High-dimension Gaussians

Separating Gaussians
Fitting Spherical Gaussian to Data

Amit Waisel
Separating Gaussians

• Heterogeneous data coming from multiple sources
• Gaussian mixture model \( p(x) = w_1 p_1(x) + w_2 p_2(x) \)
• Parameter estimation problem: given access to samples from the overall density \( p \), reconstruct the parameters for the distribution (mean and variance for each distribution: \( \mu_1, \mu_2, \sigma_1, \sigma_2 \))

• Mixed-density function
  • Gaussian \( p_i \) has its own mean \( \mu_i \) and variance \( \sigma_i \)
  • Defines the probability to sample Gaussian’s \( p_i \)’s distribution
MVN – Multivariant Normal Distribution

• Defined over vectors, not scalars
• Intuition: each coordinate in the random vector is sampled from a normal distribution
K-means

- We are given \( n \) vectors \( x_1, \ldots, x_n \) and a number \( k \)
- We would like to partition the vectors into \( k \) sets, and with each set we associate a “center” \( \mu_i \)
- The goal is to minimize the objective function
  \[
  \min_{\mu_1, \ldots, \mu_k, S_1, \ldots, S_k} \sum_{j=1}^k \sum_{i \in S_j} \|x_i - \mu_i\|^2
  \]
- We assumed that each point has to be classified to a specific cluster.
  - This is a “hard” decision, since we need to decide for each point a single cluster.
  - We have to make some assumptions (distance between the means), in order to make this problem easier
GMM - Gaussian Mixture Model

- We have $k$ Gaussian distributions, and a mixing distribution.
  - The mixing distribution gives a probability to each cluster
- To generate a point, we sample a Gaussian given the mixture distribution, and then sample the selected Gaussian to generate the point
GMM - Gaussian Mixture Model

- We have $k$ unknown clusters $S_1, \ldots, S_k$ where $S_i \sim N(\mu_i, \sigma_i^2)$
- Each point originates from cluster $j$ with probability $p_j$
- The density function for cluster $j$ is $f_j(x) = \frac{1}{d} \cdot e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}$
Separating Gaussians

- Original objective: separate the samples into their original Gaussian distributions
- Women can be high, men can be low – and we might not be able to know for sure if a specific sample belongs to a male or a female.
  - We can’t know for sure (with high probability) whether a point belongs to a specific Gaussian
- Alternative objective:
  - More difficult: mixture of two Gaussians in high-dimensions \((d\)-dimension space\), rather than 1-dimensional
  - Easier: we assume the means are well-separated compared to the variances
Separating Gaussians

• We will focus on a mixture of two spherical unit-variance Gaussians whose means are separated by a distance $\Omega\left(\frac{1}{d^4}\right)$
  • Goal: Prove that using those assumptions, we can know with high confidence the origin of a given sample. This is k-means with $k = 2$.

• Simple solution: Calculate the distance between all pairs of points.
  • Points whose distance apart is smaller are from the same Gaussian
  • Points whose distance is larger are from different Gaussians
Sample distances – one Gaussian

• Reminder: 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 \( 3e^{-c_1\beta^2} \) of the probability mass lies within the annulus \( \sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta \), where \( c \) is a fixed positive constant
  • A fixed value for \( \beta \) is good for fixed number of sampled points

• Reminder: Volume near the equator
  • For any unit-length vector \( \nu \) defining “north”, most of the volume of the unit ball lies in the thin slab of points whose dot-product with \( \nu \) has magnitude \( O \left( \frac{1}{\sqrt{d}} \right) \)
  • At least \( 1 - \frac{2}{c} e^{-\frac{c^2}{2}} \) fraction of the volume of the \( d \)-dimensional unit ball, has \( |x_1| \leq \frac{c}{\sqrt{d-1}} \)
Sample distances – one Gaussian

• Most of the Gaussian’s probability mass lies on an annulus of width $O(\beta)$ at radius $\sqrt{d}$ from the origin

• Most of its probability mass lies in a $O\left(\frac{1}{\sqrt{d}}\right)$-width slab on the equator
  • For simplicity – assume it is a constant, $O(1)$
Sample distances – one Gaussian

• Consider one spherical unit-variance Gaussian centered at the origin
  • $\mu = (0, ..., 0)$
  • $\sigma^2 = 1$

  • Density function is $f(x) = \frac{1}{\sqrt{2\pi}^d} \cdot e^{-\frac{x^2}{2}}$ (for $e^{-\frac{x^2}{2}} = \prod_i e^{-\frac{x_i^2}{2}}$)

