Berkan Ozturk V00892651
Reece Pretorius V00880300

SENG 474 Assignment 2 Part 1

1.1. SVM

a.

1. (9 pts) Consider the dataset in Fig 1, with points belonging to two classes, blue squares and red circles.

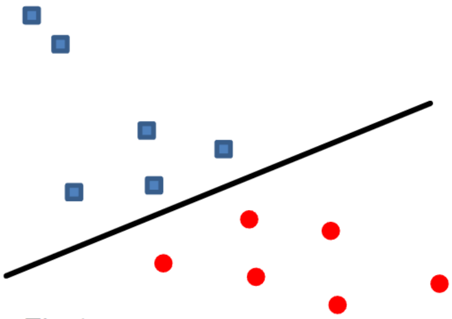

Fig. 1

b. $\frac{1}{2} * w^2 = 2$ therefore $||w|| = 2$ and we know the formula for calculating the margin, which is $1/\ ||w||$. Applying this formula, we get:

margin = ½ = 0.5

c.

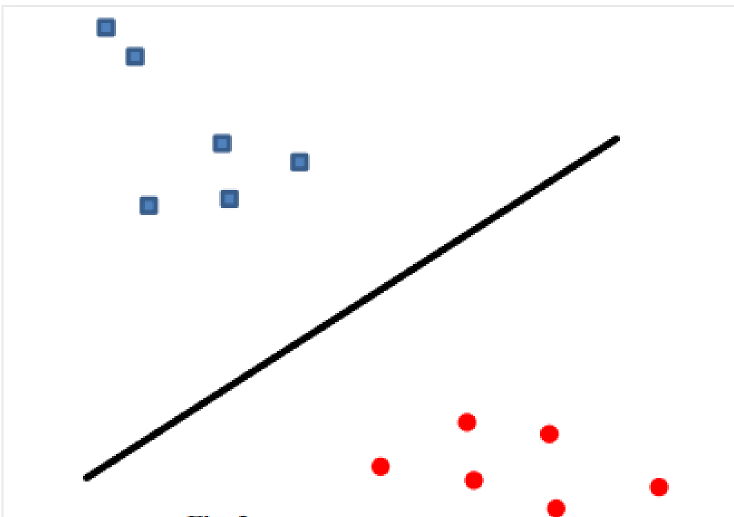

Fig. 2

The value of $\frac{1}{2} * w^2$ will be smaller than the previous answer because now we have a greater margin for error and the value of w is smaller; therefore, $\frac{1}{2} * w^2$ will be smaller.

d. Using a ruler, we found that the margin for error is now 4 times more than the margin in Figure 1. The margin in Figure 1. was ½; thus we found the margin for error to be 4/2, which is equal to 2. We know the formula for margin, which is:

margin = 1 / ||w'||

In this case, we know the value of margin. We plug in the value of margin in the equation, and we get the value of the new line coefficient vector, which is ½.

new line coefficient vector (w') = ½ = 0.5

e.

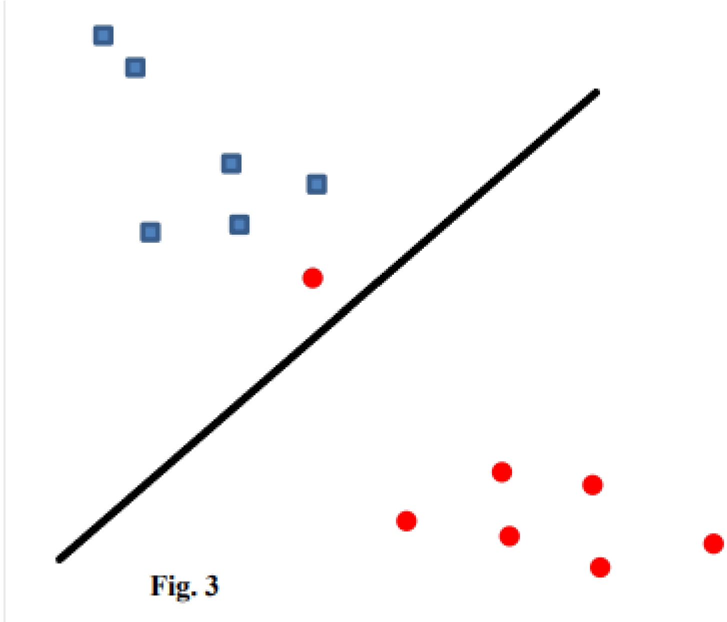

Fig. 3

This line is better than the line that perfectly separates the point because the perfect line would not give enough margin space for both blue squares and red circles. This line gives enough margin for error for both blue squares and red circles; therefore, this line fits better.

