

# **Naïve Bayes Classifier**

# Outline

- Background
- Probability Basics
- Probabilistic Classification
- Naïve Bayes with and example
- Example: Play Tennis
- Relevant Issues
- Conclusions

# Background

- There are three methods to establish a classifier
  - a) Model a classification rule directly*  
Examples: k-NN, decision trees, perceptron, SVM
  - b) Model the probability of class memberships given input data*  
Example: multi-layered perceptron with the cross-entropy cost
  - c) Make a probabilistic model of data within each class*  
Examples: naive Bayes, model based classifiers
- *a)* and *b)* are examples of **discriminative** classification
- *c)* is an example of **generative** classification
- *b)* and *c)* are both examples of **probabilistic** classification

# Probability Basics

- Prior, conditional and joint probability
  - Prior probability:  $P(X)$
  - Conditional probability:  $P(X_1 | X_2), P(X_2 | X_1)$
  - Joint probability:  $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship:  $P(X_1, X_2) = P(X_2 | X_1)P(X_1) = P(X_1 | X_2)P(X_2)$
  - Independence:  $P(X_2 | X_1) = P(X_2), P(X_1 | X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$
- Bayesian Rule

$$P(C | \mathbf{X}) = \frac{P(\mathbf{X} | C)P(C)}{P(\mathbf{X})}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

# Naïve Bayes

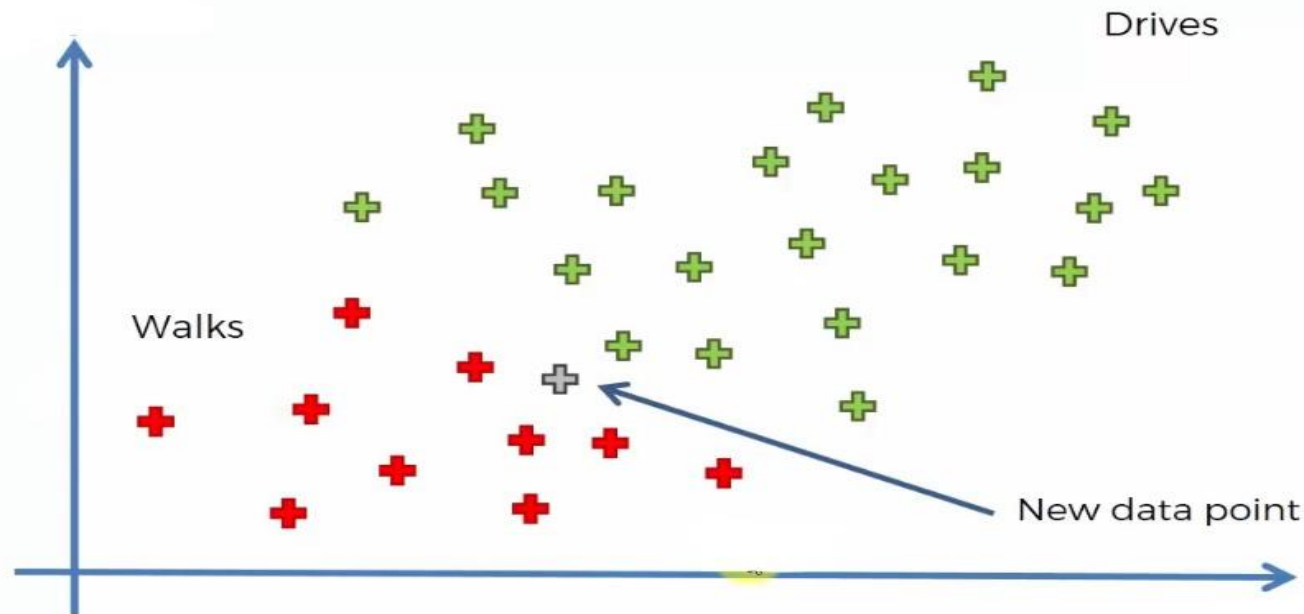
- **Naïve Bayes** is a classification algorithm that works based on the Bayes theorem.
- Bayes theorem is used to find the probability of a hypothesis with given evidence.

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

- In this, we can find the probability of A, given that B occurred. A is the hypothesis and B is the evidence.
- $P(B|A)$  is the probability of B given that A is True.
- $P(A)$  and  $P(B)$  is the independent probabilities of A and B.