• Almost all of the mass is within the slab $\{x| -\text{const} \leq x_1 \leq \text{const}\}$
Sample distances – one Gaussian

• Pick a random point $x$ from the Gaussian
• Rotate the coordinate system to make the first axis align with $x$
• Pick an independent sample $y$
  • $y$ is located on the equator, with high probability, considering $x$ as the north pole
  • $y$’s component along $x$’s direction, is $O(1)$ with high probability
Sample distances – one Gaussian

• $y$ is nearly-perpendicular to $x$
  • $|x - y|^2 \approx |x|^2 + |y|^2$

• $x$ is considered as the “north pole”
  • $x = (\sqrt{d} \pm O(\beta), 0,0,\ldots,0)$

• $y$ is nearly on the equator, we can further rotate the coordinate system so that the component of $y$ that is perpendicular to the axis of the “north pole”, is in the second coordinate
  • $y = (O(1), \sqrt{d} \pm O(\beta), 0,0,\ldots,0)$

• $|x - y|^2 = 2d \pm O(\beta \sqrt{d})$ with high probability
Sample distances – two Gaussians

• Consider two spherical unit-variance Gaussians with centers \( p, q \) separated by a distance \( \Delta \)
• We want to prove that every two random points, each selected from a different Gaussian, will have significant distance between them
  
  \[
  \approx \sqrt{\Delta^2 + 2d} \pm O(\beta \sqrt{d})
  \]
• Pick \( x \) from the 1\textsuperscript{st} Gaussian and rotate the coordinate system so \( x \) will be the north pole
  
  • Let \( z \) be the north pole of the 2\textsuperscript{nd} spherical Gaussian, using the same coordinate system
• Pick \( y \) from the 2\textsuperscript{nd} Gaussian
  
  • Most of the 2\textsuperscript{nd} Gaussian’s mass is within \( O(1) \) of the equator perpendicular to \( q – z \)
  • Most of the 2\textsuperscript{nd} Gaussian’s mass is within \( O(1) \) of the equator perpendicular to \( q – p \)
Sample distances – two Gaussians

- The distance $|z - y|$ is $O(\sqrt{2d})$
  - Two samples from the same Gaussian
- High-dimension Pythagorean Theorem

$$|x - y|^2 \approx \Delta^2 + |z - q|^2 + |q - y|^2 \approx \Delta^2 + 2d \pm O(\beta \sqrt{d})$$
  - $|z - q|^2 \approx |q - y|^2 \approx (\sqrt{d} \pm O(\beta))^2 \approx d \pm O(\beta \sqrt{d}) + \beta^2$
Sample distances - assumptions

• We have to ensure that the distance between two points picked from the same Gaussian are closer to each other, than two points picked from different Gaussians
  • The upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of the distance between points from different Gaussians
• Squared-distance between two points picked from the same Gaussian
  • $2d \pm O(\beta \sqrt{d})$
• Squared-distance between two points picked from different Gaussians
  • $\Delta^2 + 2d \pm O(\beta \sqrt{d})$
• $2d \pm O(\beta \sqrt{d}) \leq \Delta^2 + 2d \pm O(\beta \sqrt{d})$ holds for $\Delta \in \omega\left(d^{\frac{1}{4}}\right)$, as needed
Separating Gaussians - algorithm

• Calculate all pairwise distances between points
• The cluster of smallest pairwise distances must come from a single Gaussian
  • Remove these points
• The remaining points come from the second Gaussian
• We used a constant $\beta$. What happens if we take $n$ samples?
  • Any fixed $\beta$ will not be good enough. $\beta$ has to be dependent on $n$
  • $\beta = O(\sqrt{\ln n})$ is a good value for the annulus-theorem equation
    • The probability to sample the annulus is $1 - 3e^{-c_1\beta^2} = 1 - \frac{3}{n^{c_1}} = 1 - \frac{1}{poly(n)}$
Sample distances – one Gaussian, \( n \) samples

- \( y \) is nearly-perpendicular to \( x \)
  - \(|x - y|^2 \approx |x|^2 + |y|^2\)
- \( x \) is considered as the “north pole”
  - \( x = (\sqrt{d} \pm O(\sqrt{\ln n}), 0,0,...,0)\)
- \( y \) is nearly on the equator, we can further rotate the coordinate system so that the component of \( y \) that is perpendicular to the axis of the “north pole”, is in the second coordinate
  - \( y = (O(1), \sqrt{d} \pm O(\sqrt{\ln n}), 0,0,...,0)\)
- \(|x - y|^2 = 2d \pm O(\sqrt{\ln n} \sqrt{d})\) with high probability
Sample distances – two Gaussians, $n$ samples

