11 Apr 2022

Random Projections for Dimensionality Reduction

Announcements

1. Website updated with
   - lecture videos from the week before Spr Bk.
   - typeset notes from that week
2. Homework 5 updated with solution strategy for Problem (3c).

(Reminder: due Apr. 15)

Dimensionality reduction

- Data set represented by n vectors in $\mathbb{R}^d$.

  Assume n, d both large, e.g. 10^9.

  Assume $\|x-y\|_2$ is a meaningful measure of similarity of data points x, y.

  E.g. $\mathbb{R}^d$ in this example might be the output layer of a neural net that takes raw data and maps to a representation where L_2 dist. is a meaningful similarity measure.

Task. Try to find a set of points

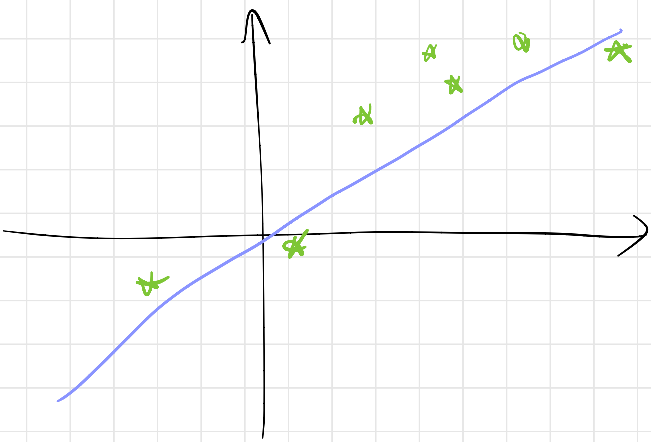
$$x_1', x_2', \ldots, x_n' \in \mathbb{R}^k$$

where $k \ll n$, such that

$$\forall i,j \quad (1-\varepsilon)\|x_i - x_j\|_2 \leq \|x_i' - x_j'\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2$$

Plan. Use a random linear transformation $\mathbb{R}^d \to \mathbb{R}^k$ represented by a matrix R with independent Gaussian entries having distribution $N(0, \frac{1}{k})$.

It'll turn out that $k > 4\ln(n/\delta)/\varepsilon^2$ suffices, but we'll just keep calling dimension k for now.



To analyze: R random matrix $k \times d$

with entries $r_{ij} \sim N(0, \frac{1}{k})$

$y = x_i - x_j \in \mathbb{R}^d$

Want to know if $\|Ry\|_2$ is likely close to $\|y\|_2$?

$\boxed{\text{Step 1}}$ Let $Y = \|Ry\|_2^2$. What is $\mathbb{E}Y$?

$$R = \begin{bmatrix} — & r_1^T & — \\ — & r_2^T & — \\ & \vdots & \\ — & r_k^T & — \end{bmatrix} \qquad Ry = \begin{pmatrix} \langle r_1, y\rangle \\ \langle r_2, y\rangle \\ \vdots \\ \langle r_k, y\rangle \end{pmatrix}$$

$$Y = \|Ry\|_2^2 = \sum_{i=1}^{k} \langle r_i, y\rangle^2 = \sum_{i=1}^{k} X_i^2 \quad (X_i = \langle r_i, y\rangle).$$

$X_1, \dots, X_k$ independent, identically distrib.

$$X_i \sim N\left(0, \frac{1}{k}\|y\|_2^2\right).$$

$$\mathbb{E}[X_i^2] = \frac{1}{k}\|y\|_2^2$$

$$\mathbb{E}[Y] = \sum_{i=1}^{k} \mathbb{E}[X_i^2] = k \cdot \frac{1}{k}\|y\|_2^2 = \|y\|_2^2.$$

Lemma 4.4 (Notes on Probability) tells us

$\forall \; 0 < \varepsilon < 1$

$$\Pr\left(Y < (1-\varepsilon)^2 \mathbb{E}Y\right) < e^{-\frac{k}{2}\varepsilon^2}$$

$$\Pr\left(Y > (1+\varepsilon)^2 \mathbb{E}Y\right) < e^{-\frac{k}{2}\varepsilon^2}$$

We want $\|R(x_i - x_j)\|_2$ to be close to $\|x_i - x_j\|_2$ for $\binom{n}{2}$ distinct pairs i,j.

The analysis above with $Y = \|R(x_i - x_j)\|_2^2$ says that the approximation for <u>one</u> <u>single</u> <u>pair</u> i,j can fail with probability $< 2e^{-\frac{1}{2}k\varepsilon^2}$.

<u>Union Bound</u>: Probability there is <u>any</u> pair i,j with $\|R(x_i - x_j)\|_2$ lying outside interval

$$\left[(1-\varepsilon)\|x_i - x_j\|_2 \,,\; (1+\varepsilon)\|x_i - x_j\|_2\right]$$

is less than $2\binom{n}{2}e^{-\frac{1}{2}k\varepsilon^2}$

Now make this less than δ...

Want

$$2\binom{n}{2} e^{-\frac{1}{2}k\varepsilon^2} < \delta$$

$$e^{-\frac{1}{2}k\varepsilon^2} < \frac{\delta}{n(n-1)}$$

$$e^{\frac{1}{2}k\varepsilon^2} > \frac{n(n-1)}{\delta}$$

$$\frac{1}{2}k\varepsilon^2 > \ln\left(\frac{n(n-1)}{\delta}\right)$$

$$k > 2\ln\left(\frac{n(n-1)}{\delta}\right)/\varepsilon^2$$

This inequality will hold when $k > 4\ln\left(\frac{n}{\delta}\right)/\varepsilon^2$.

To see that $\ln(n)/\varepsilon^2$ dimensions are really needed, suppose $x_1, \ldots, x_n$ are the standard basis vectors in $\mathbb{R}^n$. (So $d = n$.)