Bayesian Network

Zahoor Tanoli (PhD)

CUI Attock Campus

Naïve Bayes

- Founded on Bayes' Rule for probabilistic inference

- Update probability of hypotheses based on evidence

- Choose hypothesis with the maximum probability after the evidence has been incorporated

- Algorithm is particularly useful for domains with lots of features

Bayesian Classifier – Basic Equation

Class Prior Probability

Descriptor Posterior Probability

$$P(C \mid X) = \frac{P(C)\,P(X \mid C)}{P(X)}$$

Class Posterior Probability

Descriptor Prior Probability

Bayes' Rule

$$\boxed{p(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}}$$

<u>Who is who in Bayes' rule</u>

$P(h)$: prior belief (probability of hypothesis *h* before seeing any data)

$P(d \mid h)$: likelihood (probability of the data if the hypothesis *h* is true)

$P(d) = \sum_h P(d \mid h)P(h)$: data evidence (marginal probability of the data)

$P(h \mid d)$: posterior (probability of hypothesis *h* after having seen the data *d*)

Example

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 1/5$: fraction of all positive examples that have feature 1 on
$R_1(0,1) = 4/5$: fraction of all positive examples that have feature 1 off

Example

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 1/5$: fraction of all positive examples that have feature 1 on
$R_1(0,1) = 4/5$: fraction of all positive examples that have feature 1 off

$R_1(1,0) = 5/5$: fraction of all negative examples that have feature 1 on
$R_1(0,0) = 0/5$: fraction of all negative examples that have feature 1 off

Example

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 1/5$ $R_1(0,1) = 4/5$
$R_1(1,0) = 5/5$ $R_1(0,0) = 0/5$

$R_2(1,1) = 1/5$ $R_2(0,1) = 4/5$
$R_2(1,0) = 2/5$ $R_2(0,0) = 3/5$

$R_3(1,1) = 4/5$ $R_3(0,1) = 1/5$
$R_3(1,0) = 1/5$ $R_3(0,0) = 4/5$

$R_4(1,1) = 2/5$ $R_4(0,1) = 3/5$
$R_4(1,0) = 4/5$ $R_4(0,0) = 1/5$

Prediction

These R values actually represent the hypothesis and is used to classify the new input.

$R_1(1,1) = 1/5$	$R_1(0,1) = 4/5$
$R_1(1,0) = 5/5$	$R_1(0,0) = 0/5$
$R_2(1,1) = 1/5$	$R_2(0,1) = 4/5$
$R_2(1,0) = 2/5$	$R_2(0,0) = 3/5$
$R_3(1,1) = 4/5$	$R_3(0,1) = 1/5$
$R_3(1,0) = 1/5$	$R_3(0,0) = 4/5$
$R_4(1,1) = 2/5$	$R_4(0,1) = 3/5$
$R_4(1,0) = 4/5$	$R_4(0,0) = 1/5$

New $x = <0,0,1,1>$

$S(1) = R_1(0,1) * R_2(0,1) * R_3(1,1) * R_4(1,1) = .205$

$S(0) = R_1(0,0) * R_2(0,0) * R_3(1,0) * R_4(1,0) = 0$

$S(1) > S(0)$, so predict class 1

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

Bayesian Classifier – Modified Basic Equation

$$\underbrace{P(C \mid X)}_{\textbf{Posterior Probability}} = \frac{\overbrace{\boxed{P(C)}}^{\textbf{Prior Probability}}\ \overbrace{P(X \mid C)}^{\textbf{Likelihood}}}{\underbrace{\boxed{P(X)}}_{\textbf{Normalization Factor}}}$$

Bayesian Classifier - Probabilities for the weather data

Frequency Tables

Outlook	No	Yes
Sunny	3	2
Overcast	0	4
Rainy	2	3

Temp.	No	Yes
Hot	2	2
Mild	2	4
Cool	1	3

Humidity	No	Yes
High	4	3
Normal	1	6

Windy	No	Yes
False	2	6
True	3	3

Outlook	No	Yes
Sunny	3/5	2/9
Overcast	0/5	4/9
Rainy	2/5	3/9

Temp.	No	Yes
Hot	2/5	2/9
Mild	2/5	4/9
Cool	1/5	3/9

Humidity	No	Yes
High	4/5	3/9
Normal	1/5	6/9

Windy	No	Yes
False	2/5	6/9
True	3/5	3/9

Likelihood Tables

Bayesian Classifier – Predicting a new day

Outlook	Temp.	Humidity	Windy	Play
sunny	cool	high	true	?