- The distance $|z - y|$ is $O(\sqrt{2d})$
  - Two samples from the same Gaussian
- High-dimension Pythagorean Theorem
  - $|x - y|^2 \approx \Delta^2 + |z - q|^2 + |q - y|^2 \approx \Delta^2 + 2d \pm O(\sqrt{\ln n} \sqrt{d}) + O(\ln n)$
    - $|z - q|^2 \approx |q - y|^2 \approx (\sqrt{d} \pm O(\sqrt{\ln n}))^2 \approx d \pm O(\sqrt{\ln n} \sqrt{d}) + O(\ln n)$
Sample distances – assumptions, $n$ samples

• We have to ensure that the distance between two points picked from the same Gaussian are closer to each other, than two points picked from different Gaussians
  • The upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of the distance between points from different Gaussians

• Squared-distance between two points picked from the same Gaussian
  • $2d \pm O(\sqrt{\ln n} \sqrt{d})$

• Squared-distance between two points picked from different Gaussians
  • $\Delta^2 + 2d \pm O(\sqrt{\ln n} \sqrt{d})$

• $2d \pm O(\sqrt{\ln n} \sqrt{d}) \leq \Delta^2 + 2d \pm O(\sqrt{\ln n} \sqrt{d})$ holds for $\Delta \in \omega \left( (d \cdot \ln n)^{\frac{1}{4}} \right)$
Fitting a Spherical Gaussian to Data

• Given $d$-dimensional sample points $x_1, \ldots, x_n$, our objective is to find a spherical Gaussian that best fits those points
  • Find the distributions’ mean $\mu$ and variance $\sigma^2$
• Let $f$ be a Gaussian with mean $\mu$ and variance $\sigma^2$
  • $\mu$ is a $d$-dimensional vector, containing the mean values for each dimension
Fitting a Spherical Gaussian to Data

• $f$’s density function is $f(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^d} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ (remember: $x, \mu$ are vectors)

• Definition: MLE (Maximal Likelihood Estimator) of a set samples $x_1, \ldots, x_n$ is the density function $f$ that maximizes the above probability density

• We want the Gaussian which gives us the highest probability to get the data $x_1, \ldots x_n$ under its density function $f(x)$.
  • The density function becomes a function of $\mu, \sigma$ instead of $x$, because we want to maximize the likelihood

  • $F(\mu, \sigma) = f(x_1, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^\frac{d}{2}} \cdot \prod_{i=1}^{n} \left[ e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right]$ where $x_1, \ldots, x_n$ are vectors
MLE— mean value $\mu$

• Let $x_1, \ldots, x_n$ be samples in $d$-dimensional space. We will prove that $(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2$ is minimized when $\mu$ is the centroid of $x_1, \ldots, x_n$

  • Why minimize? $F(\mu, \sigma) = c \cdot e^{-\frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{2\sigma^2}}$, so maximizing $F'$s value is minimizing $(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2$

• Proof: We would like to find the minimal point of the sum, by finding its derivative

  • Note that every part of the sum is a vector, its first coordinate is the derivative by $\mu_1$.
    • Each row is a different derivative, and we have $d$ derivatives in total
  • $-2(x_1 - \mu) - \cdots - 2(x_n - \mu) = 0$
  • Solving for $\mu$ gives $\mu = \frac{1}{n}(x_1 + \cdots + x_n)$
MLE – variance \( \sigma^2 \)

• We would like to find the MLE of \( \sigma^2 \) for \( f \).

• Let \( \mu \) be the real centroid.

• Let \( \nu = \frac{1}{2\sigma^2} \) and \( a = \sum_{i=1}^{n} (x_i - \mu)^2 \) (for simplicity).

• The density function is now \( f(x) = \left[ \frac{\nu}{\pi} \right]^{dn} \cdot e^{-av} \)
  
  • To find its maximal value, we will find the derivative of \( \ln f(x) \)
  
  • \( \frac{dn}{2\nu} - a = 0 \)
  
  • \( \sigma = \frac{\sqrt{a}}{\sqrt{nd}} \)

• \( \sigma \) is the square root of the average coordinate distance squared of the samples to their mean (the definition of standard deviation!)