

Introduction to Machine Learning, Fall 2014 - Exercise session II

Rodion “rodde” Efremov

November 2, 2014

Problem 3 (3 points)

Proximity is typically defined between a pair of objects.

- (a) Give two ways in which you might define the ‘proximity’ among a set of (more than two) objects (i.e. a single measure of how similar an arbitrary number of items are all to one another)
- (b) How might you define the distance between two sets of points in Euclidian space?
- (c) How might you define the proximity between two sets of data objects? (Make no assumptions about the data objects, except that a proximity measure is defined between any pair of objects.)

(a)

If $P(\mathbf{x}, \mathbf{y})$ is a “normal” function giving the proximity of data objects \mathbf{x} and \mathbf{y} , and A is an arbitrary set of data objects (which can be indexed like A_1, A_2, \dots, A_n), I would try

$$P(A) = \left(\sum_{1 \leq i < j \leq n} P(A_i, A_j) \right) \binom{n}{2}^{-1},$$

or namely the average proximity over all pairs of data objects from A . For another one, I would choose a small $\varepsilon > 0$, and define $P(A)$ as

$$\prod_{1 \leq i < j \leq n} \left(\max(P(A_i, A_j), \varepsilon) \right)^{-\binom{n}{2}},$$

or namely the geometric mean over all pairs of data objects from A . The ε is chosen to be positive so that a single proximity value of 0 does not dominate the entire proximity of a set and bring it down to 0. (Above, it is assumed that $P(\mathbf{x}, \mathbf{y}) \geq 0$ for all data objects \mathbf{x}, \mathbf{y} .)