COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Gaussians in High Dimension

- A one dimensional Gaussian has much of its probability mass close to the origin.
- Does this generalise to higher dimensions?
- A d-dimensional spherical Gaussian with 0 means and σ^2 variance in each coordinate has density:

$$p(\mathbf{x}) = \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}}$$

- Let $\sigma^2 = 1$. Even though the probability density is high within the unit ball, the volume of of the unit ball is negligible and hence the probability mass within the unit ball is negligible.
- When the radius is $\sqrt{d}$, the volume becomes large enough to make the probability mass around the $\sqrt{d}$ radius significant.
- Even though the volume keeps increasing beyond the $\sqrt{d}$ radius, the probability density keeps diminishing. So, the probability mass much beyond the $\sqrt{d}$ radius is again negligible.

- Even though the probability density is high within the unit ball, the volume of of the unit ball is negligible and hence the probability mass within the unit ball is negligible.
- When the radius is $\sqrt{d}$, the volume becomes large enough to make the probability mass around the $\sqrt{d}$ radius significant.
- Even though the volume keeps increasing beyond the $\sqrt{d}$ radius, the probability density keeps diminishing. So, the probability mass much beyond the $\sqrt{d}$ radius is again negligible.
- This intuition is formalised in the next theorem.

Theorem (Gaussian Annulus Theorem)

For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq ||\mathbf{x}|| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.

Theorem (Gaussian Annulus Theorem)

For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq ||\mathbf{x}|| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.

- $\mathbf{E}[||\mathbf{x}||^2] = \sum_{i=1}^{d} \mathbf{E}[x_i^2] = d \cdot \mathbf{E}[x_1^2] = d$.
- So, the average squared distance of a point from center is d. The Gaussian annulus theorem essentially says that the distance of points is tightly concentrated around the distance $\sqrt{d}$ (called *radius* of Gaussian).

Random Projection and Johnson Lindenstrauss (JL)

- Typical data analysis tasks requires one to process d-dimensional point set of cardinality n where n and d are very large numbers.
- Many data processing tasks depends only on the pair-wise distances between the points (e.g., nearest neighbour search).
- Each such distance query has a significant computational cost due to the large value of the dimension d.
- Question: Can we perform dimensionality reduction on the dataset? That is, find a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ with $k << d$ such that the pairwise distances between the mapped points are preserved (in a relative sense).

Claim

There exists a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ with $k << d$ such that the pairwise distances between the mapped points are preserved (in a relative sense).

- Consider the following mapping:

$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, ..., \mathbf{u_k} \cdot \mathbf{v}),$$

where $\mathbf{u_1}, ..., \mathbf{u_k} \in \mathbb{R}^d$ are Gaussian vectors with unit variance and zero mean in each coordinate.

Claim

There exists a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ with $k << d$ such that the pairwise distances between the mapped points are preserved (in a relative sense).

- Consider the following mapping:

$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, ..., \mathbf{u_k} \cdot \mathbf{v}),$$

  where $\mathbf{u_1}, ..., \mathbf{u_k} \in \mathbb{R}^d$ are Gaussian vectors with unit variance and zero mean in each coordinate.
- We will show that $||f(\mathbf{v})|| \approx \sqrt{k}||\mathbf{v}||$.
- Due to the nature of the mapping, for any two vectors $\mathbf{v_1}, \mathbf{v_2} \in \mathbb{R}^d$ we have:

$$||f(\mathbf{v_1}) - f(\mathbf{v_2})|| \approx \sqrt{k} \cdot ||\mathbf{v_1} - \mathbf{v_2}||.$$

- So, the distance between $\mathbf{v_1}$ and $\mathbf{v_2}$ can be estimated by computing the distance between the mapped points and then dividing the result by $\sqrt{k}$.

Claim

For any $\mathbf{v} \in \mathbb{R}^d$, $||f(\mathbf{v})|| \approx \sqrt{k}||\mathbf{v}||$.

Theorem (Random Projection Theorem)

There exists a constant $c > 0$ such that for any $\varepsilon \in (0, 1)$ and $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbf{Pr}\left(\left| ||f(\mathbf{v})|| - \sqrt{k}||\mathbf{v}|| \right| \geq \varepsilon\sqrt{k}||\mathbf{v}||\right) \leq 3e^{-ck\varepsilon^2}.$$

The probability is over the randomness involved in sampling the vectors $\mathbf{u_i}$'s.

