princeton univ. F'13   cos 521: Advanced Algorithm Design

Lecture 11: High Dimensional Geometry, Curse of Dimensionality, Dimension Reduction

Lecturer: *Sanjeev Arora*   Scribe:

High-dimensional vectors are ubiquitous in algorithms and this lecture seeks to introduce some common properties of these vectors. We encounter the so-called *curse of dimensionality* which refers to the fact that algorithms are simply harder to design in high dimensions and often have a running time exponential in the dimension. We also show that it is possible to reduce the dimension of a dataset sometimes —and for some purposes.

Notation: For a vector $x \in \Re^n$ its ℓ_2-norm is $|x|_2 = (\sum_i x_i^2)^{1/2}$ and the ℓ_1-norm is $|x|_1 = \sum_i |x_i|$. For any two vectors x, y their Euclidean distance refers to $|x-y|_2$ and Manhattan distance refers to $|x-y|_1$.

We start with some useful generalizations of geometric objects to higher dimensional geometry:

- The *n-cube* in $\Re^n$: $\{(x_1...x_n) : 0 \leq x_i \leq 1\}$. To visualize this in $\Re^4$, think of yourself as looking at one of the faces, say $x_1 = 1$. This is a cube in $\Re^3$ and if you were able to look in the fourth dimension you would see a parallel cube at $x_1 = 0$. The visualization in $\Re^n$ is similar.

  The volume of the n-cube is 1.

- The unit *n-ball* in $\Re^n$: $B_n := \{(x_1...x_n) : \sum x_i^2 \leq 1\}$. Again, to visualize the ball in $\Re^4$, imagine you have sliced through it with a hyperplane, say $x_1 = 1/2$. This slice is a ball in $\Re^3$ of radius $\sqrt{1 - 1/2^2} = \sqrt{3}2$. Every parallel slice also gives a ball.

  The volume of B_n is $\frac{\pi^{n/2}}{(n/2)!}$ (assume n even if the previous expression bothers you), which is $\frac{1}{n^{\Theta(n)}}$.

0.1 An approximate way to think about B_n

A good approximation to picking a random point on the surface of B_n is by choosing random $x_i \in \{-1, 1\}$ independently for $i = 1..n$ and normalizing to get $\frac{1}{\sqrt{n}}(x_1, ..., x_n)$. A better approximation is to pick each coordinate as a gaussian with mean 0 and variance $1/n$. To get a point inside the ball, it is necessary to pick the distance from $\bar{0}$ randomly. Note that the distance is not distributed uniformly: the density at radius r is proportional to r^{n-1}.

Remark: An exact way to pick a random point on the surface of B^n is to choose x_i from the normal distribution for $i = 1..n$, and to normalize: $\frac{1}{l}(x_1, ..., x_n)$, where $l = (\sum_i x_i^2)^{1/2}$.

0.2 Funny facts

1. The volume of the unit n-ball tends to 0 as the dimension tends to ∞.

2. For any $c > 1$, a $(1 - \frac{1}{c})$ - fraction of the volume of the n-ball lies in a strip of width $O(\sqrt{\frac{c \log n}{n}})$. A strip of width a is B_n intersected with $\{(x_1, ..., x_n) | x_1 \in [-\frac{a}{2}, \frac{a}{2}]\}$.

3. If you pick 2 vectors on the surface of B_n independently, then with probability $> 1 - \frac{1}{n}$,

$$|\cos(\Theta_{x,y})| = O(\frac{\sqrt{\log n}}{n}),$$

where $\Theta_{x,y}$ is the angle between x and y. In other words, the 2 vectors are almost orthogonal w.h.p. To prove this, we use the following lemma:

Lemma 1
Suppose a is a unit vector in $\Re^n$. Let $x = (x_1, ..., x_n) \in R^n$ be chosen from the surface of B_n by choosing each coordinate at random from $\{1, -1\}$ and normalizing by factor $\frac{1}{\sqrt{n}}$. Denote by X the random variable $a \cdot x = \sum a_i x_i$. Then:

$$Pr(|X| > t) < e^{-nt^2}$$

Proof: We have:

$$\mu = E(X) = E(\sum a_i x_i) = 0$$

$$\sigma^2 = E[(\sum a_i x_i)^2] = E[\sum a_i a_j x_i x_j] = \sum a_i a_j E[x_i x_j] = \sum \frac{a_i^2}{n} = \frac{1}{n}.$$

Using the Chernoff bound, we see that,

$$Pr(|X| > t) < e^{-(\frac{t}{\sigma})^2} = e^{-nt^2}$$

□

Corollary 2
If two unit vectors x, y are chosen at random from $\Re^n$, then

$$Pr\left(|cos(\theta_{x,y})| > \sqrt{\frac{-\log \varepsilon}{n}}\right) < \varepsilon$$

Now, to get fact (3), put $\varepsilon = \frac{1}{n}$.

One consequence of the corollary is that if we pick say $\exp(0.01n)$ random vectors in n dimensions, with reasonable probability every pair of them has inner product no more than 0.1. By changing the constants, this inner product can be made an arbitrarily small constant.

1 Curse of dimensionality

Suppose we have a set of vectors in d dimensions and given another vector we wish determine its closest neighbor (in ℓ_2 norm) to it. Designing fast algorithms for this in the plane (i.e.,

$\Re^2$) uses the fact that in the plane there are only $O(1)$ distinct points whose pairwise distance is about $1 \pm \varepsilon$. In $\Re^d$ there can be $\exp(d)$ such points.

Thus most algorithms —nearest neighbor, minimum spanning tree, point location etc.— have a running time depending upon $\exp(d)$. This is the *curse of dimensionality* in algorithms. (The term was invented by R. Bellman, who as we saw earlier had a knack for giving memorable names.)

