Long Term Memory in AI - Vector Search and Databases **COS 597A Fall 2023**

Class 4 - Low Dimensional Vector Search

Lectures: Lectures: Edo Liberty and Matthijs Douze

Warning: *Please do not cite this note as a peer reviewed source. Please submit requests and corrections as issues or pull requests at github.com/edoliberty/vector-search-class-notes*

---

1 Probability Tools and Facts Recap

A variable X is a random variable if it assumes different values according to a probability distribution. For example, X can denote the outcome of a three sided die throw. The variable X takes the values $x = 1, 2, 3$ with equal probabilities.

The expectation of X is the sum over the possible values times the probability of the events.

$$\mathbb{E}[X] = \sum_{x=1}^{3} x \Pr(X = x) = 1\frac{1}{3} + 2\frac{1}{3} + 3\frac{1}{3} = 2 \tag{1}$$

Continuous random variable are described by their distribution function φ.

$$\Pr[Y \in [a, b]] = \int_a^b \varphi(t)dt.$$

For a function φ to be a valid distribution we must have:

$$\forall\, t, \quad \varphi(t) \quad \geq \quad 0 \quad \text{(where it is defined)} \tag{2}$$

$$\int_a^b \varphi(t)dt \qquad \text{is well defined for all } a \text{ and } b \tag{3}$$

$$\int_{-\infty}^{-\infty} \varphi(t)dt \quad = \quad 1 \tag{4}$$

For example consider the continuous variable Y taking values in $[0, 1]$ uniformly. That means $\varphi(t) = 1$ if $t \in [0, 1]$ and zero else. This means that the probability of Y being in the interval $[t, t + dt]$ is exactly dt. And so the expectation of Y is:

$$\mathbb{E}[Y] = \int_{t=0}^{1} t\varphi(t)dt = \int_{t=0}^{1} t \cdot dt = \frac{1}{2}t^2|_0^1 = 1/2 \tag{5}$$

Remark 1.1. *Strictly speaking, distributions are not necessarily continuous or bounded functions. In fact, they can even no be a function at all. For example, the distribution of X above includes three Dirac-δ functions which are not valid functions.*

1.1 Dependence and Independence

A variable X is said to be *dependent* on Y if the distribution of X given Y is different than the distribution of X. For example. Assume the variable X takes the value 1 if Y takes a value of less than $1/3$ and the values 2 or 3 with equal probably otherwise ($1/2$ each). Clearly, the probability of X assuming each of its

values is still 1/3. however, if we know that Y is 0.7234 the probability of X assuming the value 1 is zero. Let us calculate the expectation of X given Y as an exercise.

$$\mathbb{E}(X|Y) = \sum_{x=1}^{3} x \Pr(X = x|Y \leq 1/3) = 1 \cdot 1 \tag{6}$$

$$\mathbb{E}(X|Y) = \sum_{x=1}^{3} x \Pr(X = x|Y > 1/3) = 1 \cdot 0 + 2\frac{1}{2} + 3\frac{1}{2} = 2.5 \tag{7}$$

$E(X|Y) = 1$ for $y \in [0, 1/3]$ and $E(X|Y) = 2.5$ for $y \in (1/3, 1]$.
Remember: $\mathbb{E}(X|Y)$ is a function of y!

Definition 1.1 (Independence)**.** *Two variables are said to be Independent if:*

$$\forall y, \ \Pr[X = x|Y = y] = \Pr[X = x].$$

They are dependent otherwise.

Fact 1.1. *If two variables X and Y are Independent the so are $f(X)$ and $g(Y)$ for any functions f and g.*

Fact 1.2 (Linearity of expectation 1)**.** *For any random variable and any constant α:*

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X] \tag{8}$$

Fact 1.3 (Linearity of expectation 2)**.** *For any two random variables*

$$\mathbb{E}_{X,Y}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{9}$$

even if they are dependent.

Fact 1.4 (Multiplication of random variables)**.** *For any two* **independent** *random variables*

$$\mathbb{E}_{X,Y}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{10}$$

This does not necessarily hold if they are dependent.

Definition 1.2 (Variance)**.** *For a random variable X we have*

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \tag{11}$$

The standard deviation σ of X is defined to be $\sigma(X) \equiv \sqrt{\mathrm{Var}[X]}$.

