Lecture 21: Johnson-Lindenstrauss lemma, random projections, CNN

Nisha Chandramoorthy

November 7, 2023

Last time: PCA interpretation

- Exact when $X^\top X$ is rank n.

Last time: PCA interpretation

- Exact when $X^\top X$ is rank n.
- Maximizes variance. Let x be a random vector chosen uniformly from centered data $x_1, \cdots, x_m$. Then, for any $w \in \mathbb{R}^d$,

$$\mathrm{var}(x \cdot w) = \frac{1}{m} \sum_{i=1}^{m} (w \cdot x_i)^2.$$

Last time: PCA interpretation

- Exact when $X^\top X$ is rank n.
- Maximizes variance. Let x be a random vector chosen uniformly from centered data $x_1, \cdots, x_m$. Then, for any $w \in \mathbb{R}^d$,
$$\mathrm{var}(x \cdot w) = \frac{1}{m} \sum_{i=1}^{m} (w \cdot x_i)^2.$$
- First principal component maximizes $\mathrm{var}(x \cdot w)$ over all w with $\|w\| = 1$. (See Ex 23.4 in book.)

Last time: PCA interpretation

- Exact when $X^\top X$ is rank n.
- Maximizes variance. Let x be a random vector chosen uniformly from centered data $x_1, \cdots, x_m$. Then, for any $w \in \mathbb{R}^d$,

$$\mathrm{var}(x \cdot w) = \frac{1}{m} \sum_{i=1}^{m} (w \cdot x_i)^2.$$

- First principal component maximizes $\mathrm{var}(x \cdot w)$ over all w with $\|w\| = 1$. (See Ex 23.4 in book.)
- Informally, PCA rotates the data so that the variance is maximized along the first axis, then the second, and so on.

Last time: PCA interpretation

- Exact when $X^\top X$ is rank n.
- Maximizes variance. Let x be a random vector chosen uniformly from centered data $x_1, \cdots, x_m$. Then, for any $w \in \mathbb{R}^d$,
$$\mathrm{var}(x \cdot w) = \frac{1}{m} \sum_{i=1}^{m} (w \cdot x_i)^2.$$
- First principal component maximizes $\mathrm{var}(x \cdot w)$ over all w with $\|w\| = 1$. (See Ex 23.4 in book.)
- Informally, PCA rotates the data so that the variance is maximized along the first axis, then the second, and so on.
- Separates dissimilar points

Kernel PCA

- Let V be the matrix of the top n eigenvectors of $K = XX^\top \in \mathbb{R}^{m\times m}$.

Kernel PCA

- Let V be the matrix of the top n eigenvectors of $K = XX^\top \in \mathbb{R}^{m \times m}$.
- Then, principal vectors are $d_i^* = \frac{1}{\|X^\top v_i\|} X^\top v_i$, $i = 1, 2, \cdots, n$.

Kernel PCA

- Let V be the matrix of the top n eigenvectors of $K = XX^\top \in \mathbb{R}^{m \times m}$.
- Then, principal vectors are $d_i^* = \frac{1}{\|X^\top v_i\|} X^\top v_i$, $i = 1, 2, \cdots, n$.
- For some PD kernel, if $K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = (XX^\top)_{ij}$, can compute K only using kernel evaluations.

Graphical reprensentation of X

- Choose weighting, such as, $w_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$. As $\sigma \to 0$, $w_{ij} \to \mathbb{1}_{i=j}$. The $m \times m$ matrix W is the adjacency matrix of a graph.

Graphical reprensentation of X

- Choose weighting, such as, $w_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$. As $\sigma \to 0$, $w_{ij} \to \mathbb{1}_{i=j}$. The $m \times m$ matrix W is the adjacency matrix of a graph.
- Let D be the diagonal matrix with $D_{ii} = \sum_{j=1}^{m} w_{ij}$.

Graphical reprensentation of X

- Choose weighting, such as, $w_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$. As $\sigma \to 0$, $w_{ij} \to \mathbb{1}_{i=j}$. The $m \times m$ matrix W is the adjacency matrix of a graph.
- Let D be the diagonal matrix with $D_{ii} = \sum_{j=1}^{m} w_{ij}$.
- Graph laplacian: $L = D - W$.

