CS181 Assignment 3 - Clustering and Parameter Estimation

Lexi Ross & Ye Zhao

2013/03/08

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 $\mathbf{x} - \mathbf{y}$ are between $-\epsilon$ and ϵ, we can find out ρ by finding the probability p_m of havinng each individual dimension of $\mathbf{x} - \mathbf{y}$, 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

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

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

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