

HM2, MATLAB Project: Due on March 12, 2023

Problem 1: We have training data on 1000 pieces of fruit and among them 500 are bananas, 300 are oranges and 200 are other fruits. We consider 3 features of each fruit, whether it's long or not, sweet or not and yellow or not, as displayed in the table below.

Fruit	Long	Sweet	Yellow
Banana	400	350	450
Orange	0	150	300
Other	100	150	50

Now given an addition fruit with the features: Long, Sweet and Yellow, what is your prediction, the fruit is a banana, an orange or an other fruit? Why?

Hints: Based on the data, find the probabilities $p(\text{banana})=?$, $p(\text{orange})=?$, $p(\text{other})=?$, $p(\text{Long}|\text{banana})=?$, $p(\text{Sweet}|\text{banana})=?$, $p(\text{Yellow}|\text{banana})=?$ etc and note

$$\begin{aligned} p(\text{banana} | \text{Long, Sweet, Yellow}) \\ = \frac{p(\text{Long} | \text{banana}) p(\text{Sweet} | \text{banana}) p(\text{Yellow} | \text{banana}) p(\text{banana})}{p(\text{Long, Sweet, Yellow})} \end{aligned}$$

$$\begin{aligned} p(\text{orange} | \text{Long, Sweet, Yellow}) \\ = \frac{p(\text{Long} | \text{orange}) p(\text{Sweet} | \text{orange}) p(\text{Yellow} | \text{orange}) p(\text{orange})}{p(\text{Long, Sweet, Yellow})} \end{aligned}$$

$$p(\text{other} | \text{Long, Sweet, Yellow}) = \frac{p(\text{Long} | \text{other}) p(\text{Sweet} | \text{other}) p(\text{Yellow} | \text{other}) p(\text{other})}{p(\text{Long, Sweet, Yellow})}$$

Problem 2: Given a Data matrix with 58 rows and 2 columns that represents the Cartesian coordinates of 58 towns in West Germany:

- # 1 Augsburg, 2 Bielefeld, 3 Bochum, 4 Bremen, 5 Darmstadt, 6 Essen, 7 Freiburg,
- # 8 Hamburg, 9 Hannover, 10 Heilbronn, 10 Kaiserslautern, 132 Karlsruhe, 13 Kassel, 14 Kempten
- # 15 Koblenz, 16 Koeln, 17 Landshut, 18 Lichtenfels, 19 Mainz, 20 Muenchen, 21 Muenster
- # 22 Neuss, 23 Nuernburg, 24 Oldenburg, 25 Regensburg, 26 Rendsburg, 27 Stuttgart
- # 28 Ulm, 29 Wuerzburg, 30 Aachen, 31 Ansbach, 32 Aschaffenburg, 33 Bamberg, 34 Bayreuth
- # 35 Bonn, 36 Braunschweig, 37 Bremen, 38 Coburg, 39 Dortmund, 40 Duesseldorf

```
# 41 Duisburg, 42 Erlangen, 43 Frankfurt, 44 Fulda, 45 Fuerth, 46 Gelsen-Kirchen
# 47 Gummersburg, 48 Hagen, 49 Hersbruck, 50 Ingolstadt, 51 Kiel, 52 Mannheim
# 53 Marburg, 54 Offenburg, 55 Osnabrueck, 56 Reutlingen, 57 Saarbruecken, 58 Siegen
#
```

```
# Reference: Helmut Spaeth, ``Cluster Analysis Algorithms for Data Reduction and Classification of
Objects, Ellis Horwood, 1980, page 80.
```

```
#
```

```
City coordinates (according to the numerical order listed above) stored in the file Xdata.mat and their
affiliation (1=NW or 2=SE) stored in the file idx.mat
```

```
(1) 54.0 -65.0, (2) 0.0 71.0, (3) -31.0 53.0, (4) 8.0 111.0, (5) 1.0 -9.0, (6) -36.0 52.0, (7) -22.0 -76.0,
(8) 34.0 129.0, (9) 28.0 84.0, (10) 12.0 -38.0, (11) -21.0 -26.0, (12) -6.0 -41.0, (13) 21.0 45.0
(14) 38.0 -90.0, (15) -24.0 10.0, (16) -38.0 35.0, (17) 86.0 -57.0, (18) 58.0 -1.0, (19) -9.0 -3.0
(20) 70.0 -74.0, (21) -20.0 70.0, (22) -43.0 44.0, (23) 59.0 -26.0, (24) -5.0 114.0, (25) 83.0 -41.0
(26) 27.0 153.0, (27) 12.0 -49.0, (28) 30.0 -65.0, (29) 31.0 -12.0, (30) -57.0 28.0, (31) 44.0 -28.0
(32) 7.0 -7.0, (33) 54.0 -8.0, (34) 65.0 -8.0, (35) -35.0 25.0, (36) 46.0 79.0, (37) 5.0 118.0
(38) 56.0 4.0, (39) -21.0 54.0, (40) -40.0 45.0, (41) -43.0 51.0, (42) 57.0 -21.0, (43) 0.0 0.0
(44) 25.0 15.0, (45) 56.0 -25.0, (46) -34.0 56.0, (47) -24.0 36.0, (48) -25.0 49.0, (49) 64.0 -26.0
(50) 63.0 -48.0, (51) 37.0 155.0, (52) -5.0 -24.0, (53) 2.0 28.0, (54) -18.0 -58.0, (55) -10.0 82.0
(56) 12.0 -58.0, (57) -40.0 -28.0, (58) -16.0 28.0
```

```
The city coordinates can be plotted on a 2 dimensional plane in MATLAB by
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
clear
```

```
close all
```

```
load Xdata
```

```
load idx
```

```
hold on
```

```
for i=1:length(Xdata)
```

```
    text(Xdata(i,1),Xdata(i,2),num2str(i))
```

```

end
axis([-60 100 -100 180])
figure
hold on
for i=1:length(Xdata)
    if idx(i) > 1.5
        plot(Xdata(i,1),Xdata(i,2),'bo')
    else plot(Xdata(i,1),Xdata(i,2),'r+')
    end
end
end
testdata=[0 22];
plot(testdata(1),testdata(2),'bd','linewidth',4)

```

%%

Cities 26, 51, 24, 43, 7, 8, 55, 9, 36, 21, 2, 41, 6, 46, 3, 22, 40, 16, 39, 48, 13, 30, 35, 58, 53, 47, 15, 58, 53, 13, 36 are in the class north-west (NW) and the rests are in class south-east (SE) cities. City identification is in the data set idx.mat (load idx in MATLAB)

- 1) Determine the separating line in terms of the slope (m) and y-intercept (b).
- 2) Plot the separating line on top of cities on a two-dimensional plane.
- 3) The coordinate of the city Giessen is (0,22). Determine if Giessen is in the cluster NW or the cluster SE?