Graphical reprensentation of X

- Choose weighting, such as, $w_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$. As $\sigma \to 0$, $w_{ij} \to \mathbb{1}_{i=j}$. The $m \times m$ matrix W is the adjacency matrix of a graph.
- Let D be the diagonal matrix with $D_{ii} = \sum_{j=1}^{m} w_{ij}$.
- Graph laplacian: $L = D - W$.
- Detects local structure / clusters in data.

Laplacian eigenmaps

- Want to solve: $\min_{y_1,\cdots,y_m} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \|y_i - y_j\|^2$.

Laplacian eigenmaps

- Want to solve: $\min_{y_1, \cdots, y_m} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \|y_i - y_j\|^2$.
- optimal embeddings: $y_i = E(x_i) = U[i, -n:]$ where U is the matrix of eigenvectors of L.

Laplacian eigenmaps

- Want to solve: $\min_{y_1,\cdots,y_m} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \|y_i - y_j\|^2$.
- optimal embeddings: $y_i = E(x_i) = U[i, -n:]$ where U is the matrix of eigenvectors of L.
- For any vector v, $v^\top L v = (1/2) \sum_{i,j=1}^{m} w_{ij}(v_i - v_j)^2$.

Laplacian eigenmaps

- Want to solve: $\min_{y_1,\cdots,y_m} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \|y_i - y_j\|^2$.
- optimal embeddings: $y_i = E(x_i) = U[i, -n:]$ where U is the matrix of eigenvectors of L.
- For any vector v, $v^\top L v = (1/2) \sum_{i,j=1}^{m} w_{ij}(v_i - v_j)^2$.
- L is positive semi-definite.

Bottom n eigenvectors

- Rayleigh quotient optimality

Bottom n eigenvectors

- Rayleigh quotient optimality
- Another interpretation: top n eigenvectors of $L^\dagger$. $L^\dagger_{ij}$ represents expected time for random walk $i \to j \to i$.

Bottom n eigenvectors

- Rayleigh quotient optimality
- Another interpretation: top n eigenvectors of $L^\dagger$. $L^\dagger_{ij}$ represents expected time for random walk $i \to j \to i$.
- Kernel PCA with $K = L^\dagger$ is equivalent to Laplacian eigenmaps.

Stochastic neighbor embedding [Hinton and Roweis 2002]

- Stochastic neighbor embedding(SNE): conditional probability that x_i would pick x_j as its neighbor, given by

$$p(x_j|x_i) = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}.$$

Stochastic neighbor embedding [Hinton and Roweis 2002]

- Stochastic neighbor embedding(SNE): conditional probability that x_i would pick x_j as its neighbor, given by

$$p(x_j|x_i) = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}.$$

- For the embeddings $y_i = E(x_i)$,

$$q(y_j|y_i) = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)}.$$

Stochastic neighbor embedding [Hinton and Roweis 2002]

- Stochastic neighbor embedding(SNE): conditional probability that x_i would pick x_j as its neighbor, given by

$$p(x_j|x_i) = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}.$$

- For the embeddings $y_i = E(x_i)$,

$$q(y_j|y_i) = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)}.$$

- SNE minimizes $\sum_{i=1}^{m} D_{\mathrm{KL}}(p_i\|q_i)$, where p_i and q_i are the conditional probabilities of x_i and y_i respectively.

Stochastic neighbor embedding [Hinton and Roweis 2002]

- Stochastic neighbor embedding(SNE): conditional probability that x_i would pick x_j as its neighbor, given by

$$p(x_j|x_i) = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}.$$

- For the embeddings $y_i = E(x_i)$,

$$q(y_j|y_i) = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)}.$$

- SNE minimizes $\sum_{i=1}^{m} D_{\mathrm{KL}}(p_i\|q_i)$, where p_i and q_i are the conditional probabilities of x_i and y_i respectively.
- Penalizes large distances between x_i and x_j but also preserves local structure.

