Geometry of High Dimensional Space

Subhash Suri

September 25, 2019

1 High Dimensional Space

- One of the most useful (conceptually, visually, technically) representations of data is as a collection of vectors in some feature space. In other words, as "points" in a multi-dimensional space, typically R^d.
- Some examples: bag-of-words model for documents, pixel models for images, user-product-rankings (netflix, amazon, Yelp), time series data etc.
- In the bags-of-words (Salton's vector space model), we have a vocabulary, say, of m words, with the ith word associated with position i in the vector of dim m. A piece of text is mapped to a m-vector $(x_1, x_2, \ldots, x_m)$, where x_i may be the "frequency" of the ith word.
- Even though all the context of words is lost—we only remember how many times a word occurs, not its local context—the model empirically performs well. There are N-gram extensions whose vocabulary is the *sequence* of N words.
- This seems like a trivial idea. Why represent them as vectors, instead of just a list? After all, these vectors are likely to be quite sparse.
- Turns out there are *mathematical* advantages to using the vector representation. For instance, suppose instead of frequency, we use a 0-1 vector, indicating presence of absence of words. Then, the *dot product* of two vectors X and Y tells us how many words are in common? In general, using the frequency, we get a *correlation* measure.
- For the purpose of this course, we *delegate* the *modeling* aspects of data science domain experts. For instance, in the BoW representation, what is the best choice for each x_i? Should we use the 0-1 model, the frequency, the $\sqrt{f}$, or $\log f$, or some other measure? These considerations are left to the data *modeler*, probably a domain expert, who decides on the right representation of data. We assume an appropriate data model is

given, and focus on best algorithms we can design or best estimates we can derive from the data.

- The geometric viewpoint comes equipped with concepts that seem helpful for discovering relationships among input vectors: distances (similarity) between two vectors, nearest neighbors, clustering, separability, minimum enclosing ball, best fit subspaces etc.

- Mathematically speaking, the d-dimensional space is a straightforward generalization of our physical 2- or 3-dimensional space. But the geometry of high-dim space turns out to behave very differently and leads to many counter-intuitive (seemingly paradoxical) behavior. In the next several lectures, we explore these strange phenomena so when our algorithms employ nearest neighbors or balls of radius r, or reason about multi-variable Gaussian samples, we know what to expect.

2 d-Dim Gaussian Points

- Mixtures of d-dim Gaussians are an important an ubiquitous model for data in many domains, including AI, computer vision, medical imaging, psychology, geology etc.

- Each coordinate of the d-dim point is generated using a Gaussian, say, $N(0, 1)$, with mean 0 and variance 1.

- It turns out that there is an intimate relationship between such multivariate Gaussian data and balls (hyperspheres) in d-dimensions. We will find that in d-dim the balls, namely, the sets $\{\mathbf{x} : |\mathbf{x}|^2 = 1\}$ have many strange and counterintuitive properties, which have implications for the Gaussian mixture models.

- Let's begin with one such property. Suppose we choose two random points $\mathbf{y}$ and $\mathbf{z}$ using d-dim Gaussian distribution. What is the expected distance between them? What is the variance of this distance?

- We find that the distance $|\mathbf{y} - \mathbf{z}|^2$ is almost always (*with high probability*) about $2d$. This is not the behavior in one dimension. For instance, if y and z were uniformly distributed in $[0, 1]$, the distance will often be $1/6$, the expected value, but also close to 0 or 1 frequently.

- Similarly, if we choose a random Gaussian point in one dimension, we expect to find many points near the origin (expected distance is 0) but also many others distributed farther out. Standard deviation says that we expect distance to be more than 1 about 32% of the time, and less than 1 68% of the time.) *However, in d dimensions, we find that all points have $|z|^2 \approx d$, with high probability.*

- To get a quick feel for it, we notice that

$$|\mathbf{y} - \mathbf{z}|^2 \;=\; \sum_{i=1}^{d} (y_i - z_i)^2$$

and so the squared distance is the sum of d independent samples of a random variable $x = (y_i - z_i)^2$, the squared difference of two Gaussians.