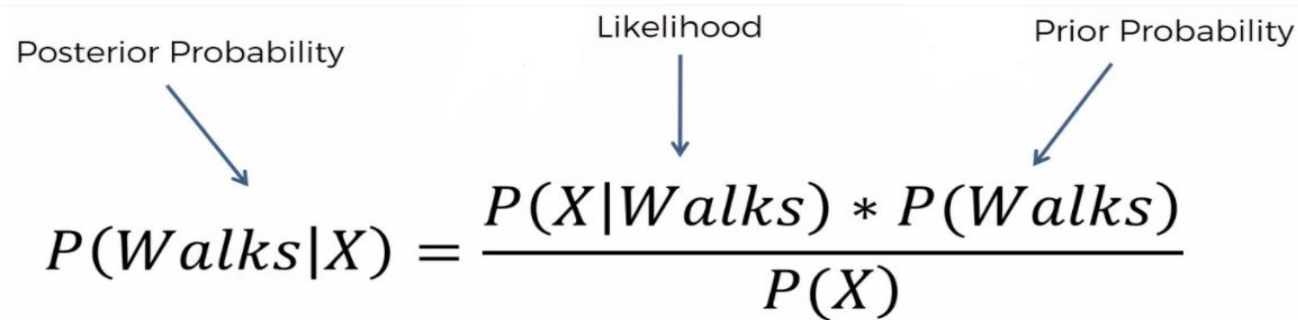
# Naïve Bayes

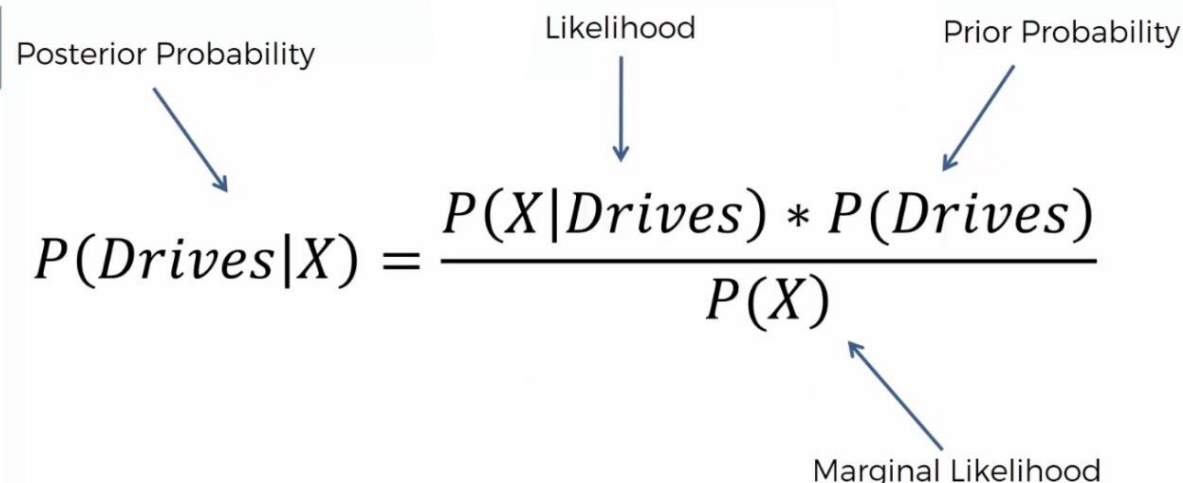
- For example, We have a dataset of employees in a company, our aim is to create a model to find whether a person is going to the office by driving or walking using salary and age of the person.
- Red points belongs to the person who walk and green points belongs to the persons who drive to their offices.
- What is the category of the new point ??



# Naïve Bayes

- Find posterior probability of walking and driving for this data point.
- The posterior probability of walking and driving for the new data point is :


$$P(Walks|X) = \frac{P(X|Walks) * P(Walks)}{P(X)}$$


$$P(Drives|X) = \frac{P(X|Drives) * P(Drives)}{P(X)}$$

# Naïve Bayes

## Steps :

**Step 1: We have to find all the probabilities required for the Bayes theorem for the calculation of posterior probability**

