CS181 Assignment 3 - Clustering and Parameter Estimation

Lexi Ross & Ye Zhao

2013/03/09

1 High Dimensional Clustering

(a) Given that $\rho = P(\max_m |x_m - y_m| \leq \epsilon)$, the probability that all the M dimensions of $\mathbf{xy}$ are between $-\epsilon$ and ϵ, we can find out ρ by finding the probability p_m of havinng each individual dimension of $\mathbf{xy}$, ie $p_m = P(\epsilon \leq x_m y_m \leq \epsilon)$. Since y_m is a uniform distribution on [0,1], we have

$$P(-\epsilon \leq x_m - y_m \leq \epsilon) = 2\epsilon \tag{1}$$

From the independence of each component, we have

$$\rho = \prod_{m=1}^{M} p_m = (2\epsilon)^M \tag{2}$$

(b) In this case since $\mathbf{x}$ is some arbitrary point in the hypercube, it is possible that the at least one of the components of $\mathbf{x}$ is within ϵ far away from the surface of the cube. Let the dimension that has x_m near to the bound, ie $x_m < \epsilon or x_m > (1\epsilon)$, then we know that the probability of $|y_m x_m| \leq \epsilon$ will be strictly less than 2ϵ since at least one side of the point is being truncated. Hence the total probabilty will be less than that of ρ.

(c) The Euclidean distance is given by

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{m=1}^{M} (x_m - y_m)^2} \tag{3}$$

Let x_m and y_m be the component that maximizes $|x_m y_m|$, hence we have

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{(x_m - y_m)^2 + \sum_{m \neq m^*, m \in M}^{M} (x_m - y_m)^2} > \sqrt{(x_m - y_m)^2} = |x_m - y_m| \tag{4}$$

$$||\mathbf{x} - \mathbf{y}|| > \max_m |x_m - y_m| \tag{5}$$

where the inequality comes from the fact that the summed square must always be bigger than or equal to zero.

Considering the geometry, $||\mathbf{x} - \mathbf{y}|| < \epsilon$ represents a hypersphere of radius ϵ centered around the point $\mathbf{x}$ and $\max_m |x_m - y_m| < \epsilon$ represents a hypercube of side 2ϵ centered around $\mathbf{x}$. In this case the probability of $\mathbf{y}$ falling into these two different regions is just equal to the M-dimensional volume of the hypercube and hypersphere respectively. We know that the hypersphere of radius ϵ can always be circumscribed within the hypercube of side 2ϵ. Hence we have

$$P(||\mathbf{x} - \mathbf{y}|| < \epsilon) < P(\max_m |x_m - y_m| < \epsilon) \leq \rho \tag{6}$$

(d) Let p be the probability that the nearest neighbor of a point $\mathbf{x}$ to be not within a radius of ϵ, ie

$$p = 1 - P(||\mathbf{x} - \mathbf{y}|| \leq \epsilon \leq 1 - \rho \tag{7}$$

Since each individual point is independent of each other, hence the probability that none of the N points will have its nearest neightbour within a radius of ϵ is p^N. Therefore, the complement of it, which is the probability that at least one of the N points will have its nearest neighbor within a radius ϵ of it will just be $1 - p^N$ which gives

$$\begin{aligned} 1 - p^N &\leq 1 - \delta && (8) \\ (1 - \rho)^N &\geq \delta && (9) \\ & && (10) \end{aligned}$$

$$\begin{aligned} \implies 1 - \delta \leq && 1 - (1 - \rho)^N && (11) \\ \delta \geq && (1 - \rho)^N && (12) \\ \frac{\log \delta}{\log(1 - \rho)} \geq && N && (13) \end{aligned}$$

$$N \geq \frac{\log \delta}{\log(1 - 2^M \epsilon^M)} \tag{14}$$

2 ML vs MAP vs FB