Khoury College of Computer Sciences
Northeastern University
CS 7870: Seminar in Theoretical Computer Science

Fall 2023
13 October 2023
Scribe: Eric Chapdelaine

Outline:

- Spherical Gaussians
- Warm-up: Clustering Mixture of Two Gaussians
- Clustering Mixture of k Guassians

The presentation of the material in this lecture is heavily drawn from Chapters 2 and 3 of the text by Blum, Hopcroft, and Kannan [BHK20].

1 Spherical Gaussians

We begin our study of high-dimensional spaces by considering points drawn from a spherical Gaussian distribution. Gaussians appear in many real-world scenarios; they also have nice properties in how they combine, which enables rigorous analyses and derivation of formal bounds. One such property is that two Gaussians, when added together, form a new Gaussian. That is, given $x \in N(\mu_1, \sigma_1)$ and $y \in N(\mu_2, \sigma_2)$, the distribution $x + y$ is also a Gaussian with mean $\mu = \mu_1 + \mu_2$ and standard deviation $\sigma = \sigma_1 + \sigma_2$ (assuming x and y are independent).

Recall the properties of the unit ball described in lectures 9 and 10. Informally, they are:

1. Volume is concentrated in an annulus close to the surface (the very edge of the ball).
2. If x is drawn uniformly at random from the unit ball, then $|x_i| \approx \frac{1}{\sqrt{d}}$ (with high probability) for each coordinate i.
3. Any two random vectors from a unit ball are (nearly) orthogonal.

The d-dimension spherical Gaussian distribution Π over $\Re^d$ has the probability density function defined as follows:

$$p(x) = ce^{-\frac{(x-\mu)^2}{2\pi^2}},$$

where μ and σ are the mean and standard deviation of the distribution.

Alternatively, we draw a point $x = (x_1, \ldots, x_d)$ according to a zero-mean unit-variance d-dimension spherical Gaussian by drawing each x_i according to a zero-mean unit-variance Gaussian.

1.1 Gaussian Annulus Theorem

Informally, the Gaussian Annulus Theorem states that the volume of a spherical Gaussian is mostly concentrated in a small annulus at radius $\sqrt{d}$ (which may not be intuitive as this is not the case in lower dimensions).

Theorem 1. If x is drawn from a 0-mean, unit-variance spherical Gaussian in dimension d, then for any $\beta \leq \sqrt{d}$

$$\mathbb{P}\left(\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta\right) \geq 1 - 3e^{-\beta^2 C} \tag{1}$$

for some constant $c > 0$.

2 Warm-up: Clustering Mixture of Two Gaussians

Suppose we have a distribution that is made up of two Gaussian distributions and we are given samples from this mixed distribution and we want to find the parameters (mean and variance) for the two Gaussians as well as their weights in the total distribution. An example of such a problem (from [BHK20]) would be people's heights as sampled from a population. We know that men, on average, tend to be taller than women and so if we were to sample the height of people (a mix of men and women) in a given population, we would expect to see two Gaussians from this sample's distribution, one for men's heights and one for women's heights. Our goal would be to find what percentage of the sample are men and women as well as the mean and variance of men's heights and women's heights.

It should be clear that if the means of the two Gaussians are far apart, it is fairly easy to distinguish. However, how separated should the two means be to allow separation merely by distance?

2.1 Properties of points from the same Gaussian

Let's draw a and b randomly from the same spherical Gaussian distribution. To answer the question above, we want to bound $|a - b|$. We first observe the following.

$$|a| \approx \sqrt{d} \pm O(1) \qquad \text{(By theorem 1)}$$
$$|b| \approx \sqrt{d} \pm O(1) \qquad \text{(By theorem 1)}$$

Thus a and b are both vectors in a ball of radius $\sqrt{d} \pm O(1)$. Since a and b are both drawn at random, we also know that they are nearly orthogonal. By then rotating these vectors so that a is the North pole and the two vectors lie on a 2-D coordinate plane, we get:

$$a = \left(\sqrt{d} \pm O(1), 0, 0, \cdots, 0\right)$$
$$b = \left(O(1), \sqrt{d} \pm O(1), 0, \cdots, 0\right) \qquad \text{(By fact 3 about unit balls)}$$

It follows that

$$\begin{aligned} |a - b|^2 &\approx \left(\sqrt{d} \pm O(1)\right)^2 + \left(\sqrt{d} \pm O(1)\right)^2 \\ &\approx 2d + O(\sqrt{d}) \end{aligned}$$

Now, what if these x and y were from different distributions?

