Norman Chung
5550926
CS 165A Homework #2

1a. (A, E), (A, G), (B, E), (B, G), (C, E), (C, G), (D, G), (E, G)

1b. i) Yes

ii) No

iii) No

iv) No

v) No

1c. P(A=1, B=0, C=0, D=1, E=0, F=1, G=1, H=0)
P(A=1) * P(B=0|A=1) * P(C=0|B=0) * P(D=1|B=0, E=0) * P(E=0) * P(F=1|D=1, G=1) * P(G=1) * P(H=0| F=1)

1d. P(F=1|A=1, B=0, C=0, D=1, E=1, G=1, H=0) =
P(A=1, B=0, C=0, D=1, E=1, G=1,H=0,F=1) / P(A=1, B=0, C=0, D=1, E=1, G=1,H=0) =
P(A=1) * P(B=0|A=1) * P(C=0|B=0) * P(D=1|B=0, E=1) * P(E=1) * P(F=1|D=1, G=1) * P(G=1) * P(H=0| F=1) + P(A=1) * P(B=0|A=1) * P(C=0|B=0) * P(D=1|B=0, E=1) * P(E=1) * P(F=0|D=1, G=1) * P(G=1) * P(H=0| F=0) / P(A=1) * P(B=0|A=1) * P(C=0|B=0) * P(D=1|B=0, E=1) * P(E=1) * P(G=1)

P(F=1|D=1, G=1) * P(H=0| F=1) + P(F=0|D=1, G=1) * P(H=0| F=0)

1e. P(D|F=0, G=0, H=0) = P(F=0|D) * (G=0|D) * P(H=0|D) * P(D) / [P(F=0) * P(G=0) * P(H=0)]
For D=1: P(D=1|F=0, G=0, H=0)=(50/116) * (48/116) * (62/116) * (116/203) = 0.055
For D=0: P(D=0|F=0, G=0, H=0)=(23/87) * (50/87) * (18/87) * (87/203) = 0.0134
0.055 > 0.0134, so D is supposed to be 1

2a. P(Word$_1$|Author) * P(Word$_2$|Author) *…* P(Word$_{N-1}$|Author) * P(Word$_N$|Author) * P(Author)

2b. P(Author = yj | Word1=x1, Word2=x2,...,WordN=xN) =
P(Word$_1$ = x$_1$ | Author = y$_j$) * P(Word$_2$ = x$_2$ | Author =$_j$) * ... * P(Word$_N$ = x$_N$ | Author = y$_j$) * P(Author = y$_j$) / [P(Word$_1$ = x$_1$) * P(Word$_2$ = x$_2$) * ... * P(Word$_N$ = x$_N$)] * [P(Author = y$_1$) + P(Author = y$_2$) + … + P(Author = y$_m$)]

2c. Given that Word$_i$ = Word$_1$, Word$_2$, … ,Word$_N$, we know that P(Word$_1$ = x$_1$ | Author) = P(Word$_1$ = x$_2$ | Author) = ... = P(Word$_1$ = x$_N$ | Author) = P(Word$_2$ = x$_1$ | Author) = P(Word$_2$ = x$_2$ | Author) = ... = P(Word$_2$ = x$_N$ | Author)
Thus, P(Word$_1$ = x$_1$ | Author = y$_j$)^n * P(Author = y$_j$) / summation of [P(Word$_i$ = x$_i$ | Author = y$_m$)^n * P(Author = y$_m$)], where m goes from 1 to k

