**Assignment 3**

**Q1:**

1. P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
2. P(Cavity):

P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2

P(~cavity) = 0.016 + 0.064 + 0.144 + 0.576 = 0.8

|  |  |
| --- | --- |
| **Cavity** | **P** |
| cavity | 0.2 |
| ~cavity | 0.8 |

1. P(Toothache|cavity):

P(toothache|cavity) = (0.108 + 0.012) / (0.2) = 0.6

P(~toothache|cavity) = (0.072 + 0.008) / (0.2) = 0.4

|  |  |
| --- | --- |
| **Toothache** | **P** |
| toothache | 0.6 |
| ~toothache | 0.4 |

1. P(Cavity|toothache V catch):

P(toothache) = (0.108 + 0.012 + 0.016 + 0.064) = 0.2

P(catch) = (0.108 + 0.016 + 0.072 + 0.144) = 0.34

P(toothache, catch) = 0.108 + 0.016 = 0.124

P(toothache V catch) = P(toothache) + P(catch) – P(toothache, catch)

= 0.2 + 0.34 – 0.124

= 0.42

P(cavity|toothache V catch) = (0.108 + 0.012 + 0.072) / (0.42) = 0.46

P(~cavity|toothache V catch) = (0.016 + 0.064 + 0.144) / (0.42) = 0.54

|  |  |
| --- | --- |
| **Cavity** | **P** |
| cavity | 0.46 |
| ~cavity | 0.54 |

**Q2**:

**Legend**:

+x = positive test

-x = negative test

+y = infected by virus

-y = not infected by virus

**Information given in question:**

**Test A:**

P(+x|+y) = 0.95

P(-x|+y) = (1 - 0.95) = 0.05

P(+x|-y) = 0.10

P(-x|-y) = (1 – 0.10) = 0.90

**Test B :**

P(+x|+y) = 0.90

P(-x|+y) = 0.10

P(+x|-y) = 0.05

P(-x|-y) = 0.95

P(+y) = 0.01

P(-y) = 0.99

**Given a positive test, what is the probability that you have the virus?**

**Test A:**

P(+y|+x) = P(+x|+y)P(+y) / P(+x)

P(+x) = P(+x|+y)P(+y) + P(+x|-y)P(-y)

= (0.95)(0.01) + (0.10)(0.99)

= 0.0095 + 0.099

= 0.109

P(+y|+x) = (0.95)(0.01) / (0.109)

P(+y|+x) = 0.088

**Given that someone tests positive with test A, the probability that they are actually infected is 0.088.**

**Test B:**

P(+y|+x) = P(+x|+y)P(+y) / P(+x)

P(+x) = P(+x|+y)P(+y) + P(+x|-y)P(-y)

= (0.90)(0.01) + (0.05)(0.99)

= 0.009 + 0.0495

= 0.0585

P(+y|+x) = (0.90)(0.01) / (0.0585)

P(+y|+x) = 0.154

**Given that someone tests positive with test B, the probability that they are actually infected is 0.154.**

**A probability of 0.154 is greater than 0.088. So Test B returning a positive result is more indicative of someone carrying the virus than Test A.Q3:**

**Legend**:

+x = positive test

-x = negative test

+y = infected by virus

-y = not infected by virus

**Information given:**

P(+x|-y) = 0.05

P(-x|-y) = (1 - 0.05) = 0.95

P(-x|+y) = 0.02

P(+x|+y) = 0.98

P(+y) = 0.0001

P(-y) = 0.9999

1. **What is the chance that you have the disease?**

P(+y|+x) = P(+x|+y)P(+y) / P(+x)

P(+x) = P(+x|-y)P(-y) + P(+x|+y)P(+y)

= (0.05)(0.9999) + (0.98)(0.0001)

= 0.05

P(+y|+x) = (0.98)(0.0001)/(0.05)

= 0.002

**Given that you have tested positive, the chance of having the disease is 0.002**

1. **The result of the second independent test is still positive. What is now your chance of having the disease?**

P(+y|+x,+x) = P(+x,+x|+y)P(+y) / P(+x,+x)

P(+x,+x) = P(+x,+x|-y)P(-y) + P(+x,+x|+y)P(+y)

= (0.05)2(0.9999) + (0.98)2(0.0001)

