Mining Massive Datasets:

Week 4: Distance Measures

Ian Quah (itq)

June 14, 2017

Note: The notes here are sparse because I know most of the material discussed so definitely go through it on your own

Topic 1

Distance Measures

Distances Measures

1. Generalized LSH is based on notion of distance between points
2. Note: Jaccard Similarity is not a true distance, 1 - Jaccard is
3. **Axioms of distance functions**

   D is distance function on x,y: D(x,y) if

   (a) D(x, y) $\geq$ 0
   (b) D(x, y) = 0 iff x == y
   (c) D(x,y) = d(y, x)
   (d) d(x, y) $\leq$ d(x, z) + d(z, y) The triangle inequality
4. **Euclidean**

   (a) has some number of real-valued dimensions and dense points
   (b) There is a notion of "average" of two points - useful for thinking about clusters
   (c) E.g: L_1, L_2, L_∞
5. **Non-Euclidean**

   (a) Based on properties of points, not location in a space
   (b) If distance measure is not Euclidean, automatically non-Euclidean
   (c) E.g: Jaccard, Cosine, Edit, Hamming Distance

Topic 2

Nearest Neighbor Learning

Instance based learning

1. Run classification again for each new example (unlike other algorithms where we estimate some parameters θ which we use to speed up classification on new params)
2. **K-nearest Neighbors**

   (a) Works for regression and classification
   (b) Keep whole training dataset
   (c) New query, q comes in
   (d) Find closest examples X
   (e) Predict y_q
3. Real world example: Collaborative filtering
4. **KNN for large datasets**

   (a) **Given**: set of point P, s.t each point $\in \mathbb{R}^d$

Topic 2 continued on next page...

(b) **Goal:** Given a query point q

(c) **NN**: Find nearest neighbor p of q in P

(d) **Range search** Find one/ all points in P within distance r from q

(e) Two types of queries when dealing with NN
1) Find K nearest to query point q
2) Find all points within some distance r to q
O(n), but with locality sensitive hashing, can be near constant