2.2 Properties of points from different Gaussians

Let's now consider two spherical, unit-variance Gaussians, p and q, that are Δ away from each other. Let's have p centered at $\vec{0} = (0, 0, \cdots, 0)$ and q centered at $\mu = (\Delta, 0, \cdots, 0)$. Suppose we pick x from p and y from q.

$$|x| \approx \sqrt{d} \pm O(1)$$
$$|y - \mu| \approx \sqrt{d} \pm O(1)$$

First, let's rotate the coordinate system so that y is at the North Pole of q, which would be at coordinate $(\Delta, \sqrt{d} \pm O(1), 0, \cdots, 0)$ and x is at $(0, 0, \sqrt{d} \pm O(1), \cdots, 0)$.

Let z be at the North Pole of p. This would be the equivalent of taking z and x from the same Gaussian. It follows that:

$$\begin{aligned}
|x - y|^2 &\approx |x - z|^2 + |z - y|^2 \\
&\approx |x|^2 + |z|^2 + |z + y|^2 \\
&\approx \left(d \pm \sqrt{d}\right) + \left(d + \sqrt{d}\right) + \Delta^2 \\
&\approx 2d \pm O\left(\sqrt{d}\right) + \Delta^2
\end{aligned}$$

If we want to distinguish two points from different Gaussians, the distance between two points from the same Gaussian must be closer to each other than two points picked from different Gaussians. It follows that

$$\begin{aligned}
2d + O(\sqrt{d}) &\leq 2d - O(\sqrt{d}) + \Delta^2 \\
&\implies \Delta^2 = \omega\left(\sqrt{d}\right) \\
&\implies \Delta = \omega\left(d^{1/4}\right).
\end{aligned}$$

3 Clustering mixture of k Gaussians

The above section implies that distance-based separation works in d dimensions as long as the separation between means are $\omega\left(d^{1/4}\right)$. The problem with this is that d can be very large, so we want to project it into a smaller-dimensional subspace. In order for us to do that, we must:

- Keep the same inter-mean distances (if distances become smaller by the projection, then it is harder to use distance-based separation)
- Maintain the properties of the Gaussians

Let's first focus on one Gaussian with mean μ. What is the best-fit line for this Gaussian if we want to project it into a 1-dimensional space?

Recall that the best-fit line is given by the vector v such that we maximize:

$$\begin{aligned}\mathbb{E}_{x\sim\pi}\left(\left(v^T x\right)^2\right) &= \mathbb{E}\left(\left(v^T(x-\mu)+v^T\mu\right)^2\right)\\ &= \mathbb{E}\left(\left(v^T(x-\mu)\right)^2\right)+\mathbb{E}\left(v^T\mu\right)^2+2\mathbb{E}\left(v^T(x-\mu)v^T\mu\right)\\ &= \sigma^2+\left(v^T\mu\right)^2\end{aligned}$$

Since we want to maximize the above, we choose $v^T = \frac{\mu}{|\mu|}$. This means that the best-fit, 1-dimensional line for a Gaussian is a line that goes through its mean.

Claim 1. Any projection of a spherical Gaussian with standard deviation σ onto a k-dimensional subspace containing the mean yields another spherical Gaussian with the same standard deviation.

Proof. We establish this using standard Gaussian properties. Recall the probability density function for the Gaussian distribution:

$$p(x) = ce^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Suppose we project to k-dimensional subspace V. First, rotate the coordinate system such that we have: $x \to (\underbrace{x'_1, x'_2, \cdots, x'_k}_{x'}, \underbrace{x'_{k+1}, \cdots}_{x''})$, where x' and x'' have mean μ' and μ'' respectively. The density function for the projected points is

$$\begin{aligned}p(x') &= ce^{\frac{-|x'-u'|^2}{2\sigma^2}}\int_{x''} e^{\frac{-|x''\mu''|^2}{2\sigma^2}}\\ &= c'e^{\frac{-|x'-u'|^2}{2\sigma^2}} \qquad \text{(Because the integral is a constant)}\end{aligned}$$

□

3.1 Algorithm

We now can cluster a mixture of k Gaussians as follows:

1. Project to best-fit, k-dimensional subspace. We use the following claim without proof (which is similar to the claim above).

   Claim 2. The best-fit k-dimensional subspace contains all k centers and therefore stays a Gaussian

2. Use distance-based clustering algorithm of Section 2 in the k-dimensional projected subspace. Since the distance between any two centers remains the same in the projection, we obtain that the clustering is correct as long as the separation between any two centers is $\omega\left(k^{1/4}\right)$.

The singular value decomposition (SVD) algorithm from the previous lecture, in fact, finds the best-fit k-dimensional subspace (for each value of k). This gives a simple efficient algorithm for clustering a mixture of k Gaussians in d dimensions, subject the condition that every pair of centers is separated by $\omega(k^{1/4})$ (note that the distance is independent of the dimension).

References

[BHK20] Avrim Blum, John Hopcroft, and Ravi Kannan. *Foundations of Data Science*. Cambridge University Press, Cambridge, 2020.