f. We prefer the line in part e because it gives enough margin for both red circles and blue squares. This line does **not** produce overfitting as it gives enough margin for both classes.

1.2. Tweet Classification

Refer to Text_Classification.ipynb and Text_Classification.pdf

SENG 474 Assignment 2 Part 2

2.1. Decision Tree using entropy

Calculations for choosing root node

Pclass:

4 types: "1st", "2nd", "3rd", "crew"

Using: =COUNTIF(B2:B2202, "1st"), =COUNTIF(B2:B2202, "2nd"), =COUNTIF(B2:B2202, "3rd"), =COUNTIF(B2:B2202, "crew") I counted the number of occurrences of each of the values.

1st = 325, 2nd = 285, 3rd = 706, crew = 885, total = 2201

=COUNTIFS(B2:B2202,"1st",E2:E2202,"yes") and =COUNTIFS(B2:B2202,"1st",E2:E2202,"no") for 1st class

=COUNTIFS(B2:B2202,"2nd",E2:E2202,"yes") and =COUNTIFS(B2:B2202,"2nd",E2:E2202,"no") for 2nd class

=COUNTIFS(B2:B2202,"3rd",E2:E2202,"yes") and =COUNTIFS(B2:B2202,"3rd",E2:E2202,"no") for 3rd class

=COUNTIFS(B2:B2202,"crew",E2:E2202,"yes") and =COUNTIFS(B2:B2202,"crew",E2:E2202,"no") for crew

gives us the remaining values we need for the entropy calculation for pclass:1st, pclass:2nd, pclass:3rd and pclass:crew:

Pclass=1st
info([**203, 122**]) = entropy(**203/325, 122/325**) = -(**203/325**)log(**203/325**) - (**122/325**)log(**122/325**) = 0.9488

Pclass=2nd
info([**118, 167**]) = entropy(**118/285, 167/285**) = -(**118/285**)log(**118/285**) - (**167/285**)log(**167/285**) = 0.9786

Pclass=3rd
info([**178, 528**]) = entropy(**178/706, 528/706**) = -(**178/706**)log(**178/706**) - (**528/706**)log(**528/706**) = 0.8146

Pclass=crew
info([**212, 673**]) = entropy(**212/885, 673/885**) = -(**212/885**)log(**212/885**) - (**673/885**)log(**673/885**) = 0.7943

info([**203, 122**], [**118, 167**], [**178, 528**], [**212, 673**]) = 0.9488*(325/2201) + 0.9786*(285/2201) + 0.8146*(706/2201) + 0.7943*(885/2201) = **0.8475**

Age:
2 types: “Adult”, “Child”

Using: =COUNTIF(C2:C2202, "adult") and =COUNTIF(C2:C2202, "child") I counted the number of occurrences of each of the values.

Adult = 2092, Child = 109

=COUNTIFS(C2:C2202,"adult",E2:E2202,"yes") and
=COUNTIFS(C2:C2202,"adult",E2:E2202,"no") for adult

=COUNTIFS(C2:C2202,"child",E2:E2202,"yes") and
=COUNTIFS(C2:C2202,"child",E2:E2202,"no") for child

gives us the remaining values we need for the entropy calculation for age:adult and age:child:

Age=Adult
info([**654, 1438**]) = entropy(**654/2092, 1438/2092**) = -(**654/2092**)log(**654/2092**) - (**1438/2092**)log(**1438/2092**) = 0.8962

Age=Child
info([**57, 52**]) = entropy(**57/109, 52/109**) = -(**57/109**)log(**57/109**) - (**52/109**)log(**52/109**) = 0.9985

info([**654, 1438**], [**57, 52**]) = 0.8962*(2092/2201) + 0.9985*(109/2201) = **0.9013**

Sex:

2 types: “Male”, “Female”

Using: =COUNTIF(D2:D2202, "male") and =COUNTIF(D2:D2202, "female") I counted the number of occurrences of each of the values.