P(**yes**|**X**) = p(sunny|yes) x p(cool|yes) x p(high|yes) x p(true|yes) x p(yes)

= 2/9 x 3/9 x 3/9 x 3/9 x 9/14 = 0.0053 => 0.0053/(0.0053+0.0206) = 0.205

P(**no**|**X**) = p(sunny|no) x p(cool|no) x p(high|no) x p(true|no) x p(no)

= 3/5 x 1/5 x 4/5 x 3/5 x 5/14 = 0.0206=0.0206/(0.0053+0.0206) = 0.795

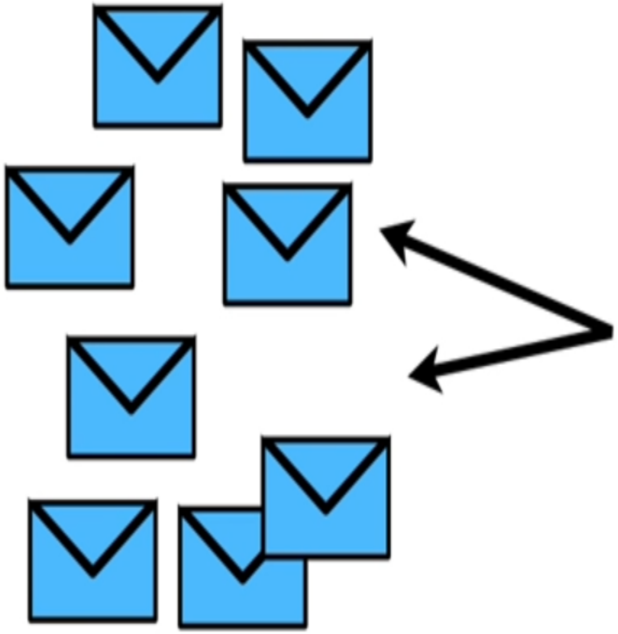
Outlook	No	Yes
Sunny	3/5	2/9
Overcast	0/5	4/9
Rainy	2/5	3/9