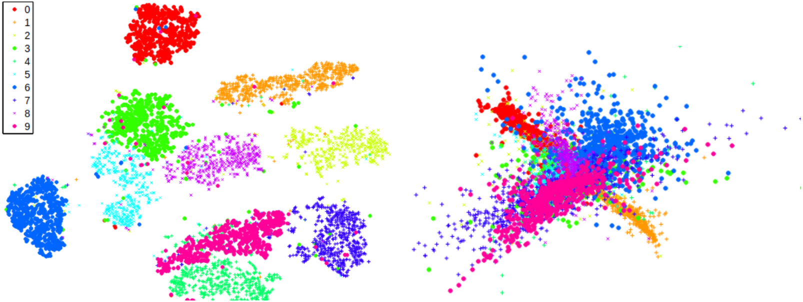
tSNE [Van der Maaten and Hinton 2008]

- tSNE cost function is $D_{\mathrm{KL}}(p\|q) = \sum_{i=1}^{m} \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $p_i j$ and q_{ij} are the joint probabilities of (x_i, x_j) and (y_i, y_j) respectively.
- Changes joint distribution to a heavy-tailed distribution, $q(y_j, y_i) = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1+\|y_i - y_k\|^2)^{-1}}$.

tSNE [Van der Maaten and Hinton 2008]

- tSNE cost function is $D_{\mathrm{KL}}(p \| q) = \sum_{i=1}^{m} \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $p_i j$ and q_{ij} are the joint probabilities of (x_i, x_j) and (y_i, y_j) respectively.
- Changes joint distribution to a heavy-tailed distribution, $q(y_j, y_i) = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1+\|y_i - y_k\|^2)^{-1}}$.
- approaches inverse square law on embedded space.

tSNE visualization



From Van der Maaten and Hinton 2008. tSNE (left) and LLE (right) on MNIST dataset.

Johnson-Lindenstrauss lemma

- Let $X \in \mathbb{R}^{m \times d}$ be a matrix of m points in $\mathbb{R}^d$.

Johnson-Lindenstrauss lemma

- Let $X \in \mathbb{R}^{m \times d}$ be a matrix of m points in $\mathbb{R}^d$.
- Let $0 < \epsilon < 1/2$, $m > 4$. Then, there exists a linear map $A : \mathbb{R}^d \to \mathbb{R}^n$ with $n = O(\epsilon^{-2} \log m)$ such that for all $x_i, x_j \in X$, $i, j \in [m]$, $(1 - \epsilon)\|x_i - x_j\|^2 \leqslant \|Ax_i - Ax_j\|^2 \leqslant (1 + \epsilon)\|x_i - x_j\|^2$.

Johnson-Lindenstrauss lemma

- Let $X \in \mathbb{R}^{m \times d}$ be a matrix of m points in $\mathbb{R}^d$.
- Let $0 < \epsilon < 1/2$, $m > 4$. Then, there exists a linear map $A : \mathbb{R}^d \to \mathbb{R}^n$ with $n = O(\epsilon^{-2} \log m)$ such that for all $x_i, x_j \in X$, $i, j \in [m]$, $(1 - \epsilon)\|x_i - x_j\|^2 \leqslant \|Ax_i - Ax_j\|^2 \leqslant (1 + \epsilon)\|x_i - x_j\|^2$.
- Informal: any set of points in high-dimensional space can be mapped to a lower-dimensional space while approximately preserving the distances between the points.

Proof

- Distortion by Gaussian random matrices: for any $x \in \mathbb{R}^d$, when the entries A_{ij} are iid standard Gaussian,

$$\mathbb{P}(n(1-\epsilon)\|x\|^2 \leqslant \|Ax\|^2 \leqslant n(1+\epsilon)\|x\|^2) \\ \geqslant 1 - 2\exp(-(\epsilon^2 - \epsilon^3)n/4).$$

Proof

- Distortion by Gaussian random matrices: for any $x \in \mathbb{R}^d$, when the entries A_{ij} are iid standard Gaussian,

$$\mathbb{P}(n(1-\epsilon)\|x\|^2 \leqslant \|Ax\|^2 \leqslant n(1+\epsilon)\|x\|^2) \geqslant 1 - 2\exp(-(\epsilon^2 - \epsilon^3)n/4).$$

- Then, deterministic statement of J-L lemma follows from union bound over all m^2 pairs of points.

To show: distortion by Gaussian random matrices

- Let A be a $n \times d$ matrix with iid standard Gaussian entries. Then, $E[(Ax)_j] = 0$ and $\mathrm{Var}((Ax)_j) = \|x\|^2$, for all $j \leqslant n$.

To show: distortion by Gaussian random matrices

- Let A be a $n \times d$ matrix with iid standard Gaussian entries. Then, $E[(Ax)_j] = 0$ and $\mathrm{Var}((Ax)_j) = \|x\|^2$, for all $j \leqslant n$.
- Thus, $1/\|x\|^2\|Ax\|^2$ is a χ^2 random variable with n degrees of freedom.