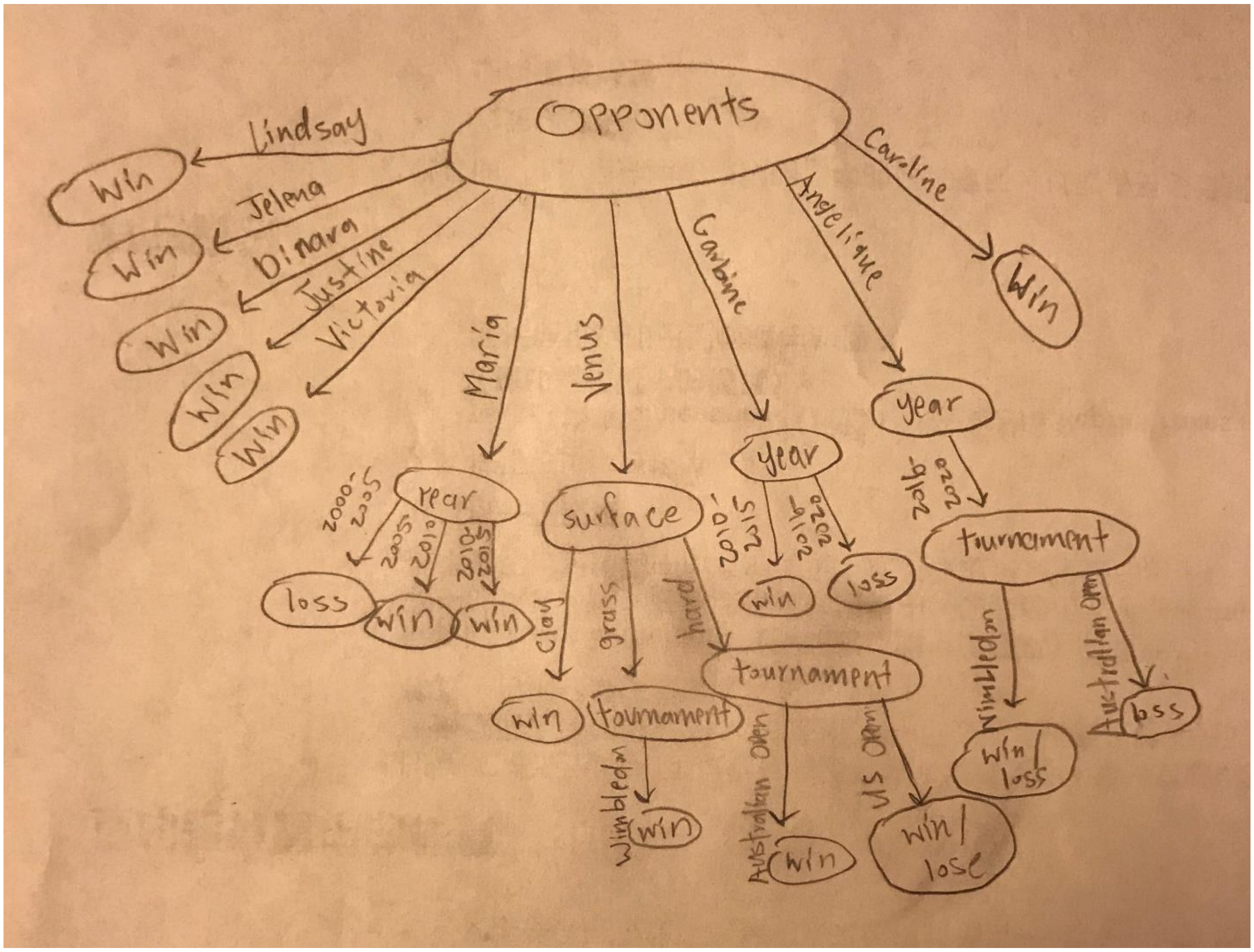
2d. k – 1

Norman Chung
5550926
CS 165A Homework #2

3a. The best attribute for root is opponents, as when you calculate the information gain, opponents has the highest gain value. The worst attribute is tournament, as it has the lowest gain value.

3b. You can split the root options into n nodes, starting from the highest gain value, and working in descending order, and you stop splitting when you reach a node with 0 entropy value.

3c.



i) She will not win.