Temp.	No	Yes
Hot	2/5	2/9
Mild	2/5	4/9
Cool	1/5	3/9

Humidity	No	Yes
High	4/5	3/9
Normal	1/5	6/9

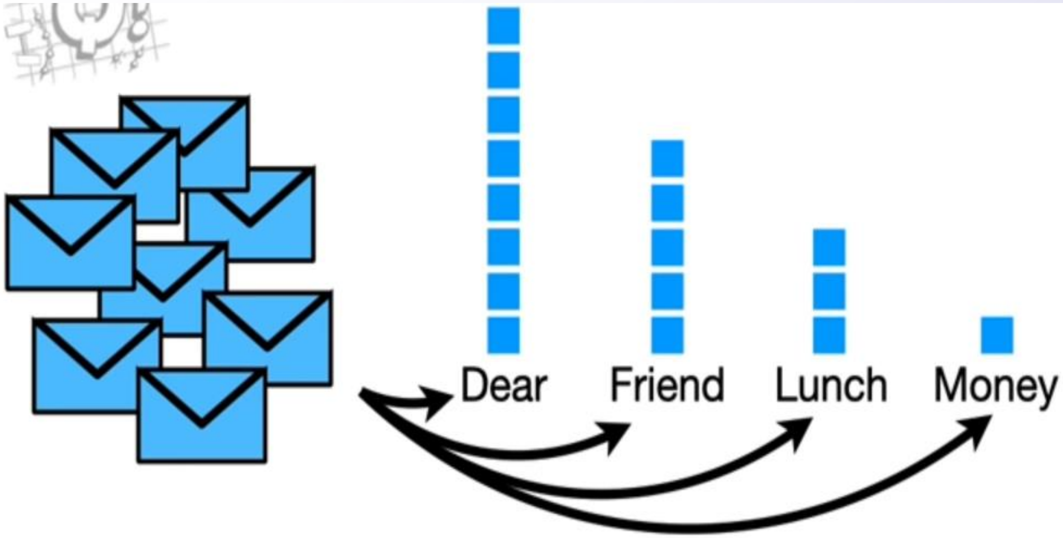
Windy	No	Yes
False	2/5	6/9
True	3/5	3/9

Spam Classification Example

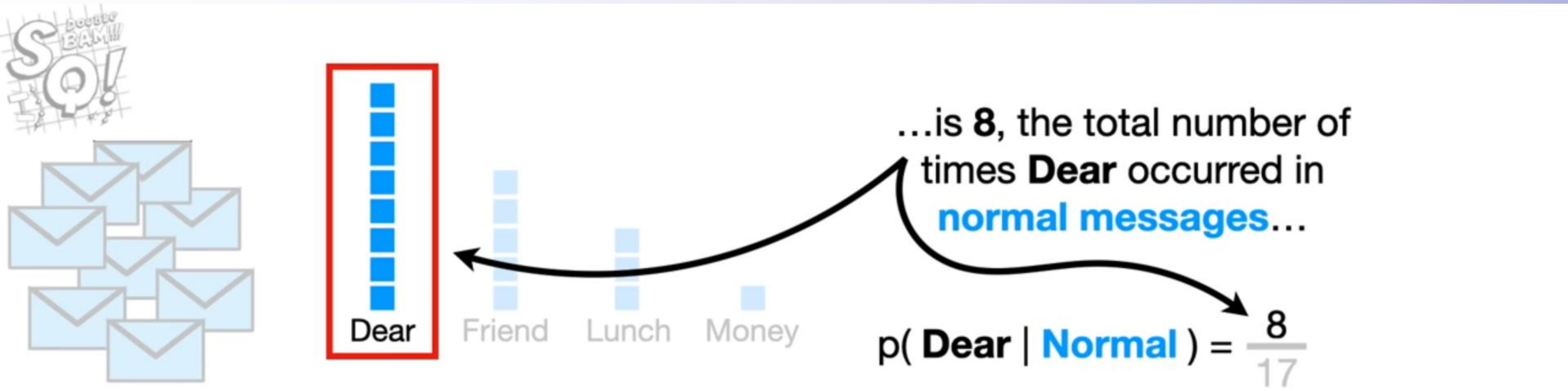


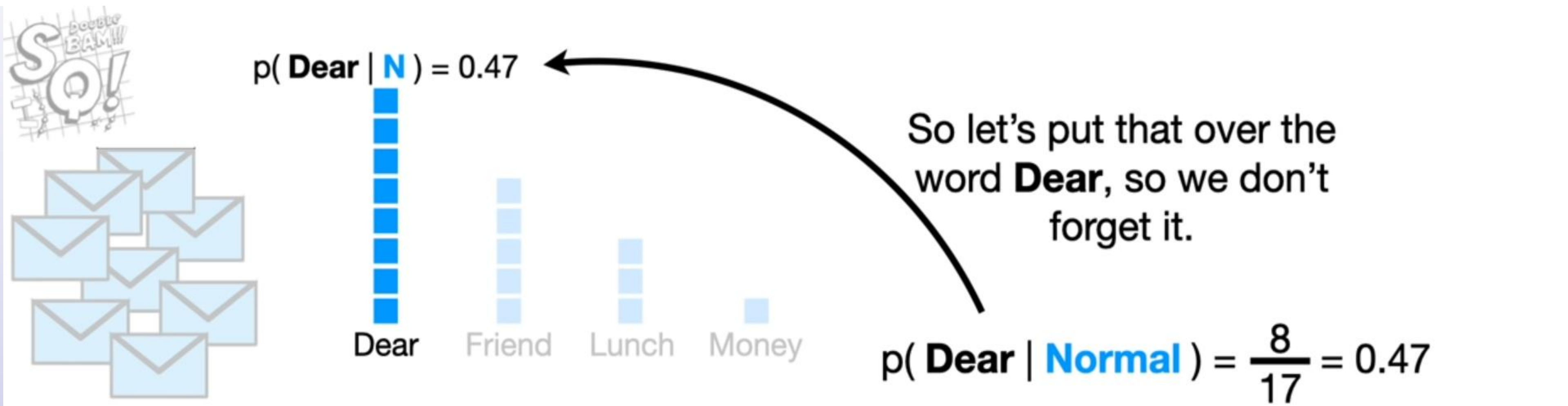
Imagine we received **normal messages** from friends and family...

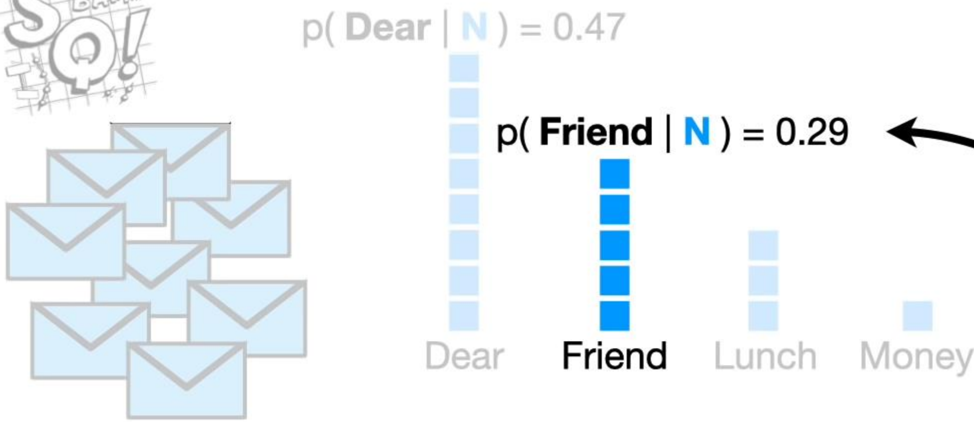
...and we also received **spam** (unwanted messages that are usually scams or unsolicited advertisements)...