To show: distortion by Gaussian random matrices

- Let A be a $n \times d$ matrix with iid standard Gaussian entries. Then, $E[(Ax)_j] = 0$ and $\mathrm{Var}((Ax)_j) = \|x\|^2$, for all $j \leqslant n$.
- Thus, $1/\|x\|^2 \|Ax\|^2$ is a χ^2 random variable with n degrees of freedom.
- Chi-squared distribution: $\rho(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, x \geqslant 0$.

To show: distortion by Gaussian random matrices

- Let A be a $n \times d$ matrix with iid standard Gaussian entries. Then, $E[(Ax)_j] = 0$ and $\mathrm{Var}((Ax)_j) = \|x\|^2$, for all $j \leqslant n$.
- Thus, $1/\|x\|^2 \|Ax\|^2$ is a χ^2 random variable with n degrees of freedom.
- Chi-squared distribution: $\rho(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, x \geqslant 0$.
- Models sum of squares of n independent standard normal random variables.

Chi-squared distribution

- Moment generating function,
  $M(t) = E[e^{tX}] = (1 - 2t)^{-n/2}$, $t < 1/2$.

Chi-squared distribution

- Moment generating function, $M(t) = E[e^{tX}] = (1-2t)^{-n/2}$, $t < 1/2$.
- Lemma 15.2 (Mohri et al): for a χ^2 random variable X with n degrees of freedom,

$$\mathbb{P}(n(1-\epsilon) \leqslant X \leqslant n(1+\epsilon)) \geqslant 1 - 2\exp(-(\epsilon^2 - \epsilon^3)n/4). \tag{1}$$

Chi-squared distribution

- Moment generating function, $M(t) = E[e^{tX}] = (1-2t)^{-n/2}, t < 1/2$.
- Lemma 15.2 (Mohri et al): for a χ^2 random variable X with n degrees of freedom,

$$\mathbb{P}(n(1-\epsilon) \leqslant X \leqslant n(1+\epsilon)) \geqslant 1 - 2\exp(-(\epsilon^2 - \epsilon^3)n/4). \tag{1}$$

- Use Markov inequality and moment generating function to prove.

Chi-squared distribution

- Moment generating function,
  $M(t) = E[e^{tX}] = (1-2t)^{-n/2}$, $t < 1/2$.
- Lemma 15.2 (Mohri et al): for a χ^2 random variable X with n degrees of freedom,

$$\mathbb{P}(n(1-\epsilon) \leqslant X \leqslant n(1+\epsilon)) \geqslant 1 - 2\exp(-(\epsilon^2 - \epsilon^3)n/4). \tag{1}$$

- Use Markov inequality and moment generating function to prove.
- Use Lemma 15.2 to prove distortion by Gaussian random matrices.

Implications: random projections

- Random projections surprisingly preserve Euclidean distances between points.

Implications: random projections

- Random projections surprisingly preserve Euclidean distances between points.
- Can be used for dimensionality reduction.

Implications: random projections

- Random projections surprisingly preserve Euclidean distances between points.
- Can be used for dimensionality reduction.
- Can also be used for speeding up nearest neighbor search (e.g. within Laplacian eigenmaps).

Implications: random projections

- Random projections surprisingly preserve Euclidean distances between points.
- Can be used for dimensionality reduction.
- Can also be used for speeding up nearest neighbor search (e.g. within Laplacian eigenmaps).

Compressed sensing revisited

- Let $A \in \mathbb{R}^{n \times d}$ be a random matrix with iid standard Gaussian entries. This is an example of a matrix that satisfies the RIP (restricted isometry property).

Compressed sensing revisited

- Let $A \in \mathbb{R}^{n \times d}$ be a random matrix with iid standard Gaussian entries. This is an example of a matrix that satisfies the RIP (restricted isometry property).
- s-RIP: for all subsets $S \subset [d]$ with $|S| \leqslant s$, there exists an $\epsilon_s > 0$ such that

$$(1 - \epsilon)\|x\|^2 \leqslant \|Ax\|^2 \leqslant (1 + \epsilon)\|x\|^2. \tag{2}$$

Compressed sensing revisited

- Let $A \in \mathbb{R}^{n\times d}$ be a random matrix with iid standard Gaussian entries. This is an example of a matrix that satisfies the RIP (restricted isometry property).
- s-RIP: for all subsets $S \subset [d]$ with $|S| \leqslant s$, there exists an $\epsilon_s > 0$ such that

$$(1-\epsilon)\|x\|^2 \leqslant \|Ax\|^2 \leqslant (1+\epsilon)\|x\|^2. \tag{2}$$

Tao 2005 If x is s-sparse, then,

$$x = \mathrm{argmin}_{z\in\mathbb{R}^d}\|z\|_1 \quad \text{s.t.} \quad Ax = Az. \tag{3}$$

Exact recovery of sparse data

- Informal: if x is s-sparse, then it can be recovered exactly from its compressed form Ax.

