**Assignment 5**

**DSC 440 Data Mining**

**Nov. 11, 2019**

**Meiying Chen**

**10.2**

(a)For A2: d(A1A2) = 5, d(B1A2) = 3\*sqrt(2), d(C1A2) = sqrt(10)

Because d(C1A2) is the smallest, A2 belongs to the third cluster.

Continue doing this for A3, B2, B3 and C2, we can get the final cluster after the first iteration:

Cluster 1: A1

Cluster 2: B1, A3, B2, B3, C2

Cluster 3: C1, A2

(b)Take mean value of each cluster, calculate the distance for every point, and reassign them to the three clusters, we can get the final cluster result:

Cluster 1: A1, C2, B1

Cluster 2: A3, B2, B3

Cluster 3: A2, C1

**10.4**

The initialization process of k-means++ guarantees datapoints with higher distances have larger probabilities to be selected as cluster center.

This speed up the convergence of k-means because it prevents the situation that selected cluster centers are too close to separate the dataset so that it needs a lot of iterations to make it right.

This also improves the final quality of clustering. Because in k-means, the choice of initial cluster center greatly influences the clustering result, if the cluster is not diverse enough, the number of datapoints in each cluster center will not be even. K-means++ maintains the diversity of cluster center, so the result improves in datapoints distribution among different clusters.

**10.6**

(a) Strength: k-means is faster as the calculation of mean value is simpler.

Weakness: mean value is less robust than medoids in terms of noise and outliners.

(b) k-means and k-medoids algorithms are partitioning-based cluster methods.

Strength: By changing cluster centers and repartitioning data points, k-means and k-medoids can roll back to former cluster status. This allows the methods to be more flexible and can adjust previous mistakes. But hierarchical clustering method cannot make this kind of adjustment.

Weakness: Partitioning-based cluster methods need to know the number of the clusters before performing the algorithm, and the cluster number is hard to decide when the dataset is complex. Hierarchical clustering can determine the number of clusters by itself.

**11.2**

(a)

Ada and Bob, the number of identical purchased items and their probability:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| dist |  |  |  |  |  |  |  |  |
| J | 3/17 | 4/16 | 5/15 | 6/14 | 7/13 | 8/12 | 9/11 | 10/10 |
| P |  | | | | | | | |

Ada and Cathy, the number of identical purchased items and their probability:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| j | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| dist |  |  |  |  |  |  |  |  |  |  |
| J | 0/20 | 1/19 | 2/18 | 3/17 | 4/16 | 5/15 | 6/14 | 7/13 | 8/12 | 9/11 |
| P |  | | | | | | | | | |

Using Euclidean distance, the probability that dist(Ada, Bob) > dist(Ada, Cathy) is

