Lecture 06: Applications I

Nikola Zlatanov

Innopolis University

Advanced Statistics

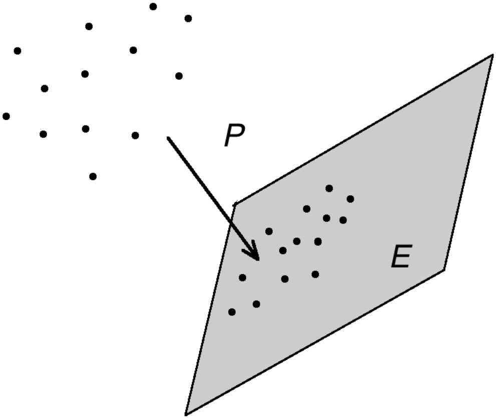
13-th of March to 20-th of March, 2023

Dimension Reduction

- We started this course by observing that high dimensions could be a major problem, due to the curse of dimensionality.
- Why don't we then fix the problem due to high-dimensions as follows: Take the high-dimensional data and transform it into a low-dimensional data.
- Let there be N high-dimensional vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$, where $\boldsymbol{x}_j^T = [x_1, x_2, ..., x_d]$, for $j = 1, 2, ..., N$, where d is very high.
- We would like to make a transformation of $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ into $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N$, where $\boldsymbol{y}_j^T = [y_1, y_2, ..., y_n]$, for $j = 1, 2, ..., N$, where $n \ll d$, and yet the geometry of $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ are preserved in $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N$.

Dimension Reduction

- First we have to define what do we mean by “the geometry of the data”: By “the geometry of the data”, we mean the pairwise distances between the original data.



Is this possible? It turns out it is possible if $n = O(\ln(N)) \ll d$.

Dimension Reduction

- Thm (Johnson-Lindenstrauss Lemma): $\forall$ vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$, where $\boldsymbol{x}_j \in \mathbb{R}^d$, for $j = 1, 2, ..., N$, there exists a linear map $T : \boldsymbol{x}_j \to \boldsymbol{y}_j$, where $\boldsymbol{y}_j \in \mathbb{R}^n$ and $n \ll d$ such that the following holds

$$\Pr\left\{(1-\delta)||\boldsymbol{x}_m - \boldsymbol{x}_j||_2 \le ||\boldsymbol{y}_m - \boldsymbol{y}_j||_2 \le (1+\delta)||\boldsymbol{x}_m - \boldsymbol{x}_j||_2\right\} \ge 1 - \epsilon \tag{1}$$

for any $j \neq m$, where $j = 1, 2, ..., N$ and $m = 1, 2, ..., N$, and small $\delta > 0$ and $\epsilon > 0$ if

$$n > \frac{1}{c}\left(\ln(N) + \frac{1}{2}\ln\left(\frac{1}{\epsilon}\right) + \frac{1}{2}\ln(2)\right)$$

holds where

$$c = \frac{\delta(2-\delta)}{k}\min\left\{\frac{\delta(2-\delta)}{k}, 1\right\}, \tag{2}$$

where k is some constant.

Dimension Reduction

Proof:

- The vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ and $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N$ are all deterministic.
- However, we will use a probabilistic method to find the mapping $T : \boldsymbol{x}_j \to \boldsymbol{y}_j$.
- Specifically, we will choose a linear map $T(\cdot)$ at random, and then we will prove that the linear map $T(\cdot)$ satisfies the properties that we seek.
- Now, a linear map is simply a multiplication of $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ by a matrix $\boldsymbol{T}$, of size $n \times d$, to obtain $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N$.
- Hence, if we will choose a linear map $T(\cdot)$ at random, this means that we should choose a matrix $\boldsymbol{T}$ at random.
- How is it possible that a randomly chosen matrix $\boldsymbol{T}$ would work?

Dimension Reduction

- Let's have a random matrix $\boldsymbol{G}$ of size $n \times d$ populated by i.i.d. Gaussian entries, i.e., the (i,j)-th element of $\boldsymbol{G}$, denoted by G_{ij}, is generated i.i.d. according to the Gaussian distribution $N(0,1)$ for $i = 1, 2, ..., n$ and $j = 1, 2, ..., d$.
- Let us have a fixed vector $\boldsymbol{z} \in \mathbb{R}^d$, with norm $||\boldsymbol{z}||_2$.
- Now, let's investigate the distribution of $\boldsymbol{Gz}$. The i-th element of $\boldsymbol{Gz}$, denoted by $(\boldsymbol{Gz})_i$ is given by

$$(\boldsymbol{Gz})_i = \sum_{j=1}^{d} G_{ij} z_j \sim N\left(0, \sum_{j=1}^{d} z_j^2\right) = N\left(0, ||\boldsymbol{z}||_2^2\right) \overset{(a)}{=} N(0,1) \quad (3)$$

where (a) holds if and only if $||\boldsymbol{z}||_2 = 1$.

Dimension Reduction

On the other hand, let's fix two vectors, $\boldsymbol{x}_m$ and $\boldsymbol{x}_l$. Then, the vector $\boldsymbol{G}\boldsymbol{x}_m - \boldsymbol{G}\boldsymbol{x}_l$, according to (3), has i.i.d. zero-mean Gaussian elements each with variance $||\boldsymbol{x}_m - \boldsymbol{x}_l||_2$.

