High dimensional geometry and Regularization

Narayana Santhanam

EE 645
Mar 3, 2023

This week

High dimensional Gaussians

Johnson Lindenstrauss Lemma

University of Hawai'i Mānoa

This week

High dimensional Gaussians

- Johnson Lindenstrauss Lemma

Ridge and Lasso

- Explanations
- Compressive sensing
- Matrix norms

High dimensional Gaussian

Multivariate Gaussian

$$f(\mathrm{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-(\mathrm{x}-\mu)^T \Sigma^{-1} (\mathrm{x}-\mu) \right)$$

$\mu = \mathbb{E}X$ (mean)
$\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$ (covariance)

High dimensional Gaussian

Multivariate Gaussian

$$f(\mathrm{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-(\mathrm{x}-\mu)^T \Sigma^{-1} (\mathrm{x}-\mu)\right)$$

$\mu = \mathbb{E}X$ (mean)
$\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$ (covariance)

Where is the probability concentrated?

University of Hawai'i Mānoa

Concentration of measure

$U \sim N(\mu, \sigma^2 I)$

University of Hawai'i Mānoa

Concentration of measure

$U \sim N(\mu, \sigma^2 I)$

Thin shell with width $\sqrt{d}$

For all $\delta > 0$,

$$\mathrm{P}\left(||U - \mu||^2 \leq \sigma^2 \left(d + 2\sqrt{d \ln \frac{1}{\delta}} \right) \right) \geq 1 - \delta$$

and

$$\mathrm{P}\left(||U - \mu||^2 \geq \sigma^2 \left(d - 2\sqrt{d \ln \frac{1}{\delta}} \right) \right) \geq 1 - \delta$$

UNIVERSITY of HAWAI'I® MĀNOA

Concentration of measure

$U \sim N(\mu, \sigma^2 I)$

Concentration of measure

$U \sim N(\mu, \sigma^2 I)$

Around equator relative to any unit vector z

For all $\delta > 0$,

$$\mathrm{P}\left(z^T(U-\mu) \leq \sigma\sqrt{2\ln\frac{1}{\delta}}\right) \geq 1-\delta$$

University of Hawai'i Mānoa

Johnson Lindenstrauss Lemma

Random projections preserve pairwise distances

For any ϵ and integer n, let $k = \frac{8 \ln n}{\epsilon^2}$. For all $z_1, \ldots, z_n \in \mathbb{R}^d$, there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all pairs z_i, z_j

$$||f(z_i) - f(z_j)||^2 \in (1 \pm \epsilon)||z_i - z_j||^2$$

These f can simply be random projections!

University of Hawai'i Mānoa

Applications of JL lemma

Regression in high dimensions

Some clustering problems
    not always: GMM faster

Sketching and streaming algorithms

University of Hawai'i Mānoa

Learning mixtures of Gaussians

Cluster n points in $\mathbb{R}^d$ into k clusters

Powerful and flexible model: Gaussian mixtures
$X \sim \sum_{i=1}^{k} \pi_i \mathcal{N}(\mu_i, \Sigma_k)$
Note: even common covariance $\Sigma_k = \Sigma$ versatile

University of Hawai'i Mānoa

Clustering in low dimensions, few clusters

$k-$means

choose centers $\mu_1, \ldots, \mu_k$ at random

assign each example to nearest mean

update centers and repeat prior step till convergence

Soft version: Expectation Maximization

Fits most likely GMM iteratively

For Gaussians, soft version of $k-$means

University of Hawai'i Mānoa

In high dimensions

Recall: most probability in $\mathcal{N}(\mu, \sigma^2 I)$ close to $\sigma\sqrt{d}$.

$$\mathsf{P}\left(||X-\mu||^2 \geq \sigma^2\left(d - 2\sqrt{d\ln\frac{1}{\delta}}\right)\right) \geq 1-\delta$$

Probability of finding a point near μ is $\exp(-\mathcal{O}(d)$

University of Hawai'i Mānoa

In high dimensions

Recall: most probability in $\mathcal{N}(\mu, \sigma^2 I)$ close to $\sigma\sqrt{d}$.

$$\mathrm{P}\left(||X-\mu||^2 \geq \sigma^2\left(d - 2\sqrt{d\ln\frac{1}{\delta}}\right)\right) \geq 1-\delta$$

Probability of finding a point near μ is $\exp(-\mathcal{O}(d)$

Need $\exp(\mathcal{O}(d)$ points to even have a point $\leq \frac{1}{2}\sigma\sqrt{d}$!