Using Jaccard similarity, the probability that J(Ada, Bob) > J(Ada, Cathy) is

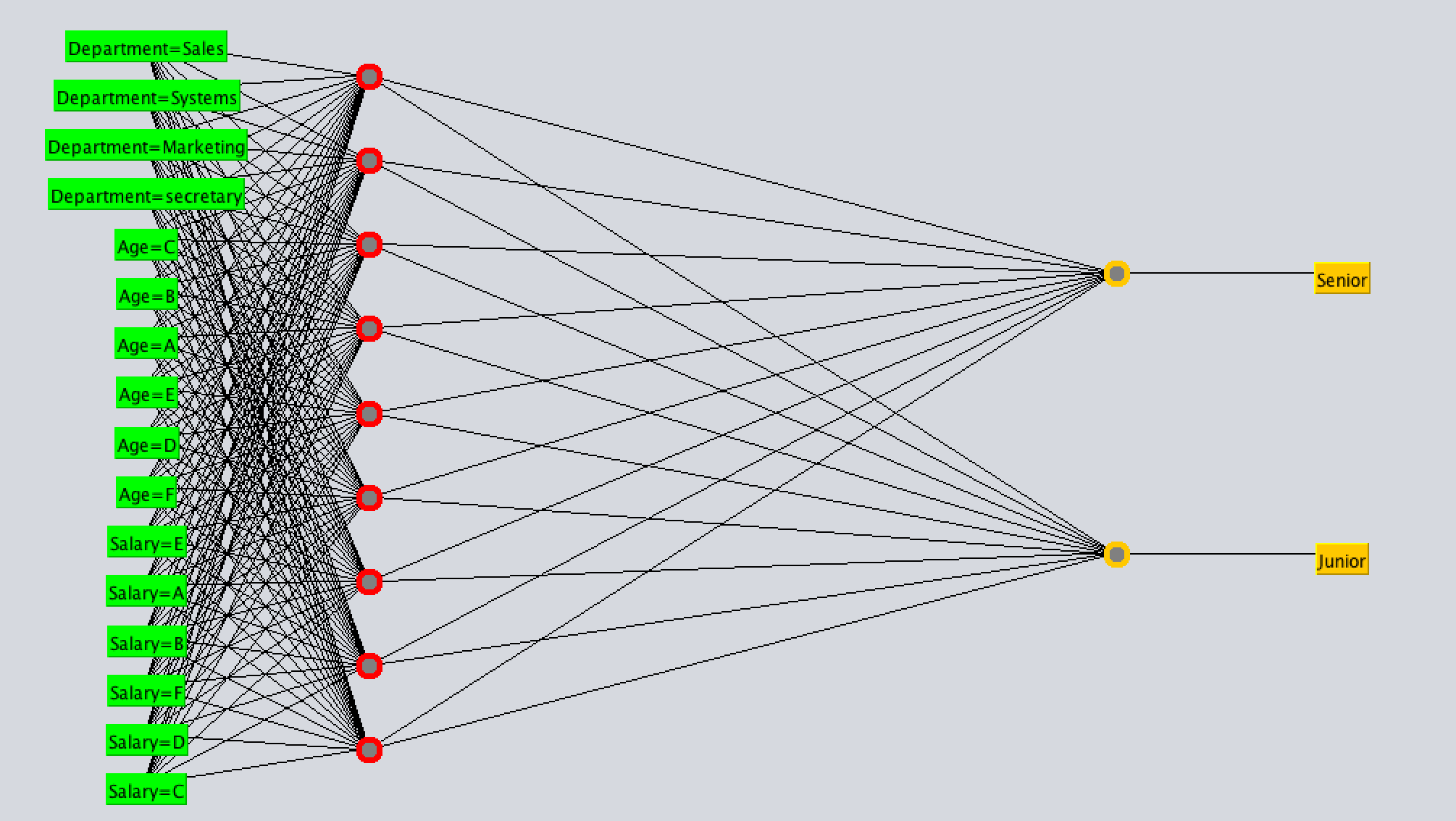
(b) The Euclidean distance is not bounded while Jaccard similarity is bounded into [0,1], both of them have realistic meaning and easy to explain, but Jaccard similarity is easier to calculate and use.

The probabilities of (Ada, Bob) > (Ada, Cathy) using different distance measure are very small.

The more similar a choice are, the smaller Euclidean distance is, but the larger Jaccard similarity is.

**9.1**

(a) Multiplayer feed forward neural network design:



(b) Training process:

Learning rate = 0.1, momentum = 0.05, after one iteration:

MSE error = 0.4051

Initial weights: W->random, b-> 0.

Weight values after first back propagation:

*Sigmoid Node 0*

*Inputs Weights*

*Threshold -0.26850603233689985*

*Node 2 -0.1536414339165637*

*Node 3 -0.17286860603738968*

*Node 4 -0.1364763907158364*

*Node 5 -0.1417153018851152*

*Node 6 -0.16754821927106747*

*Node 7 -0.14391559301750534*

*Node 8 -0.15019737513177056*

*Node 9 -0.1858561194454201*

*Node 10 -0.13526583430423164*

*Sigmoid Node 1*

*Inputs Weights*

*Threshold 0.3250999661508872*

*Node 2 0.11735720436348823*

*Node 3 0.09648946838429141*

*Node 4 0.1325624679326844*

*Node 5 0.15698257685819983*

*Node 6 0.12829954102240823*

*Node 7 0.207029580465071*

*Node 8 0.18924299523947155*

*Node 9 0.15837187964839178*

*Node 10 0.08615410010553524*

*Sigmoid Node 2*

*Inputs Weights*

*Threshold 0.038653348425512636*

*Attrib ﻿Department=Sales 0.05434560124976678*

*Attrib ﻿Department=Systems 0.050109545588148244*

*Attrib ﻿Department=Marketing -0.058778105453982016*

*Attrib ﻿Department=secretary 0.02722769706499169*

*Attrib Age=C -0.035306422313939545*

*Attrib Age=B 0.03638901057815278*

*Attrib Age=A -0.03679265696140824*

*Attrib Age=E -0.04488552333272854*

*Attrib Age=D 0.020980767695431522*

*Attrib Age=F -0.02202247968999582*

*Attrib Salary=E -0.02692232166122509*

*Attrib Salary=A 0.052523338128237396*

*Attrib Salary=B -0.006360847155370609*

*Attrib Salary=F 0.013645909211310722*

*Attrib Salary=D 0.006326506159760579*

*Attrib Salary=C 0.0187459385176509*

*Sigmoid Node 3*

*Inputs Weights*

*Threshold 0.030568650535688614*

*Attrib ﻿Department=Sales -0.010038075173673144*

*Attrib ﻿Department=Systems 0.029978347463971825*

*Attrib ﻿Department=Marketing 0.00851851742731286*

*Attrib ﻿Department=secretary -0.05763385303412723*

*Attrib Age=C 0.029922043977196826*

*Attrib Age=B 0.031451196499638075*

*Attrib Age=A 0.021062087074958543*

*Attrib Age=E 0.010894327729157069*

*Attrib Age=D 0.02397239995015439*

*Attrib Age=F 0.013304227604520234*

*Attrib Salary=E -0.037865927939002446*

*Attrib Salary=A 0.006361281116523003*

*Attrib Salary=B 0.014165063402946654*

*Attrib Salary=F -0.03393531269405591*

*Attrib Salary=D 0.04319757907148311*

*Attrib Salary=C -0.03192214927665161*

*Sigmoid Node 4*

*Inputs Weights*

*Threshold 0.025621421346336264*

*Attrib ﻿Department=Sales -0.022959579565800626*

*Attrib ﻿Department=Systems 0.028400297712035633*

*Attrib ﻿Department=Marketing -0.04148496296283999*

*Attrib ﻿Department=secretary -0.051597923726552586*

*Attrib Age=C -0.03238239430861803*

*Attrib Age=B -0.028159869376346614*

*Attrib Age=A 0.01043362167324314*

*Attrib Age=E 0.02480877631710337*

*Attrib Age=D -0.0643114556440574*

*Attrib Age=F 0.006107221343135527*

*Attrib Salary=E -0.06727112856062*

*Attrib Salary=A 0.013939539144729205*

*Attrib Salary=B -0.014690923558922342*

*Attrib Salary=F -0.03520433205299514*

*Attrib Salary=D -0.03742476273913456*

*Attrib Salary=C 0.021423337091295274*

*Sigmoid Node 5*

*Inputs Weights*

*Threshold -0.012979180996231918*

*Attrib ﻿Department=Sales 0.006976214507463208*

*Attrib ﻿Department=Systems 0.002681257165634325*

*Attrib ﻿Department=Marketing -0.030283947269863613*

*Attrib ﻿Department=secretary 0.006422783618505513*

*Attrib Age=C -0.02701986265433854*

*Attrib Age=B 0.027574556564154973*

*Attrib Age=A -0.02083584892503353*

*Attrib Age=E 0.010618497515612745*

*Attrib Age=D -0.06568762581176323*

*Attrib Age=F 0.032336785062663825*

*Attrib Salary=E -0.064217316156824*

*Attrib Salary=A 0.04516784911933486*

*Attrib Salary=B -0.008079318926876352*

*Attrib Salary=F -0.019756702283727445*

*Attrib Salary=D 0.04194935075287159*

*Attrib Salary=C -0.0337420068460443*

*Sigmoid Node 6*

*Inputs Weights*

*Threshold -0.011048713758458596*

*Attrib ﻿Department=Sales -0.021268362340937073*

*Attrib ﻿Department=Systems 0.011079090837327313*

*Attrib ﻿Department=Marketing -0.03094612448761499*

*Attrib ﻿Department=secretary 0.002085398469667209*

*Attrib Age=C 0.00979248174951219*

*Attrib Age=B 0.005323703460711975*

*Attrib Age=A -0.008390752868117322*

*Attrib Age=E 0.03279743984142937*

*Attrib Age=D -0.012389197175083987*

*Attrib Age=F 0.0245612549141821*

*Attrib Salary=E -0.06403706988049863*

*Attrib Salary=A 0.028996296095587258*

*Attrib Salary=B 0.0491891127007853*

*Attrib Salary=F -0.04563734132428389*

*Attrib Salary=D 0.025404940560991587*

*Attrib Salary=C -0.026936341502251156*