Claim

For any $\mathbf{v} \in \mathbb{R}^d$, $||f(\mathbf{v})|| \approx \sqrt{k}||\mathbf{v}||$.

Theorem (Random Projection Theorem)

There exists a constant $c > 0$ such that for any $\varepsilon \in (0, 1)$ and $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbf{Pr}\left(\left|||f(\mathbf{v})|| - \sqrt{k}||\mathbf{v}||\right| \geq \varepsilon\sqrt{k}||\mathbf{v}||\right) \leq 3e^{-ck\varepsilon^2}.$$

The probability is over the randomness involved in sampling the vectors $\mathbf{u_i}$'s.

Proof

- Claim 1: It is sufficient to prove the statement for unit vectors $\mathbf{v}$.
- For all $\mathbf{u_i}$,we have:

$$\mathbf{Var}(\mathbf{u_i} \cdot \mathbf{v}) = \mathbf{Var}(\sum_{j=1}^{d} u_{ij} v_j) = \sum_{j=1}^{d} v_j^2 \mathbf{Var}(u_{ij}) = \sum_{j=1}^{d} v_j^2 = 1.$$

- So, $f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, ..., \mathbf{u_k} \cdot \mathbf{v})$ is a k dimensional Gaussian with unit variance in each coordinate.
- The result now follows from a simple application of the Gaussian Annulus Theorem. □

Claim

For any two vectors $\mathbf{v_1}, \mathbf{v_2} \in \mathbb{R}^d$, $||f(\mathbf{v_1}) - f(\mathbf{v_2})|| \approx \sqrt{k} \cdot ||\mathbf{v_1} - \mathbf{v_2}||$.

Theorem (Johnson-Lindenstrauss (JL) Theorem)

For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ with c as in the Random Projection Theorem. For any set of n points in $\mathbb{R}^d$, the random projection $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as before has the property that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $(1 - \frac{3}{2n})$,

$$(1-\varepsilon)\sqrt{k}||\mathbf{v_i} - \mathbf{v_j}|| \leq ||f(\mathbf{v_i}) - f(\mathbf{v_j})|| \leq (1+\varepsilon)\sqrt{k}||\mathbf{v_i} - \mathbf{v_j}||.$$

Proof

- We obtain the result from the Random Projection Theorem by applying the union bound with respect to at most $\binom{n}{2} < n^2/2$ pairs of points. $\square$

Theorem (Johnson-Lindenstrauss (JL) Theorem)

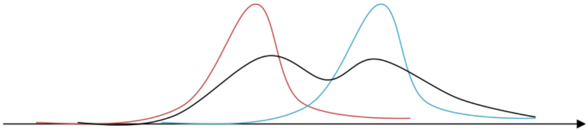
For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ with c as in the Random Projection Theorem. For any set of n points in $\mathbb{R}^d$, the random projection $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as before has the property that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $(1 - \frac{3}{2n})$,

$$(1-\varepsilon)\sqrt{k}||\mathbf{v_i} - \mathbf{v_j}|| \leq ||f(\mathbf{v_i}) - f(\mathbf{v_j})|| \leq (1+\varepsilon)\sqrt{k}||\mathbf{v_i} - \mathbf{v_j}||.$$

- Here is an application of the JL Theorem for the Nearest Neighbour (NN) problem:
  - Suppose we need to pre-process n data points $X \subseteq \mathbb{R}^d$ so that we can answer at most n' queries of the form: "find the point from X that is nearest to a given point $p \in \mathbb{R}^d$".
  - If we use a JL mapping with $k \geq \frac{3}{c\varepsilon^2} \ln(n + n')$, then we can store $f(\mathbf{x})$ for all $\mathbf{x} \in X$. For a query point $\mathbf{p}$, we just return the the point that is nearest to $f(\mathbf{p})$.

Separating Gaussians

- Mixture of Gaussians are used to model heterogenous data coming from multiple sources.
- Consider an example of height of people in a city:
  - Let $p_M(x)$ denote the Gaussian density of height of men in the city and $p_F(x)$ for women.
  - Let w_M and w_F denote the proportion of men and women in the city respectively.
  - So, the mixture model $p(x) = w_M \cdot p_M(x) + w_F \cdot p_F(x)$ is a natural way to model the density of hight of people in the city.



