Lecture 6

Prof. Nick Harvey *University of British Columbia*

Dimensionality reduction is the process of mapping a high dimensional dataset to a lower dimensional space, while preserving much of the important structure. In statistics and machine learning, this often refers to the process of finding a few directions in which a high dimensional random vector has maximimum variance. Principal component analysis is a standard technique for that purpose.

In this lecture, we consider a different sort of dimensionality reduction where the goal is to preserve *pairwise distances* between the data points. We present a technique, known as the **random projection method**, for solving this problem. The analysis of this technique is known as the **Johnson-Lindenstrauss lemma**.

In the past few lectures, our main tool has been the Chernoff bound. In this lecture we will not directly use the Chernoff bound, but the main proof uses very similar ideas.

1 Dimensionality Reduction

Suppose we have m points $x_1, \ldots, x_m \in \mathbb{R}^n$. We would like to find m points $y_1, \ldots, y_m \in \mathbb{R}^d$, where $d \ll n$, such that

$$\begin{aligned} \|y_j\| &\approx \|x_j\| \qquad \forall j \\ \|y_j - y_{j'}\| &\approx \|x_j - x_{j'}\| \qquad \forall j, j'. \end{aligned}$$

Here the notation $\|x\|$ refers to the usual Euclidean norm of the vector x. We will show that this can be accomplished while taking d to be surprisingly small.

The main result is:

Theorem 1 *Let $x_1, \ldots, x_m \in \mathbb{R}^n$ be arbitrary. Pick any $\epsilon = (0, 1)$. Then for some $d = O(\log(m)/\epsilon^2)$ there exist points $y_1, \ldots, y_m \in \mathbb{R}^d$ such that*

$$\begin{aligned} (1-\epsilon)\|x_j\| &\le \|y_j\| \le (1+\epsilon)\|x_j\| \qquad \forall j \\ (1-\epsilon)\|x_j - x_{j'}\| &\le \|y_j - y_{j'}\| \le (1+\epsilon)\|x_j - x_{j'}\| \qquad \forall j, j'. \end{aligned} \tag{1}$$

Moreover, in polynomial time we can compute a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^d$ such that, defining $y_j := L(x_j)$, the inequalities in (1) are satisfied with probability at least $1 - 2/m$.

Whereas principal component analysis is only useful when the original data points $\{x_1, \ldots, x_m\}$ are inherently low dimensional, this theorem requires *absolutely no assumption* on the original data. Also, note that the final data points $\{y_1, \ldots, y_m\}$ have no dependence on n: the original data could live in an arbitrarily high dimension!

Let me now spoil the surprise: the linear transformation L in Theorem 1 is simply multiplication by a matrix whose entries are independent Gaussian random variables.

Formally, for $i = 1, \ldots, d$, let $r_i \in \mathbb{R}^n$ be a vector whose entries are independently drawn from $N(0, 1)$, the normal distribution with mean 0 and variance 1. Define a linear map $f : \mathbb{R}^n \to \mathbb{R}^d$ as follows: the ith coordinate of $f(v)$ is simply $r_i^T v$. We now prove a lemma about f, which easily leads to our desired linear transformation L.

Lemma 2 (Johnson-Lindenstrauss) *Fix any vector $v \in \mathbb{R}^n$ with $\|v\| = 1$. For some $d = O(\log(m)/\epsilon^2)$ we have*

$$\Pr[\ 1-\epsilon \ \leq\ \frac{\|f(v)\|}{\sqrt{d}}\ \leq\ 1+\epsilon\ ]\ \geq\ 1-2/m^3.$$

Given this lemma, our main theorem follows easily.

PROOF:[of Theorem 1] Define the linear map $L(v) := f(v)/\sqrt{d}$. Since f and L are both linear, the lemma implies that for *any* $v \in \mathbb{R}^n$, we have

$$\Pr[\ (1-\epsilon)\|v\| \ \leq\ \|L(v)\|\ \leq\ (1+\epsilon)\|v\|\ ]\ \geq\ 1-2/m^3.$$

Apply this result to all vectors $v = x_j$ and all vectors $v = x_j - x_{j'}$ (with $j \neq j'$). Since there are m^2 such vectors, a union bound shows that the probability of failing to satisfy (1) is at most $2/m$. □

1.1 Discussion

First of all, you have probably noticed that we've now jumped from the world of discrete probability to continuous probability. This is to make our lives easier. The same theorem would be true if we picked the coordinates of r_i to be uniform in $\{+1, -1\}$ rather than Gaussian. But the analysis of the $\{+1, -1\}$ case is trickier, and most proofs analyze that case by showing that its failure probability is not much worse than in the Gaussian case. So the Gaussian case is really the central problem.