University of Hawai'i Mānoa

In high dimensions

Recall: most probability in $\mathcal{N}(\mu, \sigma^2 I)$ close to $\sigma\sqrt{d}$.

$$\mathrm{P}\left(||X-\mu||^2 \geq \sigma^2 \left(d - 2\sqrt{d \ln \frac{1}{\delta}} \right) \right) \geq 1-\delta$$

Probability of finding a point near μ is $\exp(-\mathcal{O}(d)$

Need $\exp(\mathcal{O}(d)$ points to even have a point $\leq \frac{1}{2}\sigma\sqrt{d}$!
Most plausible data sizes: "few scattered specks of dust in an enormous void" (Dasgupta '99)

University of Hawai'i Mānoa

In high dimensions

Recall: most probability in $\mathcal{N}(\mu, \sigma^2 I)$ close to $\sigma\sqrt{d}$.

$$\mathrm{P}\left(||X-\mu||^2 \geq \sigma^2\left(d - 2\sqrt{d\ln\frac{1}{\delta}}\right)\right) \geq 1-\delta$$

Probability of finding a point near μ is $\exp(-\mathcal{O}(d)$

Need $\exp(\mathcal{O}(d)$ points to even have a point $\leq \frac{1}{2}\sigma\sqrt{d}$!
Most plausible data sizes: "few scattered specks of dust in an enormous void" (Dasgupta '99)
Low dim algorithms need exponential in d examples

University of Hawai'i Mānoa

Key idea: Project into few dimensions

Linear projections: the projections are Gaussian too!

Key idea: Project into few dimensions

Linear projections: the projections are Gaussian too!

PCA?

University of Hawai'i Mānoa

Key idea: Project into few dimensions

Linear projections: the projections are Gaussian too!

PCA?

Can easily find cases where PCA will not work

it is possible PCA collapses components of the mixture on top of each other (or nearly so)

UNIVERSITY of HAWAI'I® MĀNOA

Key idea: Project into few dimensions

Linear projections: the projections are Gaussian too!

PCA?

Can easily find cases where PCA will not work

it is possible PCA collapses components of the mixture on top of each other (or nearly so)

For clustering, try Johnson-Lindenstrauss:

$\frac{1}{\epsilon^2} \log n$ projections retain all pairwise distances

projected space still too large

exponential in $\frac{1}{\epsilon^2} \log n$ is $n^{\frac{1}{\epsilon^2}}$

University of Hawai'i Mānoa

Key idea: Project into few dimensions

Don't worry about retaining all pairwise distances

- $\mathcal{O}(\log k)$ projections
- retain distances between means
- push points closer to mean in each cluster!

University of Hawai'i Mānoa

Key idea: Project into few dimensions

Don't worry about retaining all pairwise distances

$\mathcal{O}(\log k)$ projections

retain distances between means

push points closer to mean in each cluster!

Series of recent results on several common examples in the low-d space, how they recover parameters in high-d space

University of Hawai'i Mānoa

Distance between means

If $||\mu_1 - \mu_2|| > \Omega(d^{1/4})$, should expect to separate out clusters

Note that in this regime, the spheres are not disjoint

Yet we should expect all points in one cluster to be closer to each other than points in other clusters

University of Hawai'i Mānoa

Why Gaussian mixtures

In principle, GMs can model any continuous distribution

Two particular examples (projects):

- Asset returns (see paper on discord)
- fMRI (see paper on discord)