- $P(\text{Walks})$  is simply the probability of those who walk among all

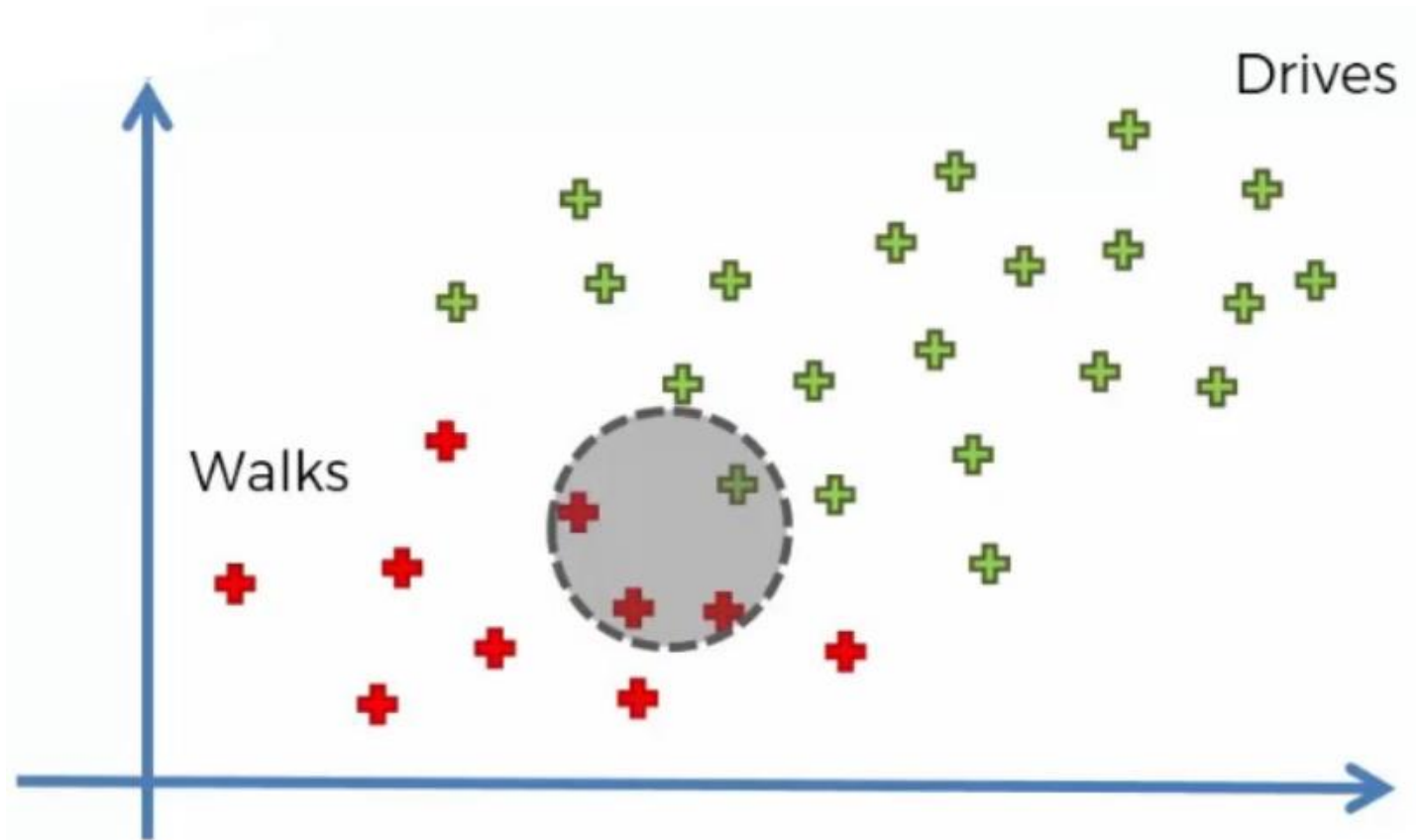
$$P(\text{Walks}) = \text{Number of Walkers} / \text{Total Observations}$$

$$P(\text{Walks}) = 10/30$$

- In order to find the marginal likelihood,  $P(X)$ , we have to consider a circle around the new data point of any radii including some red and green points.



# Naïve Bayes



**$P(X) = \text{Number of similar Observations} / \text{Total Observations}$**

**$P(X) = 4/30$**

# Naïve Bayes

$P(X|\text{Walks})$  can be found by :

$P(X|\text{Walks}) = \text{Number of similar observations among those who walk} / \text{Total number of walkers}$

$$P(X|\text{Walks}) = 3/10$$

Now, posterior probability can be calculated as follows:

$$P(\text{Walks}|X) = (3/10 * 10/30) / 4/30 = 0.75$$

**Step 2: Similarly we can find the posterior probability of Driving, and it is 0.25**

# Naïve Bayes

- **Step 3:** Compare both posterior probabilities.  
 $P(\text{walks}|X)$  has greater values and the new point belongs to the walking category.