Second of all, you might be wondering where the **random projection method** name comes from. Earlier versions of the Johnson-Lindenstrauss lemma used a slightly different function L. Specifically, they chose $L(v) = Rv$ where $R^T R$ is a *projection* onto a *uniformly random subspace* of dimension d. (Recall that an orthogonal projection matrix is any symmetric, positive semidefinite matrix whose eigenvalues are either 0 or 1.) One advantage of that setup is its symmetry: one can argue that the failure probability in Lemma 2 would be the same if one instead chose a *fixed* subspace of dimension d and a *random* unit vector v. The latter problem can be analyzed by choosing the subspace to be the most convenient one of all: the span of the first d vectors in the standard basis.

So how is our mapping L different? It is almost a projection, but not quite. When we choose R to be a matrix of independent Gaussians, it turns out that the range of $R^T R$ is indeed a uniformly random subspace, but its eigenvalues are not necessarily in $\{0, 1\}$. If we had insisted that the random vectors r_i that we choose were *orthonormal*, then we would have obtained a projection matrix. We could explicitly orthonormalize them by the Gram-Schmidt method, but fortunately that turns out to be unnecessary: the Johnson-Lindenstrauss lemma remains true, even if we ignore orthonormality of the r_i's.

Our definition of L turns out to be a bit more convenient in some algorithmic applications, because we avoid the awkward Gram-Schmidt step.

1.2 The proof

We need just one fact from probability theory: the sum of Gaussians is again Gaussian.

Fact 3 *Let X and Y be independent random variables where X has distribution $N(0, \sigma_X^2)$ and Y has distribution $N(0, \sigma_Y^2)$. Then $X + Y$ has distribution $N(0, \sigma_X^2 + \sigma_Y^2)$.*

Recall that if X has distribution $N(0, 1)$ then $\sigma \cdot X$ has distribution $N(0, \sigma^2)$. So by induction we get:

Fact 4 *Let $Y_1, \ldots, Y_m$ be independent random variables where Y_i has distribution $N(0,1)$. Then, for any scalars $\sigma_1, \ldots, \sigma_m$, the sum $\sum_i \sigma_i Y_i$ has distribution $N(0, \sum_i \sigma_i^2)$.*

The proof of Lemma 2 uses separate but similar arguments to analyze the upper and lower tails, as was the case with Chernoff bounds. We will prove only the upper tail. For convenience we square both sides, so our goal is to prove that

$$\Pr[\ \|f(v)\|^2 \ > \ (1+\epsilon)^2 d\ ] \ \leq \ 1/m^3. \tag{2}$$

Define $X_i = r_i^T v$, which is the ith coordinate of $f(v)$. By Fact 4, X_i has distribution $N(0, \sum_i v_i^2) = N(0,1)$.

We get the following expansion:

$$\|f(v)\|^2 \ = \ \sum_{i=1}^{d} (r_i^T v)^2 \ = \ \sum_{i=1}^{d} X_i^2.$$

Our goal is to prove an upper tail bound on $\|f(v)\|^2$. Fortunately, this random variable has a well-known distribution. We have just written $\|f(v)\|^2$ as the the sum-of-squares of d standard normal random variables, which is called the chi-squared distribution with parameter d. It is easy to see that

$$\mathrm{E}[\|f(v)\|^2] \ = \ \sum_{i=1}^{d} \mathrm{E}[X_i^2] \ = \ d,$$

since $\mathrm{E}[X_i^2]$ is the variance of X_i, which we have shown is 1.

So our desired inequality (2) is asking for a bound on the probability that a chi-squared random variable slightly exceeds its expectation. Since the chi-squared distribution is sum of independent random variables, we know by the the central limit theorem that it converges to a normal distribution as $d \to \infty$. We just need to quantify the rate of convergence, and this is where the Chernoff-style ideas arise.

Claim 5 *Let $Y = \sum_{i=1}^{d} X_i^2$ have the chi-squared distribution with parameter d. Set $\alpha = d(1+\epsilon)^2$. Then $\Pr[Y > \alpha] \leq \exp(-(3/4)d\epsilon^2)$.*

Applying Claim 5 to $Y = \|f(v)\|^2$ with $d = 4\ln(m)/\epsilon^2$ completes the proof of (2).

