2 \log n\right),$$

which gives the Johnson-Lindenstrauss Lemma.

4.2.2. *Proof of the main lemma.* Consider using the row vectors of R, suppose R_i is the i^{th} row of R.

$$(R\mathbf{v})_i = R_i^T \mathbf{v} = \sum_j \mathbf{v}_j R_{ij}.$$

Hence

$$(**) \Leftrightarrow \mathbf{Pr}\left[(1-\varepsilon)^2 \le 1/d \left(\sum_i (R\mathbf{v})_i^2\right) \le (1+\varepsilon)^2\right] \ge 1 - \delta.$$

Since $R_{ij} \sim N(0,1)$, we use the additive properties of Gaussians in the previous section,

$$(R\mathbf{v})_i \sim N\left(0, \sum_j \mathbf{v}_j^2\right) = N(0,1).$$

Now we only need to prove:

$$\mathbf{Pr}\left[\sum_{i=1}^{d} g_i^2 \geq (1+\varepsilon)^2 d\right] \leq \delta/2 \tag{1}$$

and

$$\mathbf{Pr}\left[\sum_{i=1}^{d} g_i^2 \leq (1-\varepsilon)^2 d\right] \leq \delta/2, \tag{2}$$

where $g_i \sim N(0,1)$ are independent and identically distributed unit Gaussians.

(1) follows from $\forall \lambda \geq 0$, by Markov's inequality,

$$\mathbf{Pr}\left[\exp\left(\lambda \sum_{i=1}^{d} g_i^2\right) \geq \exp\left(\lambda d(1+\varepsilon)^2\right)\right] \leq \exp\left(-\lambda d(1+\varepsilon)^2\right) \mathbf{E}\left[\exp\left(\lambda \sum_{i=1}^{d} g_i^2\right)\right],$$

Because g_i are independent and identically distributed,

$$\begin{aligned}
&= \exp\left(-\lambda d(1+\varepsilon)^2\right) \prod_{i=1}^{d} \mathbf{E}\left[\exp\left(\lambda g_i^2\right)\right] \\
&= \exp\left(-\lambda d(1+\varepsilon)^2\right) \left(\mathbf{E}\left[\exp\left(\lambda g^2\right)\right]\right)^d \\
&\leq \exp\left(-\lambda d(1+\varepsilon)^2 - d/2 \log(1-2\lambda)\right)
\end{aligned}$$

where $g \sim N(0,1)$. Since

$$\begin{aligned}
\mathbf{E}\left[\lambda g^2\right] &= \int_{-\infty}^{\infty} e^{\lambda x^2} \cdot \left(1/\sqrt{2\pi}\right) e^{-x^2/2} \mathrm{d}x \\
&= 1/\sqrt{2\pi} \int_{-\infty}^{\infty} \exp\left(-(1-2\lambda)x^2/2\right) \mathrm{d}x = 1/\sqrt{1-2\lambda}
\end{aligned}$$

The easiest way to see the above is that the probability distribution function of $N(0, 1/1-2\lambda)$ is

$$p(x) = \frac{\sqrt{1-2\lambda}}{\sqrt{2\pi}} \exp\left(-(1-2\lambda)x^2/2\right).$$

Now we optimize for λ. Differentiate the above,

$$-d(1+\varepsilon)^2 - \frac{d}{2} \cdot \frac{-2}{(1-2\lambda)} = 0.$$

And we get

$$\lambda = \frac{1}{2}\left(1 - \frac{1}{(1+\varepsilon)^2}\right).$$

Substituting in λ, and given $\log(1+\varepsilon) \leq \varepsilon$:

$$\begin{aligned}
&\mathbf{Pr}\left[\sum_{i=1}^{d} g_i^2 \geq (1+\varepsilon)^2 d\right] \\
&\leq \exp\left(-d\left((1+\varepsilon)^2 - 1\right)/2 - d/2 \log(1+\varepsilon)\right) \\
&\leq \exp\left(-d(\varepsilon + \varepsilon^2/2) + d\varepsilon\right) \\
&= \exp\left(-d\varepsilon^2/2\right)
\end{aligned}$$

Similarly for (2), pick $d = 2/\varepsilon^2 \log\left(2/\delta\right)$, we will get $\leq \frac{\delta}{2}$.