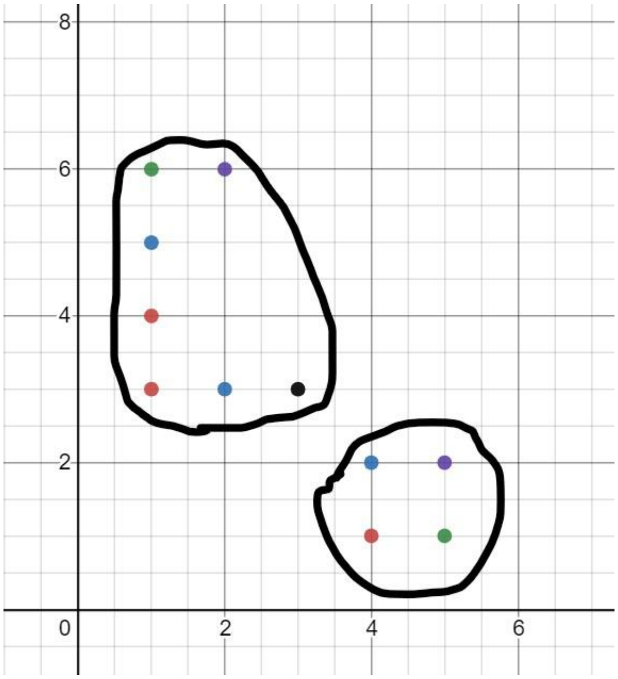
ii) She can probably win the Australian Open and Wimbledon.

4a. When k = 1, the label would have survival = 2. When k = 2, the label would have survival = 2. For k = 5, the label would have either survival = 1 or 2.
For different values of k, the results are not the same. This is because the number of neighbors can make a difference in the results.

4b. I think the best value for k should be 5. This is because there is a larger sample size as compared to when k=1 or k = 2. With this, a larger sample size would give us more data to help label unknown datapoints.

4c. Yes, it is an outlier, because the Euclidean distance between this point and the test point is much larger than in others. You can detect it by calculating its Euclidean distance.

5a. i)

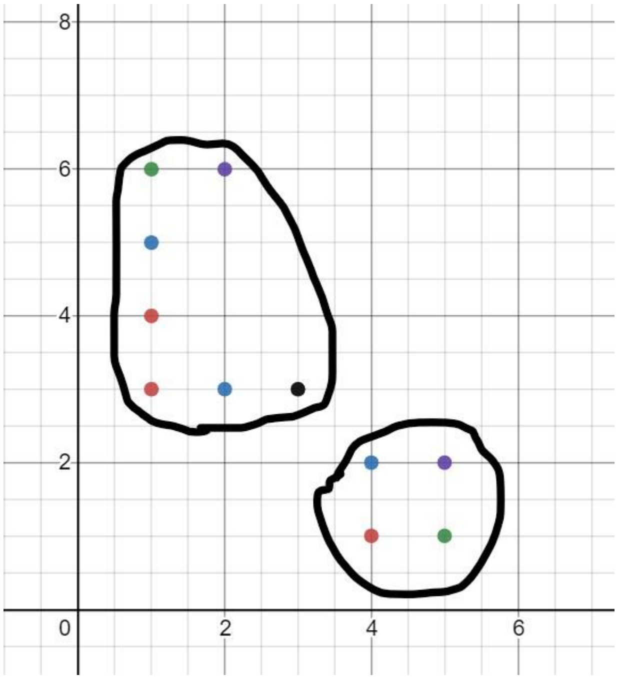


Left cluster: (1,3), (2,3), (1,4), (1,5), (1,6), (2,6), (3,3)
Right cluster: (4,1), (4,2), (5,1), (5,2)

ii) After the first iteration, the new means will be
Left x: (1 + 1 + 1 + 1 + 2 + 2 + 3) / 7 = 1.57
Left y: (6 + 5 + 4 + 3 + 6 + 3 + 3) / 7 = 4.28
Right x: (4 + 4 + 5 + 5) / 4 = 4.5
Right y: (1 + 1 + 2 + 2) / 4 = 1.5
(1.57, 4.28) and (4.5, 1.5)

iii) After the second iteration, the new means are the same: (1.57, 4.28) and (4.5, 1.5)

5b. i)



ii) For K-medoids, using the formula given, the left cluster has a cost of 14, and the right cluster has a cost of 4, giving us a total cost of 18.

iii) Given a new medoid of (2,6), the left cluster would have a new cost of 13, and the right cluster would have a new cost of 7, giving us a total of 20. The total cost increased.