Proof: Pick any parameter $t \in [0, 1/2)$. Just like with Chernoff bounds, we write

$$\Pr[Y > \alpha] \ = \ \Pr[e^{tY} > e^{t\alpha}] \ \leq \ e^{-t\alpha}\mathrm{E}[e^{tY}]. \tag{3}$$

As with Chernoff bounds, the bulk of the effort is in analyzing $\mathrm{E}[e^{tY}]$, but we can use independence to write

$$\mathrm{E}[e^{tY}] \ = \ \mathrm{E}\left[\exp\left(t\sum_{i=1}^{d} X_i^2\right)\right] \ = \ \prod_{i=1}^{d} \mathrm{E}\big[\exp(tX_i^2)\big]. \tag{4}$$

Expanding the expectation we get

$$\mathrm{E}\big[\exp(tX_i^2)\big] \ = \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ty^2)\exp(-y^2/2)\, dy \ = \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\big(-y^2(\tfrac{1}{2} - t)\big)\, dy.$$

If that $\frac{1}{2} - t$ factor were simply a $1/2$ then we could evaluate that integral using the fact that $e^{-z^2}/\sqrt{2\pi}$ is the PDF of a standard normal random variable, so it integrates to 1. We can accomplish that with a change of variables. Using $z = y\sqrt{1-2t}$, we get

$$\begin{aligned} \mathrm{E}\big[\exp(tX_i^2)\big] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(- \big(y\sqrt{1-2t}\big)^2/2\Big)\, dy \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-2t}} \int_{-\infty}^{\infty} \exp(-z^2/2)\, dz \\ &= \frac{1}{\sqrt{1-2t}}. \end{aligned}$$

Combining this with (3) and (4) we get

$$\Pr[Y > \alpha] \;\le\; e^{-t\alpha}(1-2t)^{-d/2}.$$

The last step is to plug in an appropriate choice of t. We set $t = (1 - d/\alpha)/2$, giving

$$\Pr[Y > \alpha] \;\le\; e^{-t\alpha}(1-2t)^{-d/2} \;=\; e^{(d-\alpha)/2}(d/\alpha)^{-d/2}.$$

Plugging in $\alpha = d(1+\epsilon)^2$, this becomes

$$\exp\Big(\frac{d}{2}\Big(1-(1+\epsilon)^2\Big) - \frac{d}{2}\ln\Big(\frac{1}{(1+\epsilon)^2}\Big)\Big) \;=\; \exp\Big(-d\big(\epsilon + \epsilon^2/2 - \ln(1+\epsilon)\big)\Big).$$

Using our usual techniques from Notes on Convexity Inequalities, one can show that $\ln(1+x) \le x - x^2/4$ for $x \in [0,1]$. So this shows that

$$\Pr[Y > \alpha] \;\le\; \exp\Big(-d\big(\epsilon + \epsilon^2/2 - (\epsilon - \epsilon^2/4)\big)\Big) \;\le\; \exp\big(-(3/4)d\epsilon^2\big).$$

□

2 Remarks

It turns out that the Johnson-Lindenstrauss lemma is almost optimal. Alon proved the following lower bound.

Theorem 6 (Alon) *Let $y_1, \ldots, y_{n+1} \in \mathbb{R}^d$ be vectors such that $1 \le \|y_i - y_j\| \le 1+\epsilon$ for all $i \ne j$. Then $d = \Omega(\frac{\log(n)}{\epsilon^2 \log(1/\epsilon)})$.*

To understand this theorem, let $x_1, \ldots, x_{n+1} \in \mathbb{R}^n$ be the vertices of a simplex, i.e., $\|x_i - x_j\| = 1$ for all $i \ne j$. Then, if we map the x_i's to points in $\mathbb{R}^d$ while preserving distances up to a factor $1+\epsilon$, then the dimension d must be at least $\Omega(\frac{\log(n)}{\epsilon^2 \log(1/\epsilon)})$, which is nearly what the Johnson-Lindenstrauss lemma would give. The only discrepancy is the small factor of $\log(1/\epsilon)$.

The Johnson-Lindenstrauss lemma very strongly depends on properties of the Euclidean norm. For other norms, this remarkable dimensionality reduction is not necessarily possible. For example, for the L_1 norm $\|x\|_1 := \sum_i |x_i|$, it is known that any map into $\mathbb{R}^d$ that preserves pairwise distances between n points up to a factor c must have $d = \Omega(n^{1/c^2})$. (See Brinkman-Charikar 2003 and Lee-Naor 2004.)