Yuan Wang
440 Data Mining
Prof. Jiebo Luo
Homework 5

9.1

SVM:

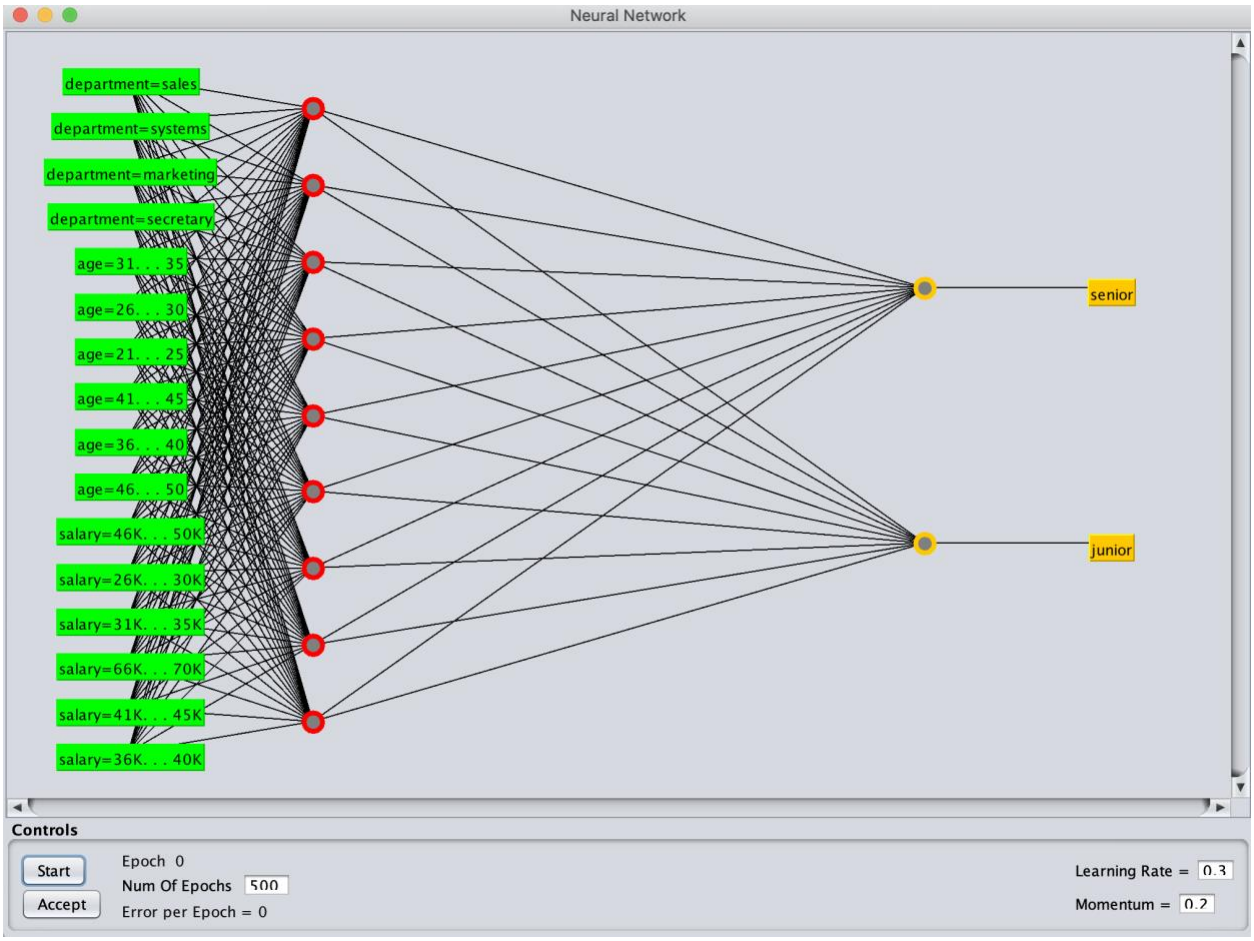
9.1_SVM.txt

Kstar:

9.1_Kstar.txt

MLNN：

9.1_MLNN.txt



10.2

Clustering analysis is finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters, and grouping dissimilar data objects in other clusters.