Hence, if we normalize $\boldsymbol{G}\boldsymbol{x}_m - \boldsymbol{G}\boldsymbol{x}_l$ by $||\boldsymbol{x}_m - \boldsymbol{x}_l||_2$, the vector

$$\frac{||\boldsymbol{G}\boldsymbol{x}_m - \boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m - \boldsymbol{x}_l||_2}$$

will have i.i.d. zero-mean Gaussian elements each with variance one. As a result, we can use the Thin-Shell Theorem, which states that

$$\Pr\left\{\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m - \boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m - \boldsymbol{x}_l||_2} - \sqrt{n}\right| \leq \delta\sqrt{n}\right\} \\ \geq 1 - 2\exp\left(-n\frac{\delta(2-\delta)}{k}\min\left\{\frac{\delta(2-\delta)}{k}, 1\right\}\right), \tag{4}$$

where k is some constant.

We will now prove the main theorem as follows. We will select a pair of vectors $\boldsymbol{x}_m$ and $\boldsymbol{x}_l$. We will prove (1) for the selected $\boldsymbol{x}_m$ and $\boldsymbol{x}_l$. Then, we will use the union bound to prove that the theorem holds for all pairs satisfying the given condition.
We start with the selected pair $\boldsymbol{x}_m$ and $\boldsymbol{x}_l$:

$$
\begin{aligned}
&\Pr\left\{\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2}-\sqrt{n}\right|\leq\delta\sqrt{n}\right\}\\
&=\Pr\left\{\big|||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2-||\boldsymbol{x}_m-\boldsymbol{x}_l||_2\sqrt{n}\big|\leq||\boldsymbol{x}_m-\boldsymbol{x}_l||_2\delta\sqrt{n}\right\}\\
&=\Pr\left\{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2\sqrt{n}(1-\delta)\leq||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2\leq||\boldsymbol{x}_m-\boldsymbol{x}_l||_2\sqrt{n}(1+\delta)\right\}\\
&=\Pr\left\{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2\sqrt{n}(1-\delta)\leq||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2\leq||\boldsymbol{x}_m-\boldsymbol{x}_l||_2\sqrt{n}(1+\delta)\right\}\\
&=\Pr\left\{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2(1-\delta)\leq\frac{1}{\sqrt{n}}||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2\leq||\boldsymbol{x}_m-\boldsymbol{x}_l||_2(1+\delta)\right\}\\
&=\text{continuation on next page} \qquad (5)
\end{aligned}
$$

Continuation of (5)

$$\overset{(a)}{=} \Pr\left\{||\boldsymbol{x}_m - \boldsymbol{x}_l||_2(1-\delta) \le ||\boldsymbol{T}\boldsymbol{x}_m - \boldsymbol{T}\boldsymbol{x}_l||_2 \le ||\boldsymbol{x}_m - \boldsymbol{x}_l||_2(1+\delta)\right\}$$
$$\overset{(b)}{\ge} 1 - 2\exp\left(-n\frac{\delta(2-\delta)}{k}\min\left\{\frac{\delta(2-\delta)}{k}, 1\right\}\right) \overset{(c)}{=} 1 - 2e^{-nc} \tag{6}$$

where (a) comes by setting

$$\boldsymbol{T} = \frac{1}{\sqrt{n}}\boldsymbol{G},$$

and (c) comes by setting

$$c = \frac{\delta(2-\delta)}{k}\min\left\{\frac{\delta(2-\delta)}{k}, 1\right\} \tag{7}$$

We now take the union bound over all pairs of N vectors

$$
\begin{aligned}
&\Pr\left\{\bigcap_{m=1}^{N}\bigcap_{l=m+1}^{N}\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2}-\sqrt{n}\right|\leq\delta\sqrt{n}\right\}\\
&=1-\Pr\left\{\bigcup_{k=1}^{N}\bigcup_{l=k+1}^{N}\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2}-\sqrt{n}\right|\geq\delta\sqrt{n}\right\}\\
&\geq 1-\sum_{k=1}^{N}\sum_{l=1}^{N}\Pr\left\{\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2}-\sqrt{n}\right|\geq\delta\sqrt{n}\right\}\\
&=1-N^2\Pr\left\{\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m-\boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m-\boldsymbol{x}_l||_2}-\sqrt{n}\right|\geq\delta\sqrt{n}\right\}\\
&\geq 1-2N^2e^{-nc}\\
&=1-e^{\ln(2)+2\ln(N)-nc}
\end{aligned}
$$

Dimension Reduction

Now, we want to set n, N, and c such that

$$\Pr\left\{\left|\frac{||\boldsymbol{G}\boldsymbol{x}_m - \boldsymbol{G}\boldsymbol{x}_l||_2}{||\boldsymbol{x}_m - \boldsymbol{x}_l||_2} - \sqrt{n}\right| \leq \delta\sqrt{n}\right\} \geq 1 - e^{\ln(2)+2\ln(N)-nc} \geq 1 - \epsilon$$

which occurs if

$$1 - e^{\ln(2)+2\ln(N)-nc} \geq 1 - \epsilon$$

or equivalently if

$$\ln(2) + 2\ln(N) - nc < \ln(\epsilon)$$

or equivalently if

$$\ln(N) < nc - \frac{1}{2}\ln\left(\frac{1}{\epsilon}\right) - \frac{1}{2}\ln(2),$$

where c is given in (7) as function of δ. Q.E.D.

Note! We have lost d. Where is d?