High dimensional geometry and Regularization

Narayana Santhanam

EE 645
Mar 3, 2023

University of Hawai'i Mānoa

This week

High dimensional Gaussians

    Johnson Lindenstrauss Lemma

This week

High dimensional Gaussians

- Johnson Lindenstrauss Lemma

Ridge and Lasso

- Explanations
- Compressive sensing
- Matrix norms

High dimensional Gaussian

Multivariate Gaussian

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

$\mu = \mathbb{E}X$ (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?

Concentration of measure

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

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

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

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 $\delta > 0$,

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

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!

Applications of JL lemma

Regression in high dimensions

Some clustering problems

  not always: GMM faster

Sketching and streaming algorithms

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

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

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 $\exp(-\mathcal{O}(d)$

In high dimensions

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

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

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

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

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)

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

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?

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)

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 $n^{\frac{1}{\epsilon^2}}$

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!

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

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

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)

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

Convex relaxation

$y = Ax$ is underdetermined

Convex relaxation

$y = Ax$ is underdetermined

infinite solutions

which solution to choose?

Finding sparsest solution too hard

NP-hard

Convex relaxation

$y = Ax$ is underdetermined

  infinite solutions

  which solution to choose?

Finding sparsest solution too hard

  NP-hard

Compressed sensing to the rescue

Convex relaxation

$\min ||x||_1$ such that $Ax = y$
$x \in \mathbb{R}^n$, S non-zero entries, A is $k \times n$ random Gaussian matrix

Convex relaxation

$\min ||x||_1$ such that $Ax = y$
$x \in \mathbb{R}^n$, S non-zero entries, A is $k \times n$ random Gaussian matrix

Can be solved fast

Convex relaxation

$\min ||\mathrm{x}||_1$ such that $A\mathrm{x} = \mathrm{y}$
$\mathrm{x} \in \mathbb{R}^n$, S non-zero entries, A is $k \times n$ random Gaussian matrix

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

Ridge and Lasso

Already noted

Matrix Completion

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