Chapter 11

Dimensionality Reduction

There are many applications where we deal with points lying in a very high dimensional Euclidean space. Storing n points in a d-dimensional space takes $O(nd)$ space, and even a linear time algorithm for processing such an input can be impractical. Many algorithms depend only on the pair-wise distance between these points. For example, the nearest-neighbour problem seeks to find the closest input point (in terms of Euclidean distance) to a query point. There is a trivial linear time algorithm to solve this problem, which just looks at every input point and computes its distance to the query point. Since the solution to this problem only depends on the distance of the query point to these n points, we ask the following question : can the points be mapped to a low dimensional space which preserves all pair-wise distances ? It is clear that d can be made at most n (just restrict to the affine space spanned by the n points), and in general one cannot do better.

Exercise 11.1 *Consider the 3 vertices of an equilateral triangle in the plane, each at distance 1 from the other two vertices. Show that it is not possible to map these 3 points to a line such that all pair-wise distances are 1.*

The above example shows that even in trivial settings, it is not possible to reduce the dimensionality of a set of points without distorting pair-wise distances. What if we are willing to incur a small amount of distortion in the pair-wise distances ? This is often an acceptable option because in many practical applications, the actual embedding of points in d dimensions is based on some rough estimates. Since the data already has some inherent noise, it should be acceptable to distort the pair-wise distances slightly.

Let us make these ideas more formal. We are given a set V of n points in a d-dimensional Euclidean space. Let f be a mapping of these points to a k-dimensional Euclidean space. We say that this mapping (or embedding) has distortion $\alpha > 1$ if

the following condition holds for every pair of distinct points $p_i, p_j \in V$:

$$\frac{1}{\alpha} \cdot ||p_i - p_j||^2 \leq ||f(p_i) - f(p_j)||^2 \leq || \leq \alpha \cdot ||p_i - p_j||^2.$$

Note that the definition deals with the square of the Euclidean distance – this turns out to be much easier to work with than Euclidean distances.

11.1 Random Projections and the Johnson Lindenstrauss Lemma

The Johnson Lindenstrauss Lemma states that for any small constant $\varepsilon > 0$ there is a linear map f into an Euclidean space of dimension $O(\log n/\varepsilon^2)$ such that the distortion is at most $(1 + \varepsilon)$. In fact the map f turns out to be quite simple. They show that f just needs to be the projection of the points on a random subspace of appropriate dimension.

As a toy example, we have two points p and q in the two dimensional plane, and suppose we try to project the points on a 1-dimensional line through the origin. We pick a suitable line L through the origin, and for a point p, define $f(p)$ as the projection of p on L. A moment's thought shows that in general such an embedding can have very high distortion. For example, suppose there are two points p and q such that the line joining them is (nearly) perpendicular to L. In this case, $f(p)$ and $f(q)$ will be very close, even though $||p - q||$ could be large. We can avoid such a situation by picking L to be a line along a *random* direction – we can easily do this by picking a random number θ in the range $[0, \pi)$ and then drawing L as the line which makes angle θ with one of the coordinate axes. Can we now compute the probability with which distance between $f(p)$ and $f(q)$ is (nearly) same as that between p and q ?

Exercise 11.2 *Assuming ε is a small positive constant, prove that the probability that $||f(p) - f(q)||^2$ is within $(1 \pm \varepsilon)||p - q||^2$ is $\theta(\sqrt{\varepsilon})$.*

The above exercise shows that this probability is quite small. How can we increase this probability to close to 1 ? One natural idea is to take several such lines, and think of projection along each line as giving one coordinate of $f(p)$. This does not make much sense when the points are already in 2-dimensions, but can lead to significant savings if the number of such lines is much less than the dimension d.

So suppose we have a set V of n points in d-dimensional Euclidean space. We pick k lines through the origin along random directions – call these lines $L_1, \ldots, L_k$. We now define $f(p)$ as a k-dimensional vector, where the i^{th} coordinate is the length

of the projection of p along L_i. The first non-trivial issue is how to pick a random direction in a d-dimensional Euclidean space. The trick is to pick a distribution whose density does not depend on a particular direction.