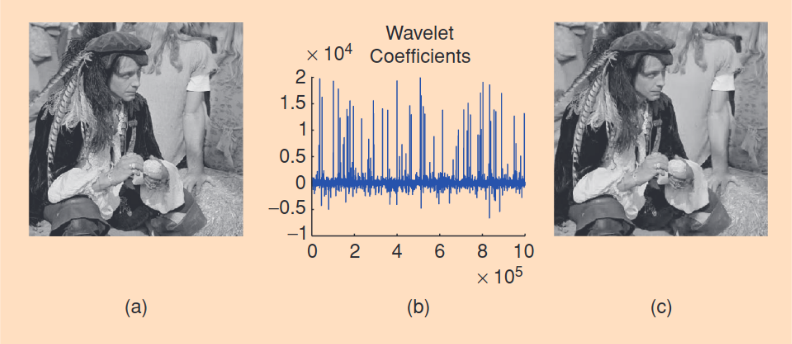
Exact recovery of sparse data

- Informal: if x is s-sparse, then it can be recovered exactly from its compressed form Ax.
- Very useful in signal processing, medical imaging, etc.

Exact recovery of sparse data

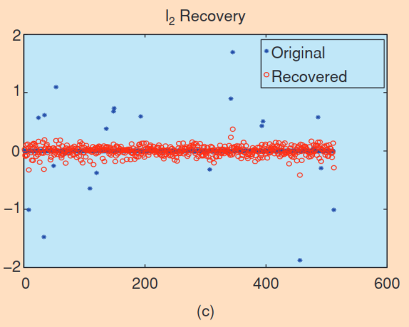
- Informal: if x is s-sparse, then it can be recovered exactly from its compressed form Ax.
- Very useful in signal processing, medical imaging, etc.
- Reconstruction obtained by solving a convex program.

Candes 2008

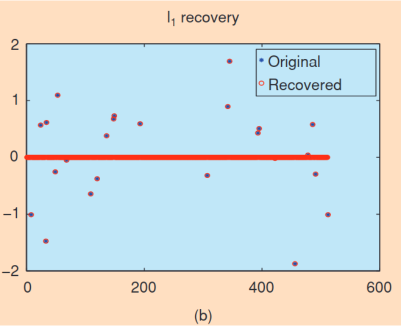


[FIG1] (a) Original megapixel image with pixel values in the range [0,255] and (b) its wavelet transform coefficients (arranged in random order for enhanced visibility). Relatively few wavelet coefficients capture most of the signal energy; many such images are highly compressible. (c) The reconstruction obtained by zeroing out all the coefficients in the wavelet expansion but the 25,000 largest (pixel values are thresholded to the range [0,255]). The difference with the original picture is hardly noticeable. As we describe in "Undersampling and Sparse Signal Recovery," this image can be perfectly recovered from just 96,000 incoherent measurements.

Candes 2008



l_2 Recovery
Original
Recovered
2
1
0
−1
−2
0
200
400
600
(c)



l_1 recovery
Original
Recovered
2
1
0
−1
−2
0
200
400
600
(b)

Convolutional Neural Networks (source: cs231n.stanford.edu)

- Suitable for image recognition. Won the 2012 ImageNet competition and subsequent ones.
- Three types of layers: convolutional, FC, pooling
- Convolutional layer: accepts a volume of size $W_1 \times H_1 \times D_1$ and outputs a volume of size $W_2 \times H_2 \times D_2$ where $W_2 = (W_1 - F + 2P)/S + 1$ and $H_2 = (H_1 - F + 2P)/S + 1$ and $D_2 = K$.
- K is number of filters, F is filter size, S is stride, P is padding.
- Pooling layer: downsamples along width and height, and optionally along depth.
- FC layer: computes class scores, resulting in volume of size $1 \times 1 \times K$.