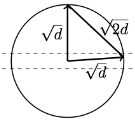
- Mixture of Gaussians are used to model heterogenous data coming from multiple sources.
- Consider an example of height of people in a city:
  - Let $p_M(x)$ denote the Gaussian density of height of men in the city and $p_F(x)$ for women.
  - Let w_M and w_F denote the proportion of men and women in the city respectively.
  - So, the mixture model $p(x) = w_M \cdot p_M(x) + w_F \cdot p_F(x)$ is a natural way to model the density of hight of people in the city.
- The parameter estimation problem is to guess the parameters of the mixture given samples from the mixture.
  - In our above example this means that we are given heights of a number of people of the city and the task is to infer w_M, w_F and the mean and variance of $p_M(x)$ and $p_F(x)$.

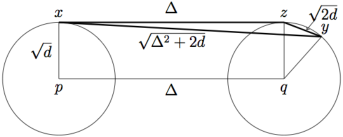
- Mixture of Gaussians are used to model heterogenous data coming from multiple sources.
- Consider an example of height of people in a city:
  - Let $p_M(x)$ denote the Gaussian density of height of men in the city and $p_F(x)$ for women.
  - Let w_M and w_F denote the proportion of men and women in the city respectively.
  - So, the mixture model $p(x) = w_M \cdot p_M(x) + w_F \cdot p_F(x)$ is a natural way to model the density of hight of people in the city.
- The parameter estimation problem is to guess the parameters of the mixture given samples from the mixture.
  - In our above example this means that we are given heights of a number of people of the city and the task is to infer w_M, w_F and the mean and variance of $p_M(x)$ and $p_F(x)$.
  - In the example, given the height of an individual can we infer whether it is a man or a woman?

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- If the means of the Gaussians are too close, then it will be hard to distinguish samples from the distributions. Suppose the distance between the means is Δ.
- We will try to design an algorithm that estimates the parameters for some minimum value on Δ.

Separating Gaussians
Parameter estimation

- We will first consider the following simpler problem of separating unit variance Gaussians:
    - Given samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- If the means of the Gaussians are too close, then it will be hard to distinguish samples from the distributions. Suppose the distance between the means is Δ.
- We will try to design an algorithm that estimates the parameters for some minimum value on Δ.
- <u>Claim 1</u>: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from the same Gaussian. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d} \pm O(1)$ w.h.p.
- <u>Claim 2</u>: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from different Gaussians. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d + \Delta^2} \pm O(1)$ w.h.p.





- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- If the means of the Gaussians are too close, then it will be hard to distinguish samples from the distributions. Suppose the distance between the means is Δ.
- We will try to design an algorithm that estimates the parameters for some minimum value on Δ.
- Claim 1: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from the same Gaussian. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d} \pm O(1)$ w.h.p.
- Claim 2: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from different Gaussians. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d + \Delta^2} \pm O(1)$ w.h.p.
- So, we can distinguish points from the same/different Gaussians based on the pairwise distance as long as $\sqrt{2d} + O(1) \leq \sqrt{2d + \Delta^2} - O(1)$ which implies that $\Delta = \omega(d^{1/4})$.
  - Since we want this for almost all point pairs there is an extra factor of $O(\sqrt{\log n})$ in Δ.

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given n samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- Let the distance between the means be $\Delta = \Omega(d^{1/4}\sqrt{\log n})$.
- Here is an algorithm for separating points from the two Gaussians.

Algorithm

- Calculate pairwise distance between all pairs of points
- The cluster of smallest pairwise distances must come from the same Gaussian. Remove these points.
- The remaining points come from the second Gaussian.

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given n samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- The parameter estimation problem was to estimate the parameters of the Gaussian that the data points are sampled.
- Since, we now have an algorithm for separating points, we should think of how to fit a spherical Gaussian to the given data.

- Given samples $\mathbf{x}_1, ..., \mathbf{x}_n$ in a d-dimensional space, we want to find the spherical Gaussian that best fits the points.
- Let f be an unknown Gaussian with mean μ and variance σ^2 in each direction.
- The probability density of picking these points from this Gaussian is given by $c \cdot exp\left(-\frac{||\mathbf{x}_1-\mu||^2+...+||\mathbf{x}_n-\mu||^2}{2\sigma^2}\right)$.
- The Maximum Likelihood Estimator (MLE) of f, given the samples $\mathbf{x}_1, ..., \mathbf{x}_n$ is the f that maximizes the above probability density.

Theorem

The maximum likelihood spherical Gaussian for a set of samples is the Gaussian with the center equal to the sample mean and standard deviation equal to the standard deviation of the sample from the true mean.

End