# Probabilistic Classification

- Establishing a probabilistic model for classification

- Discriminative model

$$P(C | \mathbf{X}) \quad C = c_1, \dots, c_L, \mathbf{X} = (X_1, \dots, X_n)$$

- Generative model

$$P(\mathbf{X} | C) \quad C = c_1, \dots, c_L, \mathbf{X} = (X_1, \dots, X_n)$$

- MAP classification rule

- **MAP: Maximum A Posterior**

- Assign  $x$  to  $c^*$  if  $P(C = c^* | \mathbf{X} = \mathbf{x}) > P(C = c | \mathbf{X} = \mathbf{x}) \quad c \neq c^*, c = c_1, \dots, c_L$

- Generative classification with the MAP rule

- Apply Bayesian rule to convert: 
$$P(C | \mathbf{X}) = \frac{P(\mathbf{X} | C)P(C)}{P(\mathbf{X})} \propto P(\mathbf{X} | C)P(C)$$

# Naïve Bayes

- Bayes classification

$$P(C | \mathbf{X}) \propto P(\mathbf{X} | C)P(C) = P(X_1, \dots, X_n | C)P(C)$$

Difficulty: learning the joint probability  $P(X_1, \dots, X_n | C)$

- Naïve Bayes classification

- Making the assumption that **all input attributes are independent**

$$\begin{aligned} P(X_1, X_2, \dots, X_n | C) &= \underline{P(X_1 | X_2, \dots, X_n; C)} P(X_2, \dots, X_n | C) \\ &= \underline{P(X_1 | C)} \underline{P(X_2, \dots, X_n | C)} \\ &= \underline{P(X_1 | C)} \underline{P(X_2 | C)} \dots \underline{P(X_n | C)} \end{aligned}$$

- MAP classification rule

$$[P(x_1 | c^*) \dots P(x_n | c^*)]P(c^*) > [P(x_1 | c) \dots P(x_n | c)]P(c), \quad c \neq c^*, c = c_1, \dots, c_L$$

# Naïve Bayes

- Naïve Bayes Algorithm (for discrete input attributes)

- **Learning Phase:** Given a training set  $S$ ,

For each target value of  $c_i$  ( $c_i = c_1, \dots, c_L$ )

$\hat{P}(C = c_i) \leftarrow$  estimate  $P(C = c_i)$  with examples in  $S$ ;

For every attribute value  $a_{jk}$  of each attribute  $x_j$  ( $j = 1, \dots, n; k = 1, \dots, N_j$ )

$\hat{P}(X_j = a_{jk} | C = c_i) \leftarrow$  estimate  $P(X_j = a_{jk} | C = c_i)$  with examples in  $S$ ;

Output: conditional probability tables; for  $x_j$ ,  $N_j \times L$  elements

- **Test Phase:** Given an unknown instance  $\mathbf{X}' = (a'_1, \dots, a'_n)$

Look up tables to assign the label  $c^*$  to  $\mathbf{X}'$  if

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_n | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_n | c)] \hat{P}(c), \quad c \neq c^*, c = c_1, \dots, c_L$$

# Example

- Example: Play Tennis

## *PlayTennis: training examples*

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

# Example

- Learning Phase

Outlook	Play=Yes	Play=No
<i>Sunny</i>	2/9	3/5
<i>Overcast</i>	4/9	0/5
<i>Rain</i>	3/9	2/5

Temperature	Play=Yes	Play=No
<i>Hot</i>	2/9	2/5
<i>Mild</i>	4/9	2/5
<i>Cool</i>	3/9	1/5

Humidity	Play=Yes	Play=No
<i>High</i>	3/9	4/5
<i>Normal</i>	6/9	1/5

Wind	Play=Yes	Play=No
<i>Strong</i>	3/9	3/5
<i>Weak</i>	6/9	2/5

$$P(\text{Play=Yes}) = 9/14 \quad P(\text{Play=No}) = 5/14$$



# Example

- Test Phase

- Given a new instance,

$\mathbf{x}' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

- Look up tables

$$P(\text{Outlook}=\text{Sunny} \mid \text{Play}=\text{Yes}) = 2/9$$

$$P(\text{Temperature}=\text{Cool} \mid \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Humidity}=\text{High} \mid \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Wind}=\text{Strong} \mid \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Outlook}=\text{Sunny} \mid \text{Play}=\text{No}) = 3/5$$

$$P(\text{Temperature}=\text{Cool} \mid \text{Play}=\text{No}) = 1/5$$

$$P(\text{Humidity}=\text{High} \mid \text{Play}=\text{No}) = 4/5$$

$$P(\text{Wind}=\text{Strong} \mid \text{Play}=\text{No}) = 3/5$$

$$P(\text{Play}=\text{No}) = 5/14$$

- MAP rule

$$P(\text{Yes} \mid \mathbf{x}'): [P(\text{Sunny} \mid \text{Yes})P(\text{Cool} \mid \text{Yes})P(\text{High} \mid \text{Yes})P(\text{Strong} \mid \text{Yes})] P(\text{Play}=\text{Yes}) = 0.0053$$

$$P(\text{No} \mid \mathbf{x}'): [P(\text{Sunny} \mid \text{No})P(\text{Cool} \mid \text{No})P(\text{High} \mid \text{No})P(\text{Strong} \mid \text{No})]P(\text{Play}=\text{No}) = 0.0206$$