5. Tightness of JL-Lemma

Alon (also Larson-Nelson) stated that if n vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n \in \mathbb{R}^d$, and $1-\varepsilon \leq \left\|\mathbf{f}_i - \mathbf{f}_j\right\| \leq 1+\varepsilon$, then

$$d = \Omega\left(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon}\right).$$

6. Rotational Invariance

Recall the $d \times n$ matrix R where $R_{ij} \sim N(0,1)$, $\forall \mathbf{x} \in \mathbb{R}^n$,

$$p(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{d/2}} \exp\left(-1/2 \left\|\mathbf{x}\right\|^2\right).$$

The probability distribution function only depends on $\|\mathbf{x}\|$, but not the direction. We call this property "rotational invariance."

Formally, if we have orthogonal matrix $O \in \mathbb{R}^{n \times n}$ (orthogonal means $O^T O = I_{n \times n}$) and a Gaussian vector $\mathbf{x} \sim N(0,1)_{n \times 1}$, then $O\mathbf{x}$, rotating the vector $\mathbf{x}$, would also be Gaussian, i.e., $O\mathbf{x} \sim N(0,1)_{n \times 1}$.

A useful application of the rotational invariance of Gaussian vectors is to sample random unit vectors. Sample Gaussian vector $\mathbf{x} \sim N(0,1)_{n \times 1}$, then $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ is a random unit vector desired.

7. Original JL-Lemma

Johnson and Lindenstrauss originally proved the lemma for uniformly random d-dimensional subspace using L such that $L^T L$ is the projection matrix onto the subspace. Dasgupta and Gupta then proved the JL-Lemma using Gaussians. The original claim and proof is the following.

7.1. **Claim.** For a fixed unit vector $\mathbf{v}$, P is a projection onto a uniformly random d-dimensional subspace, then $\|P\mathbf{v}\|^2$ is strongly concentrated. This claim is equivalent to the following statement. For a fixed projection Π onto $\{e_1, e_2, \ldots, e_d\}$ (where e_i are the first d coordinates, $\mathbf{v}$ is a uniformly random unit vector, then $\|\Pi\mathbf{v}\|^2$ is strongly concentrated.

7.2. **Proof.** We sample $\mathbf{x} \sim N(0,1)_{n \times 1}$ (independently) as an n-dimensional Gaussian vector, and let $\mathbf{v} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$. $\Pi\mathbf{v} = (v_1, v_2, \ldots, v_d)$. We want to proof that $\|\Pi\mathbf{v}\|^2 = \sum_{i=1}^{d} v_i^2$ is concentrated. Since by symmetry, $\forall i, \mathbf{E}\left[v_i^2\right] = \frac{1}{n}$, hence $\mathbf{E}\left[\sum_{i=1}^{d} v_i^2\right] = d/n$. Therefore, we want to bound (1) $\Pr\left[\sum_{i=1}^{d} v_i^2 \geq (1+\varepsilon)^2 d/n\right]$ and (2) $\Pr\left[\sum_{i=1}^{d} v_i^2 \leq (1-\varepsilon)^2 d/n\right]$.

$$
\begin{aligned}
(5) &= \mathbf{Pr}\left[\frac{\sum_{i=1}^{d} x_i^2}{\sum_{i=1}^{n} x_i^2} \geq (1+\varepsilon)^2 d/n\right] \\
&= \mathbf{Pr}\left[\frac{1}{d}\sum_{i=1}^{d} x_i^2 - \frac{(1+\varepsilon)^2}{n}\left(\sum_{i=1}^{n} x_i^2\right) \geq 0\right] \\
&= \mathbf{Pr}\left[\left(\frac{1}{d} - \frac{(1+\varepsilon)^2}{n}\right)\sum_{i=1}^{d} x_i^2 - \frac{(1+\varepsilon)^2}{n}\left(\sum_{i=d+1}^{n} x_i^2\right) \geq 0\right]
\end{aligned}
$$

Now we use the independence of x_i and Chernoff bound, which produced the same result:

$$
(5) \leq \exp\left(-\frac{d}{2}\left((1+\varepsilon)^2 - 1\right) + \frac{d}{2}\log(1+\varepsilon)^2\right)
$$

HW 2. *Consider the following two ways of sampling a d-dimensional subspace:*

(1) Sample a uniformly random unit vector u_1. For $2 \leq i \leq d$, sample a uniformly random unit vector u_i that is orthogonal to $\{u_1, u_2, \ldots, u_{i-1}\}$. Output span$(u_1, u_2, \ldots, u_d)$.

(2) Sample $g_i \sim N(0,1)_{n\times 1}$, output span$(g_1, \ldots, g_d)$.

Prove that up to zero probability events, both generate a uniformly random d-dimensional subspace.

References