University of Hawai'i Mānoa

Gaussian random matrices

If A is a $k \times n$ matrix, entries iid Gaussian

rows, cols independently chosen Gaussian multivariate

satisfy something called the Restricted isometry property

all small subset of columns approximately orthogonal

Key property used in Compressed Sensing

extends the Shannon-Nyquist theorem

used to shorten MRI acquisition on conventional equipment, network tomography, radio astronomy and optical interferometry (aperture synthesis)

Compressed Sensing

If x^* is a S-sparse signal in $\mathbb{R}^n$
$y = Ax^*$ (ie k linear measurements of x)

If k is very small, can we still find x^*?
Compare with Shannon-Nyquist sampling

University of Hawaiʻi Mānoa

Convex relaxation

$y = A\mathsf{x}$ is underdetermined

University of Hawai'i Mānoa

Convex relaxation

$y = Ax$ is underdetermined

  infinite solutions

  which solution to choose?

Finding sparsest solution too hard

  NP-hard

University of Hawai'i Mānoa

Convex relaxation

$y = Ax$ is underdetermined
    infinite solutions
    which solution to choose?

Finding sparsest solution too hard
    NP-hard

Compressed sensing to the rescue

University of Hawai'i Mānoa

Convex relaxation

Recall $y = A\mathrm{x}^*$ where $\mathrm{x}^* \in \mathbb{R}^n$ is S-sparse
A: $k \times n$ random Gaussian matrix

Solve $\hat{x} = \arg\min ||\mathrm{x}||_1$ such that $A\mathrm{x} = \mathrm{y}$

University of Hawai'i Mānoa

Convex relaxation

Recall $y = A\mathrm{x}^*$ where $\mathrm{x}^* \in \mathbb{R}^n$ is S-sparse
A: $k \times n$ random Gaussian matrix

Solve $\hat{x} = \arg\min ||\mathrm{x}||_1$ such that $A\mathrm{x} = \mathrm{y}$

Can be solved fast

University of Hawai'i Mānoa

Convex relaxation

Recall $y = A\mathrm{x}^*$ where $\mathrm{x}^* \in \mathbb{R}^n$ is S-sparse
A: $k \times n$ random Gaussian matrix

Solve $\hat{x} = \arg\min ||\mathrm{x}||_1$ such that $A\mathrm{x} = \mathrm{y}$

Can be solved fast
Solution will coincide with the sparsest x provided

- A satisfies the restricted isometry property
- $k > S \log n$
- Another project idea

University of Hawai'i Mānoa

Ridge and Lasso

Already noted, brief review

Matrix Completion

This will be our segue into next topic: NLP
Also a chance to learn about singular values

University of Hawai'i Mānoa

Matrix completion

M is a matrix of user preferences
(Netflix challenge 480189×17770, today much larger)

- user i, movie j, rating in position M_{ij}
- most observations unknown
- observations in set Ω

Complete M using Ω

University of Hawai'i Mānoa

Matrix completion

M is a matrix of user preferences
(Netflix challenge 480189×17770, today much larger)

- user i, movie j, rating in position M_{ij}
- most observations unknown
- observations in set Ω

Complete M using Ω

- no reason this is this even possible!

Reminiscent of compressed sensing

- infering sparse signal with very few measurements
- equivalent of sparsity?

"Sparsity" in matrices

Rank of a matrix, Singular Value Decomposition
Rank 1 matrices
General rank k matrices
Singular value decomposition
Outer product expression
Autoencoders

University of Hawai'i Mānoa

Singular value decomposition/Genres

$$M = U\Sigma V^T$$

Rank = number of non-zero singular values
Low rank r: few (r) singular values
$M = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T$
u_i: how much does each person genre i? v_i: how would a person who likes genre i like each movie?

University of Hawaiʻi Mānoa

Matrix completion

Works when Ω is chosen uniformly at random

The rank of M is low

Singular value are incoherent with the standard basis
    projecting the standard basis to the subspace
    of singular vectors

University of Hawai'i Mānoa

Matrix completion: nuclear norm

$$X = \arg\min ||X||_s$$

such that $X_{ij} = M_{ij},\ ij \in \Omega$ where $||X||_s$ is the nuclear norm

UNIVERSITY of HAWAI'I MĀNOA