Vector Search **Class #XXXXX Fall 2023**

Lecture number XX: Curse of Dimensionality

Lectures: Edo Liberty and Matthijs Douze

Warning: *Please do not cite this note as a peer reviewed source. If you find mistakes, please inform the authors.*

---

1 Curse of Dimensionality

A prime example for the curse of dimensionality is that a random point in $[0,1]^d$ is likely to be far from any set of n points in the unit cube. Consider the distance between the query point q and an input data vector x. We want to show that $\|x_i - q\|^2 \in \Omega(d)$.

First, notice that $\Pr[|x(j) - q(j)| \geq 1/4] \geq 1/2$. The expected distance between x and q is at least $d/8$. Since $q(j)$ are independently drown, we can apply the Chernoff bound and get that for all n points in the data set $\|x_i - q\|^2 \geq d/16$ if $d \geq \text{const} \cdot \log(n)$.

Now, consider the kd-tree data structure and algorithm run on a random query. If the radius of the ball around q is less than $d/16$ the query is "uninteresting" since it is likely to return no results at all. On the other hand, if the radius is greater than $d/16$ than the ball around q will cross all the major partitions along one of the axis. That means that the algorithm will visit at least 2^d partitions.

2 Volumes of balls and cubes

Another interesting phenomenon that occurs in high dimensions in the fact that unit spheres are exponentially smaller (in volume) than their containing boxes. Let us see this without using the explicit formulas for the volume of d dimensional spheres.

To compute the volume of a unit sphere, we perform a thought experiment. First, bound the sphere in a box (with sides of length 2). Then, pick a point in the box uniformly at random. What is the probability p that the point is also in the sphere? This is exactly the ratio between the volume of the ball and the box (2^d). More accurately, $V = p2^d$ where V is the volume of the sphere.

Now, we can bound p from above. A uniformly random chosen point from the cube is a vector $x \in \mathbb{R}^d$ such that each coordinate $x(i)$ is chosen uniformly from $[-1,1]$. The event that x is in the unit sphere is the event that $\|x\|^2 = \sum_{i=1}^d x(i)^2 \leq 1$. Let $z_i = x(i)^2$, and note that $\mathbb{E}[z(i)] = \int_{-1}^1 \frac{1}{2}t^2 dt = 1/3$. Therefore, $\mathbb{E}[\|x\|^2] = d/3$. Also,

$$\text{Var}(z_i) = \int_{-1}^{1} \frac{1}{2}t^4 dt - (1/3)^2 = 1/5 - 1/9 \leq 1/10$$

so by Chernoff's inequality. $p = \Pr[\sum_{i=1}^d x(i)^2 \leq 1] = \Pr[\sum_{i=1}^d (z_i - \mathbb{E}[z_i]) \leq 1 - d/3] \leq e^{-\frac{(d/3)^2}{4d/10}} \leq e^{-d/4}$. This concludes the observation that the fraction of the volume which is inside the sphere is exponentially small compared to the cube. A counter intuitive way of viewing this is that almost the entire volume of the cube is concentrated at the "corners".

3 Orthogonality of random vectors

It turns out that two random vectors are also almost orthogonal. We can see this in two ways.

First, we can see that the expected, dot product of any vector x with a random vector y is small. It is trivial that $\mathbb{E}[\langle x, y\rangle] = 0$ since the distribution of y is symmetric. Moreover, $\mathbb{E}[\langle x, y\rangle^2] = 1/d$. To see this,

consider $y_1, y_2, \ldots, y_d$ where $y_1 = y$ and $y_2, \ldots, y_d$ complete y to an orthogonal basis. Clearly, the distribution of all y_i are identical (but not independent) $\mathbb{E}[\langle x, y\rangle^2] = \mathbb{E}[\langle x, y_1\rangle^2] = \mathbb{E}[\frac{1}{d}\sum_{i=1}^{d}\langle x, y_i\rangle^2] = \frac{1}{d}\|x\|^2 = \frac{1}{d}$.

It is not hard to show that in fact for any vector x, if y is chosen uniformly at random from the unit sphere then $\Pr[\langle x, y\rangle \geq \frac{t}{\sqrt{d}}] \leq e^{-t^2/2}$. First, replace that uniform distribution over the unit sphere with an i.i.d. distribution of Gaussians $y(i) \sim \mathcal{N}(0, \frac{1}{\sqrt{d}})$. Note that $E[\|y\|^2] = 1$, moreover, from the sharp concentration of the χ^2 distribution we know that $E[\|y\|^2] \approx 1$. For convenience we will assume that $E[\|y\|^2] = 1$ and will ignore the small inaccuracy. Moreover, due to the rotational invariance of the Gaussian distribution we have that any direction is equally likely and thus this new distribution approximates the uniform distribution over the sphere. Next, notice that due to the rotational invariance $\langle x, y\rangle \sim \mathcal{N}(0, \frac{\|x\|}{\sqrt{d}}) = \mathcal{N}(0, \frac{1}{\sqrt{d}})$. Therefore, letting $Z \sim \mathcal{N}(0, 1)$ we have $\Pr[\langle x, y\rangle \geq \frac{t}{\sqrt{d}}] = \Pr[Z \geq t] \leq e^{-t^2/2}$.

References