1. Partitioning methods
   Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where ci is the centroid or medoid of cluster Ci)

   $$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

   Example: k-means, k-medoids (PAM), CLARA
2. Hierarchical methods
   A hierarchical method creates a hierarchical decomposition of the given set of data objects. It has two approaches: agglomerative (bottom-up) and divisive (top-down). The bottom-up approach starts with each object forming a separate group. It then merges the objects close to one another, util all of the groups are merged into one, or until a termination condition holds. The top-down approach starts with all objects in the same

cluster. In each successive iteration, a cluster is split up into small clusters, until each object is in one cluster or until a termination condition holds.
Example: Diana, Agnes, BRICH, CAMELEON

3. Density-based methods
   Clustering based on density (local cluster criterion) such as density-connected points. It is to continue growing cluster as long as the density in its "neighborhood" exceeds some threshold. For each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. This method is used to filter out noise and discover clusters of arbitrary shape.
   Example: DBSACN, OPTICS, DenClue, GraphCut
4. Grid-based methods
   This kind of methods quantize the object space into a finite number of cells that form a grid structure. All the clustering operations are performed on the grid structure. The main advantage is its fast-processing time, which typically independent of the number of data objects and depend only on the number of cells in each dimension in the quantized space.
   Example: STING, WaveCluster, CLIQUE
5. Model-based methods
   This method hypothesizes a model for each of the clusters and find the best fit of the data to the given model. It locates cthe clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking "noise" or outliers into account and thus yielding robust clustering methods.
   Example: EM, SOM, COBWEB

10.4

Both algorithms put the first centroid randomly, but K-means++ choose the next centroid based on a probability that depends on the distance to the first point, the further apart the point is the more probable it is. Then repeat the process, the probability of each point is based on its distance to the closest centroid to that point. Therefore, the centroids that K-means++ picks is more reasonable and is more closed to optimum. This introduces an overhead in the initialization of the algorithm, but it reduces the probability of a bad initialization leading to a bad clustering result. Furthermore, although K-means cost more time in the beginning to pick up more reasonable centroids, the convergence tends to be faster and better.

10.6

(a)

K-means is more efficient but is sensitive to outliers or noise.

K-medoids is less sensitive to outliers or noise but its cost is expensive since it needs to calculate each sample points to get the most center point.

(b)

Partition:

Its quality is good in general, can undo what was done (moving objects around clusters); it requires the number of clusters to be known. It's good for spherical shaped cluster

Hierarchical:

The quality could be poor, cannot undo what was done, does not require the number of clusters to be known; more efficient and parallel, may find only arbitrary shaped clusters.

11.2

The probability of Ada and Bob choose i common products is: ($3 \leq i \leq 10$)

$Pr_{(A,B)}(i) = \frac{\binom{997}{7}\binom{7}{i-3}\binom{990}{10-i}}{\binom{997}{7}\binom{997}{7}} = \frac{\binom{7}{i-3}\binom{990}{10-i}}{\binom{997}{7}}$

The probability of Ada and Bob choose j common products is: ($1 \leq j \leq 10$)

$Pr_{(A,C)}(j) = \frac{\binom{997}{7}\binom{10}{j}\binom{990}{10-j}}{\binom{997}{7}\binom{1000}{10}} = \frac{\binom{10}{j}\binom{990}{10-j}}{\binom{1000}{10}}$

i	3	4	5	6	7	8	9	10
dist(Ada, Bob)	$\sqrt{14}$	$\sqrt{12}$	$\sqrt{10}$	$\sqrt{8}$	$\sqrt{6}$	2	1	0
J(Ada, Bob)	3/17	4/16	5/15	6/14	7/13	8/12	9/11	10/10
Pr(A,B)	0.95	6.8×10^{-3}	4.1×10^{-5}	2.1×10^{-7}	8.5×10^{-10}	2.6×10^{-12}	5.2×10^{-15}	5.3×10^{-18}

j	1	2	3	4	5
dist(Ada, Cathy)	$\sqrt{18}$	$\sqrt{16}$	$\sqrt{14}$	$\sqrt{12}$	$\sqrt{10}$
J(Ada, Cathy)	1/19	2/18	3/17	4/16	5/15
Pr(A,C)	9.2×10^{-13}	8.4×10^{-5}	6.9×10^{-7}	4.9×10^{-9}	3×10^{-11}
j	6	7	8	9	10
dist(Ada, Cathy)	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{4}$	$\sqrt{1}$	0
J(Ada, Cathy)	6/14	7/13	8/12	9/11	10/10
Pr(A,C)	1.5×10^{-13}	6.1×10^{-16}	1.9×10^{-18}	3.8×10^{-21}	3.8×10^{-24}

The probability of dist(Ada, Bob) > dist(Ada, Cathy) is when i ($3 \leq i \leq 10$) and j = i + 1 ($4 \leq j \leq 10$),

Thus, sum of Pr(A,B)*Pr(A,C) = 4.7×10^{-9}

The probability that J(Ada, Bob) > J(Ada, Cathy) is when i ($3 \leq i \leq 10$) and j = i - 1 ($4 \leq j \leq 9$),

Thus, sum of Pr(A,B)*Pr(A,C) = 8.9×10^{-3}

1. The probability is very small when two customers purchase same products among a large number of products
2. The larger difference, the larger Euclidean distance, but smaller Jaccard similarity
3. When the data have high dimensions, Jaccard similarity is easy to use because its range is [0, 1] but the range of Euclidean distance is [0, infinite]