Recall that the normal distribution with 0 mean and variance 1, denoted by $N(0,1)$, has density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

We define a multi-dimensional normal distribution $X = (x_1, \ldots, x_d)$, where the variables are independent and each of them has distribution $N(0,1)$ (such a set of variables are called i.i.d. $N(0,1)$ random variables). The joint distribution of X is given by

$$\phi(X) = \frac{1}{(2\pi)^{d/2}} e^{-(x_1^2+\ldots+x_d^2)/2} = \frac{1}{(2\pi)^{d/2}} e^{-||X||^2/2}.$$

Note that this distribution just depends on the length of X and is independent of the direction of X. In other words, here is how we pick a line along a random direction: sample d i.i.d. N(0,1) random variables $x_1, \ldots, x_d$. Consider the line through the origin and the vector $(x_1, \ldots, x_d)$.

Having resolved the issue of how to pick a line along a uniformly random direction, we can now define what the embedding f does. Recall that f needs to project a point along k such lines. Thus, if p is a point with coordinates $\mathbf{p} = (p_1, \ldots, p_d)$, then $f(p) = R \cdot \mathbf{p}$, where R is a $k \times d$ matrix with entries being i.i.d. $N(0,1)$ random variables. Note that each row of R gives a line along a random direction, and each coordinate of $R \cdot \mathbf{p}$ is proportional to the projection of p along the corresponding line. To understand the properties of this embedding, we first need to understand some basic facts about the normal distribution. We use $N(\mu, \sigma^2)$ to denote a Normal distribution with mean μ and variance σ^2. Recall that the distribution of $N(\mu, \sigma^2)$ is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-u)^2/2\sigma^2}.$$

Exercise 11.3 *If X is an $N(0,1)$ random variable, and a is a real number, prove that aX is distributed as $N(0, a^2)$. Using this prove that if X, Y are two independent $N(0,1)$ random variables, and a and b are two real numbers, them $aX + bY$ has distribution $N(0, a^2 + b^2)$. Finally, use induction to prove that if $X_1, \ldots, X_d$ are d i.i.d. $N(0,1)$ random variables, then $a_1X_1 + \ldots + a_dX_d$ has distribution $N(0, ||a||^2)$, where a denotes the vector $(a_1, \ldots, a_d)$.*

The above exercise shows that the projection of a vector a along a uniformly random direction also has normal distribution. Using this fact, we can now calculate the expected length of $f(p)$ for a point p. Indeed, each coordinate of $f(p)$ is the

projection of p along a random direction (given by row i of R, denoted by R_i). Therefore, using the above exercise, we get

$$E[||f(p)||^2] = \sum_{i=1}^{k} E[(R_i \cdot p)^2] = k \cdot ||p||^2$$

We would like to normalize $f(p)$ such that $E[||f(P)||^2]$ is the same as that of $||p||^2$. Therefore, we redefine $f(p)$ as $\frac{1}{\sqrt{k}} \cdot R \cdot \mathbf{p}$. Now, the above calculations show that $E[||f(p)||^2] = ||p||^2$. We would now like to prove that $||f(p)||^2$ is closely concentrated around its mean with high probability. More precisely, we want to show that given an error parameter $\varepsilon > 0$ (which should be thought of as a small constant),

$$\Pr[||f(p)||^2 \notin (1 \pm \varepsilon)||p||^2] \leq 1/n^3.$$

Once we show this, we will be done. We can replace p by $p_i - p_j$ for all distinct pair of points p_i, p_j in V. Thus, for any distinct pair of points p_i, p_j, the distance between them gets distorted by more than $(1+\varepsilon)$-factor with probability at most $1/n^3$. But now, notice that there are at most n^2 such pairs we need to worry about. So, using union bound, the probability that there exists a pair p_i, p_j in V for which $||f(p_i) - f(p_j)||^2$ is not in the range $(1 \pm \varepsilon)||p_i - p_j||^2$ is at most

$$n^2 \cdot 1/n^3 = 1/n.$$

Thus, the embedding has distortion at most $(1+\varepsilon)$ with probability at least $1 - 1/n$ (in particular, this shows that there *exists* such an embedding).