- When d is large, the Law Of Large Numbers tells us that the sum is very close to the expectation, *with high probability.*

- We will study this in more detail later, but for now recall that LLN states

$$Pr\left[\left|\frac{x_1 + x_2 \ldots + x_n}{n} \;-\; E[x]\right| \;\geq\; \varepsilon\right] \;\leq\; \frac{\mathrm{Var(x)}}{n\varepsilon^2}$$

In our case, we can show that the $Var(x) = 2$, and the number of *samples* is d, the number of coordinates, and therefore the prob. that the squared distance deviates from the expectation by more than εd will be less than $\frac{2}{d\varepsilon^2}$, which goes to zero for large d.

- Similarly, the squared distance of a random point $\mathbf{z}$ from the origin is $|\mathbf{z}|^2 = \sum_{i=1}^{d} z_i^2$, which is the sum of d independent samples. By LLN, the prob. that this sum deviates from expected value d by more than ε is vanishingly small.

- These results suggest that geometry of high dimensions is quite different from geometry of low dimensions. In order to gain some insight into it, we look at the behavior of the *unit ball* in d dimensions, specifically in relation to the unit cube.

3 d-Dim Balls and Cubes

- Consider the cube with side length 1 in d dimensions.

- Its d-dim volume is 1 *for all dim d.*

- If increase the side length to 2, the volume of the cube becomes 2^d.

- In general, the volume of cube with side length r is r^d.

- Now, consider a d-dim ball of radius 1. Suppose its volume is V.

- The volume of a ball of radius r is $r^d \times V$.
  This follows from integration in calculus where we "integrate" the volume of an arbitrary solid by dividing it into cubes of side length dx.

- So the ball's volume also grows exponentially with the radius.
- But the important distinction is between the volumes of the Cube and the Ball of *equal* side length and radius, say, 1.
- The volume of the unit ball is π in $2d$, $4\pi/3$ in $3d$, and

$$\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$$

  in d dimensions where Γ is the generalized factorial function for non-integer values.

- See the picture. One cannot draw the high dim cube in 2D, so this picture is only meant to convey a rough idea of how various parts of the cube sit in terms of distances, but should not be taken literally. After all, the hypercube is *convex*, and does not look like this.
- In 2 dimension, the cube lies entirely inside the unit ball. In dim $d = 4$, the corners of the cube touch the ball. In higher dimensions, the cube begins to jut out of the ball.
- So, while the unit cube's volume remains constant, the unit ball's volume initially grows with dimension d (until about $d = 5$), and then begins to *decrease* and it essentially goes to 0 as $d \to \infty$!

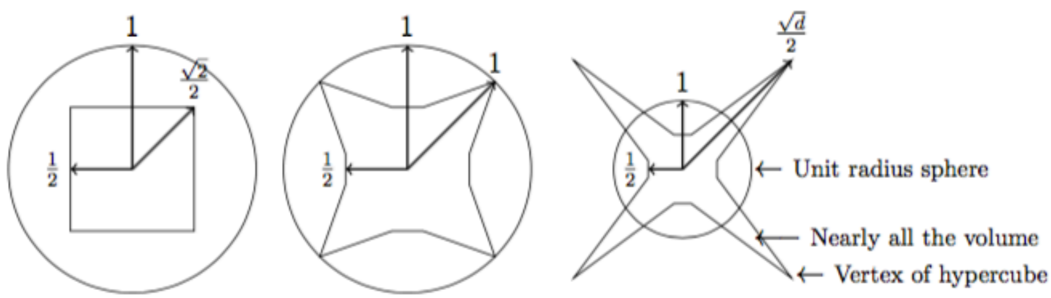

Figure 2.4: Illustration of the relationship between the sphere and the cube in 2, 4, and d-dimensions.