Naive Bayes Learning

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(Rain \mid Wet) = \frac{P(W|R)P(R)}{P(W)}$$

fact → $P(W)$ เปลี่ยนตามสถานการณ์ เช่น ฤดูฝน ค่านี้น้อยมาก เพราะตกวันเดียว

ทราบว่าฟันเปียก ฝนตกด้วยค่าความน่าจะเป็นเท่าไร

$$P(Covid \mid +) = \frac{P(+|Covid)P(Covid)}{P(+)}$$

$P(+)$ → จับทุกคนตรวจ ATK แล้วเป็น + ซึ่งเราไม่รู้ค่า

$$P(\neg Covid \mid +) = \frac{P(+|\neg Covid)P(\neg Covid)}{P(+)}$$

Posterior Hypothesis จับ 2 อันเทียบกัน

100 คน เป็น + 99 คน

100 คนที่ไม่เป็น หลุด 3 คน

$P(+|Covid)P(Covid)$

$(0.99)(0.01)$

$= 0.0099$

$P(+|\neg Covid)P(\neg Covid)$

ลบทิ้ง $P(Covid)$ ได้

$(0.03)(0.99)$

$= 0.0297$ ※

ผล ไม่ผิดน้อยมาก แต่ $P(Covid)$ ที่เปลี่ยนตามสถานการณ์แตกต่างมาก

ถ้าหมอการเมืองโกงกิน จะตกนรก

$P(Chest \mid Hell)$

$$\frac{P(H \mid C)\,P(C)}{P(H)}$$

$(0.95)(0.9)$

$= 0.855$

$P(\neg Chest \mid Hell)$

อาจตกนรกเพราะสาเหตุอื่น

$$\frac{P(H \mid \neg C)\,P(\neg C)}{P(H)}$$

$(0.2)(0.1)$

$= 0.02$

$(0.95)(x) = (0.2)(1-x)$

$0.95x = 0.2 - 0.2x$

$1.15x = 0.2$

$x = \frac{0.2}{1.15}$

ต้องไม่ถึง 20 % ถึงจะเชื่อได้ว่า ไม่ได้ทำผิด เมืองน่าอยู่

	0.7 | 0.3	0.8 | 0.2
ถูกใส่ร้าย	0.56	0.14
ผิดจริง	0.01	0.24

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

ถ้า 2 เหตุการณ์ที่อิสระต่อกันเกิดพร้อมกัน

ค่าความน่าจะเป็นคือ 2 เหตุการณ์คูณกัน

D15 Overcast Mild Normal Weak ?

$P(Y \mid \text{overcast}, \text{Mild}, \text{Normal}, \text{Weak})$

$P(N \mid \text{overcast}, \text{Mild}, \text{Normal}, \text{Weak})$

$P(Y \mid \text{overcast}, \text{Mild}, \text{Normal}, \text{Weak})$

$$\frac{P(\text{overcast, mild, Normal, weak} \mid Y)\,P(Y)}{P(\text{overcast, Mild, Normal, Weak})}$$

$P(N \mid \text{overcast}, \text{Mild}, \text{Normal}, \text{Weak})$

$$\frac{P(\text{overcast, mild, Normal, weak} \mid N)\,P(N)}{P(\text{overcast, Mild, Normal, Weak})}$$

$$P(\text{overcast} \mid Y)\,P(\text{Mild} \mid Y)\,P(\text{Normal} \mid Y)\,P(\text{Weak} \mid Y)\,P(Y)$$

$$= \left(\frac{4+1}{9+1}\right)\left(\frac{4+1}{9+1}\right)\left(\frac{4+1}{9+1}\right)\left(\frac{6+1}{9+1}\right)\left(\frac{9+1}{14+1}\right)$$

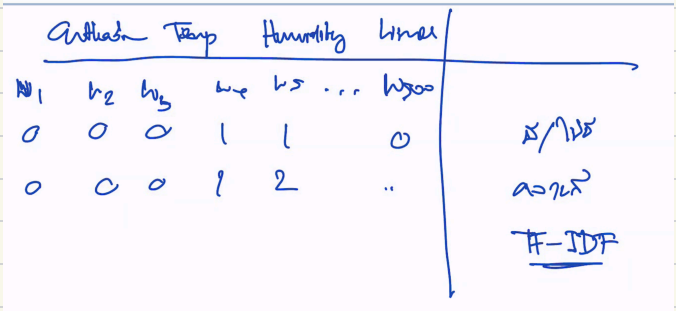
$$P(\text{overcast} \mid N)\,P(\text{Mild} \mid N)\,P(\text{Normal} \mid N)\,P(\text{Weak} \mid N)\,P(N)$$

$$= \left(\frac{0+1}{5+1}\right)\left(\frac{2+1}{5+1}\right)\left(\frac{1+1}{5+1}\right)\left(\frac{2+1}{5+1}\right)\left(\frac{5+1}{14+1}\right)$$

$$\frac{0 + 0.05}{5 + 0.05}$$

TF - IDF

└ ความถี่ที่ไม่สำคัญ




```python
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score

from sklearn.model_selection import train_test_split

vectorizer=CountVectorizer(tokenizer=lambda x: x.split())
tfidf=TfidfTransformer()
X=vectorizer.fit_transform(pos_data)
X=tfidf.fit_transform(X)
X_train, X_test, Y_train, Y_test = train_test_split(X,pos_target, test_size = 0.10, random_state = 42)
clf=MultinomialNB().fit(X_train,Y_train)
Y_predict = clf.predict(X_test)
accuracy_score(Y_test, Y_predict)
```

```python
X.todense()
```

```
matrix([[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1],
        [0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0],
        [1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0]])
```

Not TF - IDF

```python
Y.todense()
```

ลดความสำคัญลง

```
matrix([[0.25117442, 0.        , 0.33026418, 0.33026418, 0.25117442,
         0.25117442, 0.19505935, 0.33026418, 0.25117442, 0.25117442,
         0.        , 0.25117442, 0.        , 0.19505935, 0.33026418,
         0.33026418],
        [0.        , 0.40786601, 0.        , 0.        , 0.31019261,
         0.        , 0.24089223, 0.        , 0.31019261, 0.31019261,
         0.40786601, 0.31019261, 0.40786601, 0.24089223, 0.        ,
         0.        ],
        [0.44652407, 0.        , 0.        , 0.        , 0.        ,
         0.44652407, 0.34676577, 0.        , 0.        , 0.        ,
```

Use TF - IDF