Definition 1.3 (Additivity of variances)**.** *For any two* **independent** *variables X and Y we have*

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] \tag{12}$$

Fact 1.5 (Markov's inequality)**.** *For any positive random variable X:*

$$\Pr(X > t) \leq \frac{E[X]}{t} \tag{13}$$

Fact 1.6 (Chebyshev's inequality)**.** *For any random variable X*

$$\Pr[|X - E[X]| > t] \leq \frac{\sigma^2(X)}{t^2} \tag{14}$$

Lemma 1.1 (Chernoff's bound). *Let $X_1, \ldots, X_n$ be independent Bernoulli trials $\Pr[X_i = 1] = p_i$. And let $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X] = \sum_{i=1}^{n} p_i$.*

$$\Pr[X > (1+\varepsilon)\mu] \leq e^{-\mu\varepsilon^2/4} \tag{15}$$

$$\Pr[X < (1-\varepsilon)\mu] \leq e^{-\mu\varepsilon^2/2} \tag{16}$$

Lemma 1.2 (The union bound). *For any set of m events $A_1, \ldots, A_m$:*

$$\Pr[\cup_{i=1}^{m} A_i] \leq \sum_{i=1}^{m} \Pr A_i.$$

In words, the probability that one or more events happen is at most the sum of the individual event probabilities.

2 kd-trees

Nearest neighbor search is a fundamental computational building block in computer vision, graphics, data mining, machine learning, and many other sub-fields. As an example, consider a simple k-nearest-neighbor-classifier which, for each point predicts its class by the a majority vote over its neighbors' classes. As simplistic as this classifier sounds, it actually performs very well in many scenarios.

We could easily compute this by scanning the data points and finding those which minimize $\|q - x_i\|$. But, of course, it would be significantly better if we could reduce the dependence on n, or in other words, avoid scanning the data for every query point. One possible definition of the nearest neighbor problem is as follows:

Definition 2.1. Nearest Neighbor Search: *Given a set of points $\{x_1, \ldots, x_n\} \in \mathbb{R}^d$ preprocess them into a data structure X of size $\mathrm{poly}(n, d)$ in time $\mathrm{poly}(n, d)$ such that nearest neighbor queries can be performed in logarithmic time. In other words, given a search query point q a radius r and X one can return all x_i such the $||q - x_i|| \leq r$. The search for x_i should require at most $\mathrm{poly}(d, \log(n))$ time.*

3 kd-trees

First, we shall review a well known an widely used algorithm for this problem which is called kd-trees [1]. The data structure holds the points in a tree. Each subtree contains all the points in an axis aligned bounding box (maybe infinite). In each depth in the tree the bounding boxes are split along an axis (in a round robin order) at the median value of the points in the box. Splitting stops when the number of points in the box is sufficiently small, say one.

It is quite simple to prove that inserting and deleting points requires $O(log(n))$ time. Thus, the entire construction time is bounded by $O(n \log(n))$.

Searching however is quite a different matter. Let us assume w.o.l.g. that all points are inside the unit cube. A harmless assumption because we can scale the entire data set by a constant factor. Also, notice that each point in the date whose bounding box intersects the query sphere must be examined. Moreover, examining each data point can generate at most $O(log(n))$ computations. We will therefore only ask ourselves, how many point boxes have a non zero intersection with the query sphere.

To make this exercise simpler, consider a simpler case and a variation of kd-trees. First, all n data points are distributed uniformly and i.i.d. over $[0, 1]^d$. Also, the space partition splits every box exactly in the middle when the cut is made along the axes in a round robin order. Note that for such random data, our simple algorithm produces a very simple partition to the one kd-tree would have generated since the median point is approximately also in the middle of the box. (this holds as long as the number of points in a box is large enough)

Let's assume each box is split t times (the depth of the tree is t). That means that each axis was split in half t/d (assuming t/d is an integer) times. Therefore, the dimension of our box are $[2^{-t/d}, 2^{-t/d}, \ldots, 2^{-t/d}]$. What is the probability that a (randomly located) ball of radius r will hit this box?

This probability that the box is "hit" by a random query sphere is at most the volume of a sphere of radius $\sqrt{d}2^{-t/d-1} + r$ since if the query falls outside this ball it will not intersect the box. Assume, for now, that $\sqrt{d}2^{-t/d-1} \leq r$ and so the probability of this event is bounded by the volume of a ball of radius $2r$. The volume of such a ball is $\pi^{d/2}(2r)^d/(d/2)!$ (assuming, for convenience, that d is even). The dependance on the radius is therefore $O(r^d)$. Thus, we can expect to inspect only a $O(r^d)$ fraction of the boxes (and thus points). Since $r \leq 1$ this can be a significant speedup.

We saw that separating the space into boxes enabled us to search in it more efficiently. We, however, made some heavy assumptions. One such is that $\sqrt{d}2^{-t/d-1} \leq r$. This assumption fails when the dimension grows. To see this, consider that $t \approx \log(n)$ to give boxes of volume roughly $1/n$. Requiring that $\sqrt{d}2^{-\log(n)/d} \leq \sqrt{d}/2$ we get that $\log(n) \geq d$ or that n must be exponential in d, $n \geq 2^d$. Therefore, we only gain if the number of points in exponential in the dimension. This is a result of the so called "curse of dimensionality" which is discussed next.

4 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.

5 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".

6 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

[1] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18:509–517, September 1975.