0368-3248-01-Algorithms in Data Mining **Fall 2013**

Lecture 5: Random-projection

Lecturer: Edo Liberty

Warning: *This note may contain typos and other inaccuracies which are usually discussed during class. Please do not cite this note as a reliable source. If you find mistakes, please inform me.*

We will give a simple proof of the following, rather amazing, fact. Every set of n points in a Euclidian space (say in dimension d) can be embedded into the Euclidian space of dimension $k = O(\log(n)/\varepsilon^2)$ such that all pairwise distances are preserved up distortion $1 \pm \varepsilon$. We will prove the construction of [1] which is simpler than the one in [2].

Random projection

We will argue that a certain distribution over the choice of a matrix $\mathbb{R} \in \mathbb{R}^{k \times d}$ gives that:

$$\forall x \in \mathbb{S}^{d-1} \quad \Pr\left[\left| ||\frac{1}{\sqrt{k}}Rx|| - 1 \right| > \varepsilon\right] \le \frac{1}{n^2} \tag{1}$$

Before we pick this distribution and show that Equation 1 holds for it, let us first see that this gives the opening statement.

Consider a set of n points $x_1, \ldots, x_n$ in Euclidian space $\mathbb{R}^d$. Embedding these points into a lower dimension while preserving all distances between them up to distortion $1 \pm \varepsilon$ means approximately preserving the norms of all $\binom{n}{2}$ vectors $x_i - x_j$. Assuming Equation 1 holds and using the union bound, this property will fail to hold for at least one $x_i - x_j$ pair with probability at most $\binom{n}{2}\frac{1}{n^2} \le 1/2$. Which means that all $\binom{n}{2}$ point distances are preserved up to distortion ε with probability at least $1/2$.

1 Matrices with normally distributed independent entries

We consider the distribution of matrices R such that each $R(i,j)$ is drawn independently from a normal distribution with mean zero and variance 1, $R(i,j) \sim \mathcal{N}(0,1)$. We show that for this distribution Equation 1 holds for some $k \in O(\log(n)/\varepsilon^2)$.

First consider the random variable $z = \sum_{j=1}^{d} r(j)x(j)$ where $r(j) \sim \mathcal{N}(0,1)$. To understand how the variable z distributes we recall the two-stability of the normal distribution. Namely, if $z_3 = z_2 + z_1$ and $z_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ then,

$$z_3 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}).$$

In our case, $r(i)x(i) \sim \mathcal{N}(0, x_i)$ and therefore, $z = \sum_{i=1}^{d} r(i)x(i) \sim \mathcal{N}(0, \sqrt{\sum_{i=1}^{d} x_i^2}) \sim \mathcal{N}(0,1)$. Now, note that each element in the vector Rx distributes exactly like z. Defining k identical copies of z, $z_1, \ldots, z_k$, We get that $||\frac{1}{\sqrt{k}}Rx||$ distributes exactly like $\sqrt{\frac{1}{k}\sum_{i=1}^{k} z_i^2}$. Thus, proving Equation 1 reduces to showing that:

$$\Pr\left[\left|\sqrt{\frac{1}{k}\sum_{i=1}^{k} z_i^2} - 1\right| > \varepsilon\right] \le \frac{1}{n^2} \tag{2}$$

for a set of independent normal random variables $z_1, \dots, z_k \sim \mathcal{N}(0,1)$. It is sufficient to demanding that $\Pr[\sum_{i=1}^k z_i^2 \geq k(1+\varepsilon)^2]$ and $\Pr[\sum_{i=1}^k z_i^2 \leq k(1-\varepsilon)^2]$ are both smaller than $1/2n^2$. We start with bounding the probability that $\sum_{i=1}^k z_i^2 \geq k(1+\varepsilon)$ (this is okay because $k(1+\varepsilon) < k(1+\varepsilon)^2$).

$$\Pr[\sum z_i^2 \geq k(1+\varepsilon)] = \Pr[e^{\lambda \sum z_i^2} \leq e^{\lambda k(1+\varepsilon)}] \leq (\mathbb{E}[e^{\lambda z^2}])^k / e^{\lambda k(1+\varepsilon)}$$

Since $z \sim \mathcal{N}(0,1)$ we can compute $\mathbb{E}[e^{\lambda z^2}]$ exactly:

$$\mathbb{E}[e^{\lambda z^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda t^2} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(t\sqrt{1-2\lambda})^2}{2}} dt = e^{\frac{1}{2}\log(1-2\lambda)}$$

The final step is by substituting $t' = t\sqrt{1-2\lambda}$ and recalling that $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t'^2}{2}} dt' = 1$. Finally, using the fact that $log(\frac{1}{1-2\lambda}) \leq 2\lambda + 4\lambda^2$ for $\lambda \in [0, 1/4]$ we have:

$$\mathbb{E}[e^{\lambda z^2}] = \frac{1}{\sqrt{1-2\lambda}} = e^{\frac{1}{2}\log(\frac{1}{1-2\lambda})} \leq e^{\lambda + 2\lambda^2}$$

Substituting this into the equation above we have that:

$$\Pr \leq e^{k(\lambda+2\lambda^2) - k\lambda(1+\varepsilon)} = e^{2k\lambda^2 - k\lambda\varepsilon} = e^{-k\varepsilon^2/8}$$

for $\lambda \leftarrow \varepsilon/4$. Finally, our condition that

$$\Pr[\sum_{i=1}^{k} z_i^2 \geq k(1+\varepsilon)] \leq e^{-k\varepsilon^2/8} \leq 1/2n^2$$

is achieved by $k = c\log(n)/\varepsilon^2$. Calculating for $\Pr[\sum_{i=1}^k z_i^2 \leq k(1-\varepsilon)]$ in the same manner shows that $k = c\log(n)/\varepsilon^2$ is also sufficient for this case. This completes the proof.

References

[1] S. DasGupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report, UC Berkeley*, 99-006, 1999.

[2] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.