Information Retrieval

Vector classification, SVM, k means

1. Consider the vectors a = $(0.5\ 1.5)^T$, ~x = $(2\ 2)^T$, b =$(4\ 4)^T$, and c = $(8\ 6)^T$. which of the three vectors a, b, and c is (i) most similar to x according to dot product similarity, (ii) most similar to x according to cosine similarity, (iii)closest to x according to Euclidean distance?
2. Show that the decision boundaries in Rocchio classification are, as in kNN, given by the Voronoi tessellation.
3. Consider the data points {(3,1),(3,-1),(6,1),(6,-1)} labelled positively and the points {(1,0),(0,1),(0,-1),(-1,0)} labelled negatively. Find the Support Vectors for a linear SVM.
4. Apply k means on the following {2,4,8,13,14,18} with k=2 and initial cluster centers 2, 13.
5. Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points *a=(x1, y1)* and *b=(x2, y2)* is defined as: $\rho(a, b) = |x2 - x1| + |y2 - y1|$. Take the initial cluster centers to be (2, 10), (5, 8) and (1, 2)

Solutions

1. (i)c (dot products: 0.05, 0.16, 0.28) (ii)b (cosines: 0.9805, 1.0, 0.9899) (iii)a (distances: 0.1118, 0.2828, 0.7211)
2. Rocchio can be viewed as kNN classification with the centroids being the training points
3. (1,0),(3,1) and (3,-1)
4. {2,4,8},{13,14,18}, 3 iterations required
5. Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

Cluster 1	Cluster 2	Cluster 3
(2, 10)	(8, 4)	(2, 5)
	(5, 8)	(1, 2)
	(7, 5)	
	(6, 4)	
	(4, 9)	

Two more iterations are required. After the 2nd, the clusters are {A1,A8}, {A3,A4,A5,A6}, {A2,A7}, with centers (3,9.5), (6.5,5,25),(1.5,3.5)
After the third, the clusters are {A1,A4,A8},{A3,A5,A6},{A2,A7}
with centers (3.66,9), (7,4.33) and (1.5,3.5)