- Given the fact  $P(\text{Yes} \mid \mathbf{x}') < P(\text{No} \mid \mathbf{x}')$ , we label  $\mathbf{x}'$  to be “No”.

# Relevant Issues

- Violation of Independence Assumption
  - For many real world tasks,  $P(X_1, \dots, X_n | C) \neq P(X_1 | C) \dots P(X_n | C)$
  - Nevertheless, naïve Bayes works surprisingly well anyway!
- Zero conditional probability Problem
  - If no example contains the attribute value  $X_j = a_{jk}$ ,  $\hat{P}(X_j = a_{jk} | C = c_i) = 0$
  - In this circumstance,  $\hat{P}(x_1 | c_i) \dots \hat{P}(a_{jk} | c_i) \dots \hat{P}(x_n | c_i) = 0$  during test
  - For a remedy, conditional probabilities estimated with

$$\hat{P}(X_j = a_{jk} | C = c_i) = \frac{n_c + mp}{n + m}$$

$n_c$  : number of training examples for which  $X_j = a_{jk}$  and  $C = c_i$

$n$  : number of training examples for which  $C = c_i$

$p$  : prior estimate (usually,  $p = 1/t$  for  $t$  possible values of  $X_j$ )

$m$  : weight to prior (number of "virtual" examples,  $m \geq 1$ )

# Relevant Issues

- Continuous-valued Input Attributes
  - Conditional probability modeled with the normal/ Gaussian distribution

$$\hat{P}(X_j | C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

$\mu_{ji}$  : mean(average) of attribute values  $X_j$  of examples for which  $C = c_i$

$\sigma_{ji}$  : standard deviation of attribute values  $X_j$  of examples for which  $C = c_i$

- Learning Phase: for  $\mathbf{X} = (X_1, \dots, X_n)$ ,  $C = c_1, \dots, c_L$   
Output:  $n \times L$  normal distributions and  $P(C = c_i) \ i = 1, \dots, L$
- Test Phase: for  $\mathbf{X}' = (X'_1, \dots, X'_n)$ 
  - Calculate conditional probabilities with all the normal distributions
  - Apply the MAP rule to make a decision

# **Continuous-valued Input Attributes**

# Continuous-valued Input Attributes

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

# Calculating Mean and Variance

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

$$P(\text{Male}) = 4/8 = 0.5$$

$$P(\text{Female}) = 4/8 = 0.5$$

**Male:**

$$\text{Mean (Height)} = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$\begin{aligned}\text{Variance (Height)} &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{(6-5.855)^2 + (5.92-5.855)^2 + (5.58-5.855)^2 + (5.92-5.855)^2}{4-1} \\ &= 0.035055\end{aligned}$$

# Calculating Posterior Probability

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

$$P(\text{Male}) = 4/8 = 0.5$$

$$P(\text{Female}) = 4/8 = 0.5$$

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$\text{Posterior (Male)} = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{\text{Evidence}}$$

$$\text{Posterior (Female)} = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{\text{Evidence}}$$

Gaussian Distribution Equation

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



# Calculating Posterior Probability

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$

$$P(W|M) = 5.9881e^{-6}$$

$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$

$$P(W|F) = 1.6789e^{-2}$$

$$P(FS|F) = 2.8669e^{-1}$$

$$\text{Posterior (Male)} = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{\text{Evidence}} = 0.5 * 1.5789 * 5.9881e^{-6} * 1.3112e^{-3} = 6.1984e^{-9}$$

$$\text{Posterior (Female)} = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{\text{Evidence}} = 0.5 * 2.2346e^{-1} * 1.6789e^{-2} * 2.8669e^{-1} = 5.377e^{-4}$$



# Conclusions

- Naïve Bayes based on the independence assumption
  - Training is very easy and fast; just requiring considering each attribute in each class separately
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- A popular generative model
  - Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  - Many successful applications, e.g., spam mail filtering
  - Apart from classification, naïve Bayes can do more...

# References

- <http://intranet.cs.man.ac.uk/mlo/comp20411/>
- <https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/>