

Midterm 2019 Answers (Part A Only)

I. I. $P(x = \text{"so"}) | y = \text{"Intent to Buy"}) = \frac{P(x = \text{"so"}, y = \text{"Intent to Buy"})}{P(y = \text{"Intent to Buy"})}$

$$= \frac{\frac{1}{6}}{\frac{1}{2}} = \boxed{\frac{1}{3}}$$

II. $P(x = \text{"see"}) = \frac{2}{6} = \boxed{\frac{1}{3}}$

III. $P(x_i = \text{"see"}, x_{ii} = \text{"move"}) = \boxed{\frac{2}{6}}$ (\downarrow comments where "see" and "move" appear in same sentence)

IV. $P(y = \text{"No Intent"}) | x = \text{"bad"}) = \frac{P(y = \text{"No Intent"}, x = \text{"bad"})}{P(x = \text{"bad"})}$

$$= \frac{\frac{1}{6}}{\frac{2}{6}} = \boxed{\frac{1}{2}}$$

B. events A and B are independent if $P(A, B) = P(A)P(B)$

$$P(x_i = \text{loc}, x_j = \text{move}) = P(x = \text{loc}) P(x = \text{move})$$

\downarrow \downarrow

should equal
this if
independent $\rightarrow \frac{1}{6}$ = $\frac{1}{6}$ $\frac{1}{6}$

$P(x = \text{loc}, x = \text{move}) = \frac{1}{6}$ (\downarrow only 1 row where "loc" and "move" appear together).

C. priors:

$$P(y = \text{No Intent}) = \frac{3}{6} = \boxed{\frac{1}{2}}$$

$$P(y = \text{Intent}) = \boxed{\frac{1}{2}}$$

likelihood:

$$P(x | y = \text{No Intent}) = P(x = \text{look} | y = \text{No}) \times P(x = \text{bad} | y = \text{No})$$

$$\boxed{\frac{1}{9}} = \frac{1}{3} \quad \quad \quad \frac{1}{3}$$

$$p(x|y=Intent) = p(x=look|y=Intent) \times p(x=bad|y=Intent)$$

$$\boxed{\frac{1}{9}} = \frac{1}{3} \quad \frac{1}{3}$$

evidence:

$$\boxed{\frac{1}{9}} = p(x|y=Intent)p(y=Intent) + p(x|y=No)p(y=No)$$

$$\boxed{\frac{1}{9}} = \left(\frac{1}{9}\right) \left(\frac{1}{2}\right) \quad \left(\frac{1}{9}\right) \left(\frac{1}{2}\right)$$

posterior: $p(y=Intent|x) = \frac{p(x|y=Intent)p(y=Intent)}{p(x|y=Intent)p(y=Intent) + p(x|y=No)p(y=No)}$

$$\boxed{\frac{1}{2}} = \frac{\left(\frac{1}{9}\right)\left(\frac{1}{2}\right)}{\left(\frac{1}{9}\right)}$$

same for $p(y=No|Intent|x)$

2. A.i. word count:

	old	blue	navy	sweater	trendy	wood	chair	decor	jeans
A	0	0	0	0	1	1	1	0	0
B	2	1	1	0	0	0	0	1	1
C	1	1	0	1	0	0	0	1	0

B. one-hot:

same as A, except the 2 becomes a 1.

	old	blue	navy	sweater	trendy	wood	chair	decor	jeans
A	0	0	0	0	1	1	1	0	0

	IDF	old	blue	navy	sweater	trendy	wood	chair	decor	jeans
TF _A	2	2	2.5	2.5	2	2.5	2.5	2	2.5	0
TF _B	0	0	0	0	1	1	1	0	0	0
TF _C	2	1	1	0	0	0	0	1	1	0

this is the same as word count

$$IDF = 1 + \frac{N}{df(f)+1}$$

For words that appear 1 time:
(navy, sweater, wood, chair, jeans):