Male = 1731, Female = 470

=COUNTIFS(D2:D2202,"female",E2:E2202,"yes") and
=COUNTIFS(D2:D2202,"female",E2:E2202,"no") for female

=COUNTIFS(D2:D2202,"male",E2:E2202,"yes") and
=COUNTIFS(D2:D2202,"male",E2:E2202,"no") for male

gives us the remaining values we need for the entropy calculation for sex:female and sex:male:

Sex=female
info([**344, 126**]) = entropy(**344/470, 126/470**) = -(**344/470**)log(**344/470**) - (**126/470**)log(**126/470**) = 0.8387

Sex=male
info([**367, 1364**]) = entropy(**367/1731, 1364/1731**) = -(**367/1731**)log(**367/1731**) - (**1364/1731**)log(**1364/1731**) = 0.7453

info([**344, 126**], [**367, 1364**]) = 0.8387*(470/2201) + 0.7453*(1731/2201) = **0.7652**

- **Root Node:** Based on the entropy calculations of the 3 attributes Pclass, Age and sex the lowest entropy was calculated from the sex column so the root node for the decision tree will be: “Sex” with branches “male” and “female”.

First level calculations:

Using the same function in excel to count, we did the entropy calculations for the first level as follows:

[Sex:female] Pclass=1st
info([**141, 4**]) = entropy(**141/145, 4/145**) = -(**141/145**)log(**141/145**) - (**4/145**)log(**4/145**) = 0.1821

[Sex:female] Pclass=2nd
info([**93, 13**]) = entropy(**93/106, 13/106**) = -(**93/106**)log(**93/106**) - (**13/106**)log(**13/106**) = 0.5369

[Sex:female] Pclass=3rd
info([**90, 106**]) = entropy(**90/196, 106/196**) = -(**90/196**)log(**90/196**) - (**106/196**)log(**106/196**) = 0.9952

[Sex:female] Pclass=crew
info([**20, 3**]) = entropy(**20/23, 3/23**) = -(**20/23**)log(**20/23**) - (**3/23**)log(**3/23**) = 0.5586

Expected info: 0.1821*(145/470) + 0.5369*(106/470) + 0.9952*(196/470) + 0.5586*(23/470) = **0.6196**

[Sex:female] Age=adult
info([**316, 109**]) = entropy(**316/425, 109/425**) = -(**316/425**)log(**316/425**) - (**109/425**)log(**109/425**) = 0.8214

[Sex:female] Age=child
info([**28, 17**]) = entropy(**28/45, 17/45**) = -(**28/45**)log(**28/45**) - (**17/45**)log(**17/45**) = 0.9565

Expected info: 0.8214*(425/470) + 0.9565*(45/470) = **0.8343**

- Therefore given that the expected info for Pclass is lower it is the feature we choose for the node following Sex:female.

[Sex:male] Pclass=1st
info([**62, 118**]) = entropy(**62/180, 118/180**) = -(**62/180**)log(**62/180**) - (**118/180**)log(**118/180**) = 0.9290

[Sex:male] Pclass=2nd
info([**25, 154**]) = entropy(**25/179, 154/179**) = -(**25/179**)log(**25/179**) - (**154/179**)log(**154/179**) = 0.5834

[Sex:male] Pclass=3rd
info([**88, 422**]) = entropy(**88/510, 422/510**) = -(**88/510**)log(**88/510**) - (**422/510**)log(**422/510**) = 0.6635

[Sex:male] Pclass=crew
info([**192, 670**]) = entropy(**192/862, 3/862**) = -(**192/862**)log(**192/862**) - (**670/862**)log(**670/862**) = 0.7651

Expected info: 0.9290*(180/1731) + 0.5834*(179/1731) + 0.6635*(510/1731) + 0.7651*(862/1731) = **0.7334**