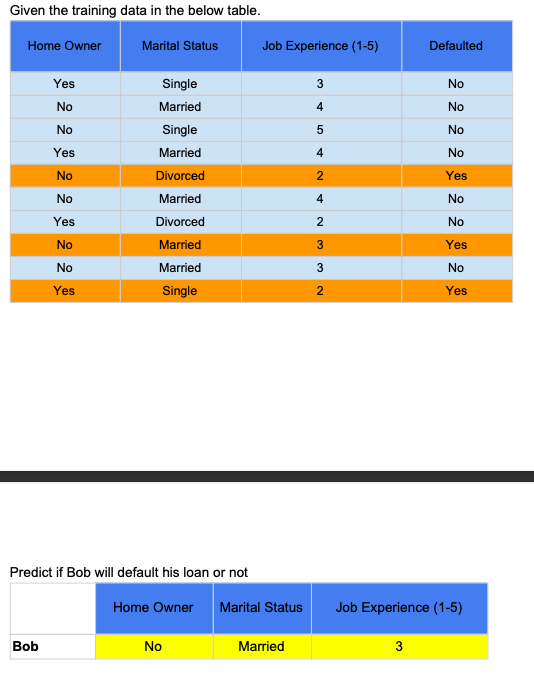
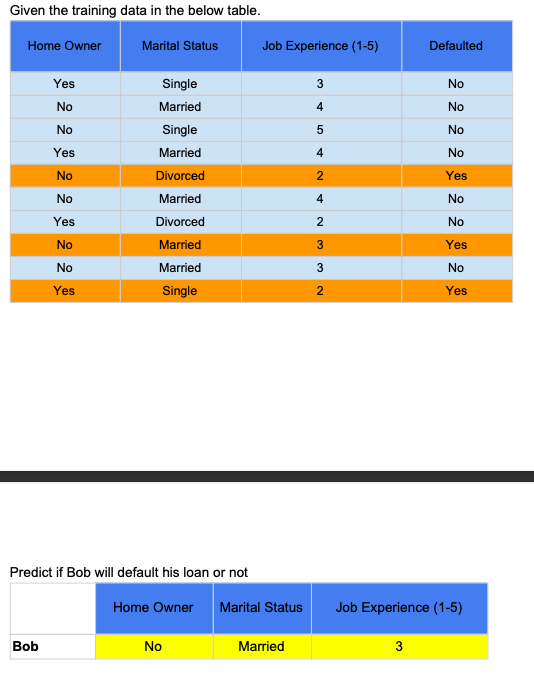
= 0025

P(+y|+x,+x) = (0.98)2(0.0001) / 0.0025

= 0.038

**Given two positive tests, there is a 0.038 chance of having the disease.**

**Q4 Classifiers:**

****



(D)

(J)

(M)

(H)

**\* Rows 8 and 9 are identical, but with different labels, so for the following classifiers I chose to ignore row 9.**

**\*\* Also note that I added single-letter abbreviations to each column name to be used as variables in the following calculations**

**Naïve Bayes Classifier**

P(D=No|Bob) = P(Bob|D=No) P(D=No)

= P(H=No|D=No) P(M=Married|D=No) P(J=3|D=No) P(D=No)

= (3/6)(3/6)(1/6)(6/9)

= 0.028

Given Bob’s attributes, the probability that he will not default his loan is 0.028.

P(D=Yes|Bob) = P(Bob|D=Yes) P(D=Yes)

= P(H=No|D=Yes) P(M=Married|D=Yes) P(J=3|D=Yes) P(D=Yes)

= (2/3)(1/3)(1/3)(3/9)

= 0.024

Given Bob’s attributes, the probability that he will default his loan is 0.024

0.028 > 0.024 Therefore, Naïve Bayes will predict that Bob will not default his loan.