Let us now prove that the length of $f(p)$ is tightly concentrated around its mean. First observe that $||f(p)||^2$ is the sum of k independent random variables, namely, $(R_1 \cdot p)^2, \ldots, (R_k \cdot p)^2$, each of which has mean $||p||^2$. Therefore, as it happens in Chernoff-Hoeffding bounds, we should expect the sum to be tightly concentrated around its mean. However, in the setting of Chernoff-Hoeffding bounds, each of these random variables have bounded range, whereas here, each of the variables $(R_i \cdot p)^2$ lie in an unbounded range. Still, we do not expect these random variables to deviate too much from their mean because $(R_i \cdot p)$ has normal distribution and we know that normal distribution decays very rapidly as we go away from the mean by a distance more than its variance. One hope would be to carry out the same steps as in the proof of the Chernoff-Hoeffding bound, and show that they go through in the case of sum of independent random variables with normal distribution.

Theorem 11.1 *Let $X_1, \ldots, X_k$ be i.i.d. $N(0, \sigma^2)$ random variables. Then, for any constant $\varepsilon < 1/2$,*

$$\Pr[(X_1^2 + \ldots + X_k^2)/k \geq (1+\varepsilon)\sigma^2] \leq e^{-\varepsilon^2 k/4},$$

and

$$\Pr[(X_1^2 + \ldots + X_k^2)/k \leq (1-\varepsilon)\sigma^2] \leq e^{-\varepsilon^2 k/4}.$$

It follows that if we pick k to be $12 \log n/\epsilon^2$, then the probability that $||f(p)-f(q)||$ differs from $||p-q||$ by more than $(1 \pm \varepsilon)$ factor is at most $1/n^3$. Since we are only concerned about at most n^2 such pairs, the embedding has distortion at most $1+\varepsilon$ with probability at least $1 - 1/n$. We now prove the above theorem.

Proof: We prove the first inequality, the second one is similar. Let Y denote $(X_1^2 + \cdots + X_k^2)/k$. Then $E[Y] = \sigma^2$. Therefore, as in the proof of Chernoff bounds,

$$\Pr[Y > (1+\varepsilon)\sigma^2] = \Pr[e^{sY} > e^{s(1+\varepsilon)\sigma^2}] \leq \frac{\mathbb{E}[e^{sY}]}{e^{s(1+\varepsilon)\sigma^2}}, \tag{11.1.1}$$

where $s > 0$ is a suitable parameter, and we have used Markov's inequality in last inequality. Now, the independence of the variables $X_1, \ldots, X_k$ implies that

$$\mathbb{E}[e^{sY}] = \prod_{i=1}^{k} \mathbb{E}[e^{sX_i^2/k}].$$

For a parameter α and $N(0, \sigma^2)$ normal random variable X,

$$E[e^{\alpha X^2}] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{\alpha x^2} \cdot e^{-x^2/2\sigma^2} dx = (1 - 2\alpha\sigma^2)^{-1/2}.$$

To evaluate the integral, you can use the fact that $\int_{-\infty}^{+\infty} e^{-x^2/2} = \sqrt{2\pi}$. Therefore, we can express the right hand side in (11.1.1) as

$$\frac{(1 - 2s\sigma^2/k)^{-k/2}}{e^{s(1+\varepsilon)\sigma^2}}.$$

Now, we would like to find the parameter s such that the above expression is minimized. By differentiating the above expression with respect to s and setting it to 0, we see that the right value of s is $\frac{k\varepsilon}{2\sigma^2(1+\varepsilon)}$. Substituting this in the above expression, we see that $\Pr[Y > (1+\varepsilon)\sigma^2]$ is at most $e^{k/2\ln(1+\varepsilon) - k\varepsilon/2}$. Using the fact that $\ln(1+\varepsilon) \leq \varepsilon - \varepsilon^2/2$, if $\varepsilon < 1/2$, we get the desired result. □