P5 PartA GroupPi

Lydia, Jordan, Alan Bouwman

April 2021

1. Group Pi. Lydia Savatsky, Jordan DeYonker, Alan Bouwman
2.

   a. Bernoulli Naive Bayes

      i $P(X_{peony}|Class = 2) = \frac{4+1}{14+2} = \frac{5}{16}$

      ii $P(X_{crocus}|Class = 2) = \frac{1+1}{14+2} = \frac{2}{16} = \frac{1}{8}$

      iii $P(X_{peony}|Class = 1) = \frac{1+1}{8+2} = \frac{2}{10} = \frac{1}{5}$

   b. Multinomial Naive Bayes

      i $P(X_{peony}|Class = 2) = \frac{4+1}{14+14} = \frac{5}{28}$

      ii $P(X_{crocus}|Class = 2) = \frac{1+1}{14+14} = \frac{2}{28} = \frac{1}{14}$

      iii $P(X_{peony}|Class = 1) = \frac{1+1}{8+14} = \frac{2}{22} = \frac{1}{11}$

   c. Prediction using Bernoulli Naive Bayes
      $P(X_{daffodil}|Class = 1) = \frac{0+1}{8+2} = \frac{1}{10}$
      $P(X_{corcus}|Class = 1) = \frac{0+1}{8+2} = \frac{1}{10}$
      $P(X_{daisy}|Class = 1) = \frac{0+1}{8+2} = \frac{1}{10}$
      $P(X_{tulip}|Class = 1) = \frac{1+1}{8+2} = \frac{2}{10} = \frac{1}{5}$
      $P(X_{clematis}|Class = 1) = \frac{1+1}{8+2} = \frac{2}{10} = \frac{1}{5}$
      $P(X_{peony}|Class = 1) = \frac{1+1}{8+2} = \frac{2}{10} = \frac{1}{5}$
      $P(X_{daffodil}|Class = 2) = \frac{0+1}{14+2} = \frac{1}{16}$
      $P(X_{corcus}|Class = 2) = \frac{1+1}{14+2} = \frac{2}{16} = \frac{1}{8}$
      $P(X_{daisy}|Class = 2) = \frac{0+1}{14+2} = \frac{1}{16}$
      $P(X_{tulip}|Class = 2) = \frac{0+1}{14+2} = \frac{1}{16}$
      $P(X_{clematis}|Class = 2) = \frac{2+1}{14+2} = \frac{3}{16}$
      $P(X_{peony}|Class = 2) = \frac{4+1}{14+2} = \frac{5}{16}$
      $P(X_{daffodil}|Class = 3) = \frac{0+1}{7+2} = \frac{1}{9}$
      $P(X_{corcus}|Class = 3) = \frac{0+1}{7+2} = \frac{1}{9}$
      $P(X_{daisy}|Class = 3) = \frac{1+1}{7+2} = \frac{2}{9}$
      $P(X_{tulip}|Class = 3) = \frac{1+1}{7+2} = \frac{2}{9}$

$P(X_{clematis}|Class = 3) = \frac{0+1}{7+2} = \frac{1}{9}$
$P(X_{peony}|Class = 3) = \frac{0+1}{7+2} = \frac{1}{9}$
Thus,

$$P(Class = 1|doc) = \frac{1}{4} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} = 2.0 \times 10^{-6}$$

$$P(Class = 2|doc) = \frac{1}{2} \cdot \frac{1}{16} \cdot \frac{1}{8} \cdot \frac{1}{16} \cdot \frac{1}{16} \cdot \frac{3}{16} \cdot \frac{5}{16} = 8.94 \times 10^{-7}$$

$$P(Class = 3|doc) = \frac{1}{4} \cdot \left(\frac{1}{9}\right)^4 \cdot \left(\frac{2}{9}\right)^2 = 1.88 \times 10^{-6}$$

We will predict class 1.

d. Prediction using Multinomial Naive Bayes
$P(X_{daffodil}|Class = 1) = \frac{0+1}{8+14} = \frac{1}{22}$
$P(X_{corcus}|Class = 1) = \frac{0+1}{8+14} = \frac{1}{22}$
$P(X_{daisy}|Class = 1) = \frac{0+1}{8+14} = \frac{1}{22}$
$P(X_{tulip}|Class = 1) = \frac{1+1}{8+14} = \frac{2}{22} = \frac{1}{11}$
$P(X_{clematis}|Class = 1) = \frac{1+1}{8+14} = \frac{2}{22} = \frac{1}{11}$
$P(X_{peony}|Class = 1) = \frac{1+1}{8+14} = \frac{2}{22} = \frac{1}{11}$
$P(X_{daffodil}|Class = 2) = \frac{0+1}{14+14} = \frac{1}{28}$
$P(X_{corcus}|Class = 2) = \frac{1+1}{14+14} = \frac{2}{28} = \frac{1}{14}$
$P(X_{daisy}|Class = 2) = \frac{0+1}{14+14} = \frac{1}{28}$
$P(X_{tulip}|Class = 2) = \frac{0+1}{14+14} = \frac{1}{28}$
$P(X_{clematis}|Class = 2) = \frac{2+1}{14+14} = \frac{3}{28}$
$P(X_{peony}|Class = 2) = \frac{4+1}{14+14} = \frac{5}{28}$
$P(X_{daffodil}|Class = 3) = \frac{0+1}{7+14} = \frac{1}{21}$
$P(X_{corcus}|Class = 3) = \frac{0+1}{7+14} = \frac{1}{21}$
$P(X_{daisy}|Class = 3) = \frac{1+1}{7+14} = \frac{2}{21}$
$P(X_{tulip}|Class = 3) = \frac{1+1}{7+14} = \frac{2}{21}$
$P(X_{clematis}|Class = 3) = \frac{0+1}{7+14} = \frac{1}{21}$
$P(X_{peony}|Class = 3) = \frac{0+1}{7+14} = \frac{1}{21}$
Thus,

$$P(Class = 1|doc) = \frac{1}{4} \cdot \frac{1}{22} \cdot \frac{1}{22} \cdot \frac{1}{22} \cdot \frac{1}{11} \cdot \frac{1}{11} \cdot \frac{1}{11} = 1.764 \times 10^{-8}$$

$$P(Class = 2|doc) = \frac{1}{2} \cdot \frac{1}{28} \cdot \frac{1}{14} \cdot \frac{1}{28} \cdot \frac{1}{28} \cdot \frac{3}{28} \cdot \frac{5}{28} = 3.11 \times 10^{-8}$$

$$P(Class = 3|doc) = \frac{1}{4} \cdot \left(\frac{1}{21}\right)^4 \cdot \left(\frac{2}{21}\right)^2 = 1.166 \times 10^{-8}$$

We would choose class 2 this time.

3.

a. Term Document Matrix:

	1	2	3
bat	1	3	0
cat	3	0	1
fat	1	0	1
mat	0	1	1
pat	0	1	1
rat	1	1	1
sat	0	0	1

b. TF-IDF Matrix:
We start with the document frequency vector:

word	df
bat	2
cat	2
fat	2
mat	2
pat	2
rat	3
sat	1

Now here is the TF-IDF matrix:

word	1	2	3
bat	0.053	0.106	0
cat	0.106	0	0.053
fat	0.053	0	0.053
mat	0	0.053	0.053
pat	0	0.053	0.053
rat	0	0	0
sat	0	0	0.144

c. The term-document pair with the highest TF-IDF value is $(sat, 3)$.