Data mining: lecture 6

Edo liberty

We will give a simple proof of the following, rather amazing, fact. Every set of n points in Euclidian space (say in dimension $\mathbb{R}^d$) can be embedded into the Euclidian space of dimension $O(\log(n)/\varepsilon^2)$ such that all pairwise distances are preserved up distortion $1 \pm \varepsilon$.

Random projection

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

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

Before we show this distribution and show that Equation 1 holds for it, let us first see that this will 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 $\binom{n}{2}\frac{1}{n^2} \leq 1/2$. Which means that all $\binom{n}{2}$ point distances are preserved up to distortion ε with probability at least 1/2.

1 I.i.d gaussian distribution

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 will show that for this distribution Equation 1 holds for some $k \in O(\log(n)/\varepsilon^2)$.

First consider the random variable $z = \sum_{i=1}^{d} r(i)x(i)$ where $r(i) \sim \mathcal{N}(0,1)$. 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 $z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ then, $z_3 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. In our case, $r(i)x(i) \sim \mathcal{N}(0, x_i^2)$ and therefore, $z = \sum_{i=1}^{d} r(i)x(i) \sim \mathcal{N}(0, \sum_{i=1}^{d} x_i^2) = \mathcal{N}(0, ||x||^2) = ||x|| \cdot \mathcal{N}(0,1)$. 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:

$$||\frac{1}{\sqrt{k}}Rx|| \sim \sqrt{\frac{1}{k}\sum_{i=1}^{k} z_i^2} =\sim ||x||\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2}$$

where $y_i \sim \mathcal{N}(0,1)$. Thus, proving Equation 1 reduces to showing that:

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

It is now straight forward to show since the sum of k squared normal variables is a very known distribution called chi-square with k degrees of freedom. (χ_k^2). More accurately, it is defined by $\chi_k^2 = \sum_{i=1}^{k} y_i^2$ where $y_i \sim \mathcal{N}(0,1)$ which is exactly what we have. Since χ_k^2 is a sum of independent random variables, due to the central limit theorem, χ_k^2 converges to a normally distributed quantity as k grows. We will use here a slightly different property: $\sqrt{\chi_k^2} \sim_{k\to\infty} \mathcal{N}(\sqrt{k}, 1/2)$. Somewhat sloppily, we will assume that k is large enough so that assuming $\sqrt{\chi_k^2} \sim \mathcal{N}(\sqrt{k-1/2}, 1/2) \approx \mathcal{N}(\sqrt{k}, 1/2)$ is harmless. I that case, $\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2} \sim \mathcal{N}(1, \frac{1}{2k})$ and $\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2} - 1 \sim \mathcal{N}(0, \frac{1}{2k})$. Thus, we only need to show that for a random variable $Z \sim \sqrt{2k}\left[\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2} - 1\right] \sim$ $, \sqrt{2k}\mathcal{N}(0, \frac{1}{2k}) \sim \mathcal{N}(0,1)$ it holds that

$$\Pr\left[|Z| > \varepsilon\sqrt{2k}\right] \leq \frac{1}{n^2} \tag{3}$$

We now use a bound on the error function: $\int_{t=\varepsilon\sqrt{2k}}^{\infty} \frac{1}{\sqrt{2\Pi}} e^{-t^2/2} dt = erf(\varepsilon\sqrt{2k}) \leq e^{-\varepsilon^2 k}$. Since $\Pr[Z > \varepsilon\sqrt{2k}] = \Pr[Z < -\varepsilon\sqrt{2k}]$we demand that $e^{-\varepsilon^2 k} \leq \frac{1}{2n^2}$. This yields the bound $k \geq \frac{2\log(n)+1}{\varepsilon^2}$.