In machine learning and statistics sometimes the term refers to the fact that available data is too sparse in high dimensions; different take on the same underlying phenomenon.

I hereby coin a new term: *Blessing of dimensionality.* This refers to the fact that many phenomena become much clearer and easier to think about in high dimensions because one can use simple rules of thumb (e.g., Chernoff bounds, measure concentration) which don't hold in low dimensions.

2 Dimension Reduction

Now we describe a central result of high-dimensional geometry (at least when distances are measured in the ℓ_2 norm). Problem: Given n points $z^1, z^2, ..., z^n$ in $\Re^n$, we would like to find n points $u^1, u^2, ..., u^n$ in $\Re^m$ where m is of low dimension (compared to n) and the metric restricted to the points is almost preserved, namely:

$$\|z^i - z^j\|_2 \leq \|u^i - u^j\|_2 \leq (1+\varepsilon)\|z^j - z^j\|_2 \ \forall i, j. \tag{1}$$

The following main result is by Johnson & Lindenstrauss :

Theorem 3
In order to ensure (1), $m = O(\frac{\log n}{\varepsilon^2})$ suffices.

The following ideas do not work to prove this theorem (as we discussed in class): (a) take a random sample of m coordinates out of n. (b) Partition the n coordinates into m subsets of size about n/m and *add* up the values in each subset to get a new coordinate.

Proof: Choose m vectors $x^1, ..., x^m \in \Re^n$ at random by choosing each coordinate randomly from $\{\sqrt{\frac{1+\varepsilon}{m}}, -\sqrt{\frac{1+\varepsilon}{m}}\}$. Then consider the mapping from $\Re^n$ to $\Re^m$ given by

$$z \longrightarrow (z \cdot x^1, z \cdot x^2, \ldots, z \cdot x^m).$$

In other words $u^i = (z^i \cdot x^1, z^i \cdot x^2, ..., z^i \cdot x^m)$ for $i = 1, \ldots, k$. We want to show that with positive probability, $u^1, ..., u^k$ has the desired properties. This would mean that there exists at least one choice of $u^1, ..., u^k$ satisfying inequality 1. To show this, first we write the expression $\|u^i - u^j\|$ explicitly:

$$\|u^i - u^j\|^2 = \sum_{k=1}^{m} \left(\sum_{l=1}^{n} (z_l^i - z_l^j) x_l^k \right)^2.$$

Denote by z the vector $z^i - z^j$, and by u the vector $u^i - u^j$. So we get:

$$\|u\|^2 = \|u^i - u^j\|^2 = \sum_{k=1}^{m} \left(\sum_{l=1}^{n} z_l x_l^k \right)^2.$$

Let X_k be the random variable $(\sum_{l=1}^{n} z_l x_l^k)^2$. Its expectation is $\mu = \frac{1+\varepsilon}{m}\|z\|^2$ (can be seen similarly to the proof of lemma 1). Therefore, the expectation of $\|u\|^2$ is $(1+\varepsilon)\|z\|^2$. If we show that $\|u\|^2$ is concentrated enough around its mean, then it would prove the theorem. More formally, this is done in the following Chernoff bound lemma. □

Lemma 4
There exist constants $c_1 > 0$ and $c_2 > 0$ such that:

1. $Pr[\|u\|^2 > (1+\beta)\mu] < e^{-c_1\beta^2 m}$
2. $Pr[\|u\|^2 < (1-\beta)\mu] < e^{-c_2\beta^2 m}$

Therefore there is a constant c such that the probability of a "bad" case is bounded by:

$$Pr[(\|u\|^2 > (1+\beta)\mu) \vee (\|u\|^2 < (1-\beta)\mu)] < e^{-c\beta^2 m}$$

Now, we have $\binom{n}{2}$ random variables of the type $\|u_i - u_j\|^2$. Choose $\beta = \frac{\varepsilon}{2}$. Using the union bound, we get that the probability that any of these random variables is not within $(1 \pm \frac{\varepsilon}{2})$ of their expected value is bounded by

$$\binom{n}{2} e^{-c\frac{\varepsilon^2}{4}m}.$$

So if we choose $m > \frac{8(\log n + \log c)}{\varepsilon^2}$, we get that with positive probability, all the variables are close to their expectation within factor $(1 \pm \frac{\varepsilon}{2})$. This means that for all i,j:

$$(1-\frac{\varepsilon}{2})(1+\varepsilon)\|z^i - z^j\|^2 \le \|u^i - u^j\|^2 \le (1+\frac{\varepsilon}{2})(1+\varepsilon)\|z^i - z^j\|^2$$

Therefore,

$$\|z_i - z_j\|^2 \le \|u^i - u^j\|^2 \le (1+\varepsilon)^2\|z^i - z^j\|^2,$$

and taking square root:

$$\|z^i - z^j\| \le \|u^i - u^j\| \le (1+\varepsilon)\|z^i - z^j\|,$$

as required.

Question: The above dimension reduction preserves (approximately) ℓ_2-distances. Can we do dimension reduction that preserves ℓ_1 distance? This was an open problem for many years until Brinkman and Charikar showed in 2004 that no such dimension reduction is possible.

2.1 Locality preserving hashing

Suppose we wish to hash high-dimensional vectors so that nearby vectors tend to hash into the same bucket. To do this we can do a random projection into say the cube in 5 dimensions. We discretise the cube into smaller cubes of size ε. Then there are $1/\varepsilon^5$ smaller cubes; these can be the buckets.

This is simplistic; more complicated schemes have been constructed. Things get even more interesting when we are interested in ℓ_1-distance.