**Decision Tree Classifier:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Defaulted | |  |  | Defaulted | |  |  | Defaulted | |
| Homeowner | No | Yes |  | Marital Status | No | Yes |  | Job Experience | No | Yes |
| Yes | 3 | 1 |  | Single | 2 | 1 |  | 2 | 1 | 2 |
| No | 3 | 2 |  | Married | 3 | 1 |  | 3 | 1 | 1 |
|  |  |  |  | Divorced | 1 | 1 |  | 4 | 3 | 0 |
|  |  |  |  |  |  |  |  | 5 | 1 | 0 |

From the tables, the only certainty right away is that people with 4 or more years of job experience never defaulted their loans. (Can classify 4 people)

Of the five remaining people, there are two who do not own a home (rows 5 and 8), and both of them defaulted their loans. (Can classify 2 people)

Of the three remaining, #7 is divorced, and they did not default their loan. None are married, so it cannot be predicted whether a married person would default their loan or not.

Of the 2 remaining, #1 with 3 years of job experience defaulted their loan whereas #10 with 2 years did not.



Bob:

Job experience >=4?

No

Home owner?

No

* Yes

The decision tree classifier predicts that Bob will default his loan.

**KNN Classifier (k=3)**

Distance, D, will be calculated by:

H + M + J

H = 0 if Home owner status matches

H = 1 if Home owner status doesn’t match

M = 0 if Marital status matches

M = 1 if Marital status doesn’t match

J = |JobExperience\_x – JobExperience\_y|

(ie. J is the number of years that the job experience differs from Bob)

**All distances from Bob:**

#1 = 1 + 1 + 0 = 2

#2 = 1

#3 = 3

#4 = 2

#5 = 2

#6 = 1

#7 = 3

#8 = 0

#10 = 3

**Closest 3:**

1. No, Married, 3, Yes (row 8, Distance = 0)
2. No, Married, 4, No (row 2, Distance = 1)
3. No, Married, 4, No (row 6, Distance = 1)

The first neighbour is obviously the closest to Bob because all their attributes match exactly. Row #2 and #6 differ in only the job experience years by 1 year. There are no other rows that match in at least 2 attributes (Therefore, they have a distance from Bob of at least 2).

Of the three nearest neighbours, 2/3 did not default their loans. Therefore it is predicted that Bob will not default his loan also.

**k-Means (k=2)**

Distance, D, will be calculated by:

H + M + J

H = 0 if Home owner status matches

H = 1 if Home owner status doesn’t match

M = 0 if Marital status matches

M = 1 if Marital status doesn’t match

J = |JobExperience\_x – JobExperience\_y|

(ie. J is the number of years that the job experience differs from the mean)

**1st iteration:**

Random starting points:

X1 = point 1 (Yes, Single, 3)

X2 = point 5 (No, Divorced, 2)

Distances to cluster centres:

D(point2, X1) = 1 + 1 + 1 = 3

D(point2, X2) = 0 + 1 + 2 = 3

Coin flip to assign… Point2 assigned to cluster 1

D(point3, X1) = 3

D(3, X2) = 4

Point3 assigned to cluster 1

D(4, X1) = 2

D(4, X2) = 4

Point4 assigned to cluster 1

D(6, X1) = 3

D(6, X2) = 3

Point6 assigned to cluster 2

D(7, X1) = 2

D(7, X2) = 1

Point7 assigned to cluster 2

D(8, X1) = 2

D(8, X2) = 2

Point8 assigned to cluster 2

D(10, X1) = 1

D(10, X2) = 2

Point 20 assigned to cluster 1

New clusters:

1. 1,2,3,4,10
2. 5,6,7,8

**2nd iteration:**

New means (cluster centres):

Means are points in which each attribute is calculated by finding the mode of the points in the cluster. Job experience is the calculated average

Cluster 1:

Home owner: Mode = Yes

Marital status: Mode = Single

Job experience: (3+4+5+4+2)/5 = 3.6

Cluster 2:

Home owner: No

Marital status: Married (random choice because there is a tie with “divorced”)

Job experience: 2.8

**After calculating distances…**

Point1 assigned to cluster 1

Point2 assigned to cluster 2

Point3 assigned to cluster 1

Point4 assigned to cluster 1

Point5 assigned to cluster 2

Point6 assigned to cluster 2

Point7 assigned to cluster 1

Point8 assigned to cluster 2

Point10 assigned to cluster 1

New Clusters:

1. 1,3,4,7,10
2. 2,5,6,8

New means:

Cluster 1:

Home owner = Yes

Marital Status = Single

Job Experience = 3.2

Cluster 2:

Home owner = No

Marital Status = Married

Job experience = 3.3