[Sex:male] Age=adult
info([**338, 1329**]) = entropy(**338/1667, 1329/1667**) = -(**338/1667**)log(**338/1667**) - (**1329/1667**)log(**1329/1667**) = 0.7274

[Sex:male] Age=child
info([**29, 35**]) = entropy(**29/64, 17/64**) = -(**29/64**)log(**29/64**) - (**35/64**)log(**35/64**) = 0.9937

Expected info: 0.7274*(1667/1731) + 0.9937*(64/1731) = **0.7372**

- Therefore given that the expected info for Pclass is slightly lower than the expected info for Age, Pclass is the feature we choose for the node following Sex:male. With the final root and first level tree shown in **figure 1** below.

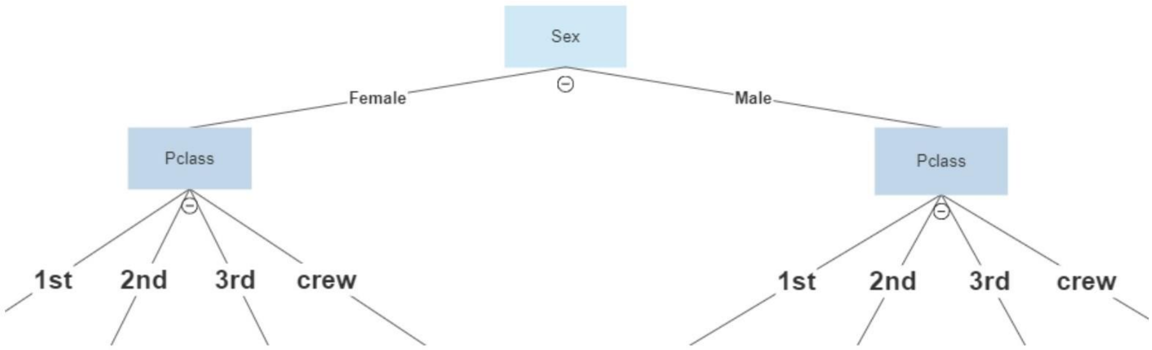

Figure 1: Tree root and first level

2.2. Naive Bayes method

P(survived = yes) = 711 / 2201 = 0.32303

P(survived = no) = 1490 / 2201 = 0.67697

pclass	Yes	No
1st	203 / 711	122 / 1490
2nd	118 / 711	167 / 1490
3rd	178 / 711	528 / 1490
crew	212 / 711	673 / 1490

age	Yes	No
adult	654 / 711	1438 / 1490
child	57 / 711	52 / 1490

sex	Yes	No
female	344 / 711	126 / 1490
male	367 / 711	1364 / 1490

v_{NB}(yes) = P(survived = yes) * P(pclass = 2nd | yes) * P(age = child | yes) * P(sex = male | yes) = **0.00221852171**

v_{NB}(no) = P(survived = no) * P(pclass = 2nd | no) * P(age = child | no) * P(sex = male | no) = **0.00242405017**

After normalization:

P(survived = yes | E) = v_{NB}(yes) / v_{NB}(yes) + v_{NB}(no) = **0.47786480488 = %48**

P(survived = no | E) = v_{NB}(no) / v_{NB}(yes) + v_{NB}(no) = **0.52213519416 = %52**

- **Based on the probabilities, the 2nd class male child won't survive.**

v_{NB}(yes) = P(survived = yes) * P(pclass = 2nd | yes) * P(age = adult | yes) * P(sex = female | yes) = **0.02385936901**

v_{NB}(no) = P(survived = no) * P(pclass = 2nd | no) * P(age = adult | no) * P(sex = female | no) = **0.00619231902**

After normalization:

P(survived = yes | E) = v_{NB}(yes) / v_{NB}(yes) + v_{NB}(no) = **0.79394438629 = %79**

P(survived = no | E) = v_{NB}(no) / v_{NB}(yes) + v_{NB}(no) = **0.2060556137 = %21**

- **Based on the probabilities, the 2nd class female adult will survive.**

Based on our calculations, we can also derive the table below.

pclass	age	sex	survived
2nd	child	male	**no**
2nd	adult	female	**yes**