*Sigmoid Node 7*

*Inputs Weights*

*Threshold 0.05243809928953863*

*Attrib ﻿Department=Sales -0.0028839958256765205*

*Attrib ﻿Department=Systems -0.03487698037616963*

*Attrib ﻿Department=Marketing -0.00344964845987625*

*Attrib ﻿Department=secretary -0.058424064979996304*

*Attrib Age=C -0.06990004383692128*

*Attrib Age=B 0.058126765896145516*

*Attrib Age=A -0.015401500591844355*

*Attrib Age=E 0.011006841707771194*

*Attrib Age=D -0.050474252502331875*

*Attrib Age=F -0.029382902594712653*

*Attrib Salary=E -0.09105728947155484*

*Attrib Salary=A 0.07112061752691698*

*Attrib Salary=B -0.012549373930123535*

*Attrib Salary=F -0.006694426796025129*

*Attrib Salary=D 0.0302636156561919*

*Attrib Salary=C 0.03289645185702459*

*Sigmoid Node 8*

*Inputs Weights*

*Threshold 0.03716876286494403*

*Attrib ﻿Department=Sales 0.01597725913632879*

*Attrib ﻿Department=Systems -0.011844421708797868*

*Attrib ﻿Department=Marketing -0.05680055327387138*

*Attrib ﻿Department=secretary -0.02052475728654725*

*Attrib Age=C -0.03328777267265564*

*Attrib Age=B 0.06271813282102778*

*Attrib Age=A 0.006290729630263818*

*Attrib Age=E 0.007830196826772304*

*Attrib Age=D -0.04067850730489039*

*Attrib Age=F 0.02517295161105183*

*Attrib Salary=E -0.0070622816467351916*

*Attrib Salary=A -0.019352455913961782*

*Attrib Salary=B 0.0650263495377443*

*Attrib Salary=F 0.0179353197268242*

*Attrib Salary=D 0.0386029201442317*

*Attrib Salary=C -0.06693488117648713*

*Sigmoid Node 9*

*Inputs Weights*

*Threshold 0.014977177092840358*

*Attrib ﻿Department=Sales -0.01759614326096123*

*Attrib ﻿Department=Systems 0.01277406143287983*

*Attrib ﻿Department=Marketing -0.03940808929201179*

*Attrib ﻿Department=secretary 0.025762441970107283*

*Attrib Age=C 0.01812765276524513*

*Attrib Age=B -0.01379520274958834*

*Attrib Age=A 0.00678781474727209*

*Attrib Age=E -0.047323849335668314*

*Attrib Age=D -0.015261474739819016*

*Attrib Age=F 0.017671623261090202*

*Attrib Salary=E -0.08825332080678693*

*Attrib Salary=A 0.04924607547113777*

*Attrib Salary=B 0.014258696383286595*

*Attrib Salary=F -0.05663697719158005*

*Attrib Salary=D -0.0024072310316824936*

*Attrib Salary=C -0.00612771930633431*

*Sigmoid Node 10*

*Inputs Weights*

*Threshold -0.030782801677984733*

*Attrib ﻿Department=Sales -0.016139024714893688*

*Attrib ﻿Department=Systems 0.01279216702884576*

*Attrib ﻿Department=Marketing 0.021979004825723868*

*Attrib ﻿Department=secretary 0.012060425628644773*

*Attrib Age=C -0.039993472220185046*

*Attrib Age=B -0.0066111599104027455*

*Attrib Age=A -0.03476569630789775*

*Attrib Age=E 0.03957296340471577*

*Attrib Age=D -0.02908776186408368*

*Attrib Age=F 0.00910110083915472*

*Attrib Salary=E -0.018381112562647062*

*Attrib Salary=A -0.012255679517140343*

*Attrib Salary=B -0.031230216476898658*

*Attrib Salary=F 0.010016990812493675*

*Attrib Salary=D 0.014138839608671998*

*Attrib Salary=C -0.03779657482166918*

After 100 iterations:

MSE reduced to 0.0279

Accuracy = 99.4118 %

(c)\* SVM(BinarySMO in WEKA):

Kernel used: Linear Kernel

Classifier for classes: Senior, Junior

Attributes weights:

-0.271 \* (normalized) ﻿Department=Sales

+ 0.436 \* (normalized) ﻿Department=Systems

+ 0.1618 \* (normalized) ﻿Department=Marketing

+ -0.3268 \* (normalized) ﻿Department=secretary

+ -0.3387 \* (normalized) Age=C

+ 0.953 \* (normalized) Age=B

+ 0.9544 \* (normalized) Age=A

+ -0.3394 \* (normalized) Age=E

+ -0.7713 \* (normalized) Age=D

+ -0.4579 \* (normalized) Age=F

+ -0.6327 \* (normalized) Salary=E

+ 0.1311 \* (normalized) Salary=A

+ 1.3667 \* (normalized) Salary=B

+ -1.3404 \* (normalized) Salary=F

+ 0.9332 \* (normalized) Salary=D

+ -0.4579 \* (normalized) Salary=C

+ 0.243

MAE: 0.0059

Accuracy: 99.4118 %

Other classifier like Logistic Regression:

Logistic Regression with ridge parameter of 1.0E-8

Coefficients...

Class

Variable Senior

==============================================

﻿Department=Sales 7.8115

﻿Department=Systems -16.9648

﻿Department=Marketing 4.4046

﻿Department=secretary 12.9138

Age=C 7.1188

Age=B -10.5308

Age=A -29.9266

Age=E 42.1782

Age=D 44.0235

Age=F 42.4323

Salary=E 24.2832

Salary=A -10.098

Salary=B -21.0905

Salary=F 24.2934

Salary=D -91.5933

Salary=C 42.4323

Intercept -13.0612

MAE: 0.0094

Accuracy: 99.4118 %