$$1 + \frac{3}{7+1} = \underline{\underline{2.5}}$$

For words that appear 2 times:
(old, blue, discount, trendy)

$$1 + \frac{3}{3} = 2$$

	old	blue	navy	sweater	trendy	wood	chair	discount	jeans
TFIDF _A	0	0	0	0	2	2.5	2.5	0	0
TFIDF _B	4	2	2.5	0	0	0	0	2	2.5
TFIDF _C	2	2	0	0	2.5	0	0	0	2
B.	(using word count) old	blue	navy	sweater	trendy	wood	chair	discount	jeans
TFIDF _q	1	0	0	0	0	0	0	0	1
TFIDF _q	2	0	0	0	0	0	0	0	2.5

$$\text{distance}(Q, B) = \sqrt{(2-1)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2}$$

$$= \sqrt{1+1+1+1}$$

$$= 2$$

TF _B	2	1	1	0	0	0	1	1
TF _q	1	0	0	0	0	0	0	1

query: OLD JEANS

$$\begin{aligned}
 \text{distance}(Q, C) &= \\
 &= \sqrt{(1-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2} \\
 &= \sqrt{1+1+1+1+1} \\
 &= \sqrt{5} \\
 &= 2.23
 \end{aligned}$$

$TF_C = 110100010$
$TF_q = 100000001$

Since $\text{distance}(Q, B) < \text{distance}(Q, C)$, recommend product B.

C. query: OLD JEANS (changed from original query)

$$\cos(Q, A) = \frac{Q \cdot A}{\|Q\| \times \|A\|}$$

$$Q \cdot A \approx 0 \quad \text{dot product is } 0, \text{ so similarity is also } 0$$

$$\cos_{sim}(A, Q) = 0$$

$$\cos_{sim}(B, Q) = \frac{Q \cdot B}{\|Q\| \cdot \|B\|}$$

$$\text{Norm: } \sqrt{x_1^2 + x_2^2 + \dots + x_i^2} \quad (\text{definition})$$

$$\begin{aligned}
 \|Q\| &= \sqrt{2^2 + (2.5)^2} \\
 &= \sqrt{10.25} = 3.
 \end{aligned}$$

TFIDF_B 4 2 2.5 0 0 0 0 2 2.5

from previous question 2

$$\|B\| = \sqrt{4^2 + 2^2 + 2.5^2 + 0^2 + 2.5^2}$$

$$\begin{aligned}\|B\| &= \sqrt{16 + 4 + 6.25 + 0 + 6.25} \\ &= \sqrt{36.5} = \approx 6\end{aligned}$$

TFIDF_Q 2 0 0 0 0 0 0 0 2.5

from above

$$\begin{aligned}Q \cdot B &= 4 \cdot 2 + 2.5 \cdot 2.5 \quad \leftarrow \text{dot product of } Q \text{ and } B. \\ &= 8 + 6.25 = 14\end{aligned}$$

$$\cos_{sim}(Q, B) = \frac{14}{3 \times 6} = .875$$

$$\cos_{sim}(Q, B) > \cos(Q, A)$$

Therefore, recommend product B.

d. Reasons why cosine similarity is used over Euclidean distance:

① accounts for different document lengths

② uses dot product, which handles high-dimensional data better.

3. a.

- I. accuracy = $\boxed{11/15}$
- II. precision = $\boxed{2/4}$
- III. recall = $\boxed{2/4}$

- IV. $F1 \text{ score} = 2 \cdot \frac{\text{Precision} \times \text{recall}}{\text{precision} + \text{recall}}$

$$= 2 \cdot \frac{1/4}{1/2 + 1/2} = \boxed{1/2}$$

b.

		T	NT	
		T	2	2
Actual	T	2	9	
	NT			
Predicted				

c. Recall: the model should be more accepting of false positives in order to catch more threats, since the cost of a false negative is extremely high.

d. Any model getting below 73.3% ($11/15$) accuracy is more or less useless; since you can achieve this accuracy by simply predicting "no threat" to every data point.