So, the first thing we do is make a **histogram** of all the words that occur in the normal messages from friends and family.

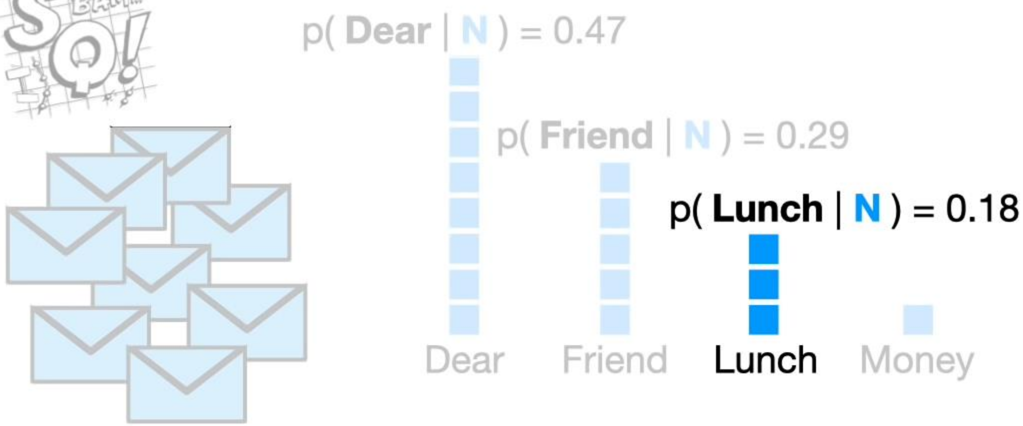

Dear
Friend
Lunch
Money
…is 8, the total number of times Dear occurred in normal messages…
p(Dear | Normal) = 8 / 17


p(Dear | N) = 0.47
Dear
Friend
Lunch
Money
So let's put that over the word Dear, so we don't forget it.
p(Dear | Normal) = 8 / 17 = 0.47



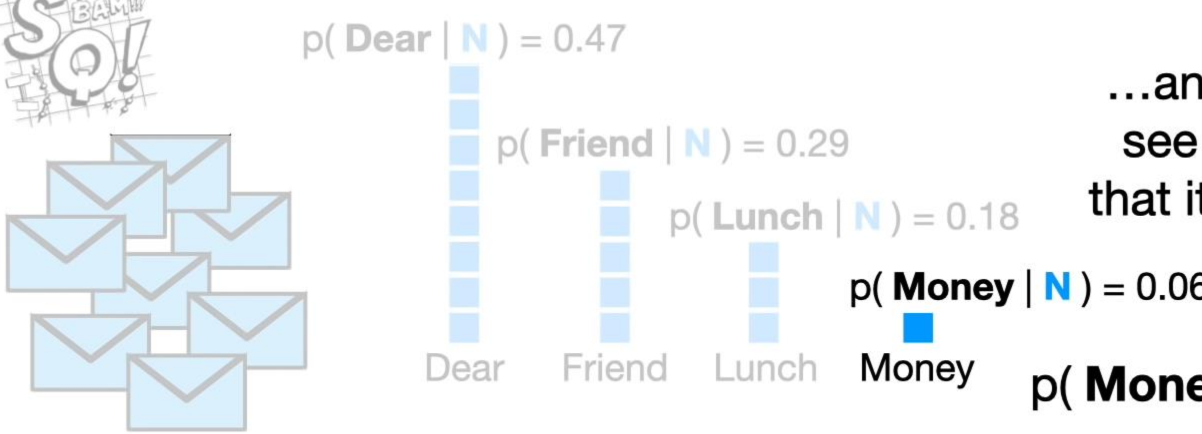
So let's put that over the word **Friend**, so we don't forget it.

$$p(\textbf{Friend} \mid \textbf{Normal}) = \frac{5}{17} = 0.29$$



Likewise, the probability that we see the word **Lunch**, given that it is in a **normal message** is **0.18**…

$$p(\textbf{Lunch} \mid \textbf{Normal}) = \frac{3}{17} = 0.18$$

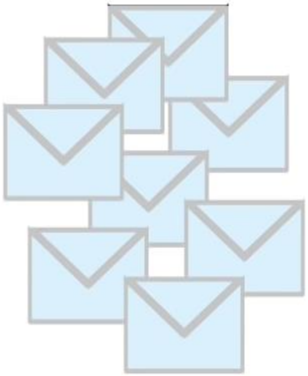


…and the probability that we see the word **Money**, given that it is in a **normal message** is **0.06**.

$$p(\textbf{Money} \mid \textbf{Normal}) = \frac{1}{17} = 0.06$$

$$p(\textbf{Dear} \mid \textbf{Spam}) = \frac{2}{7} \boxed{= 0.29}$$

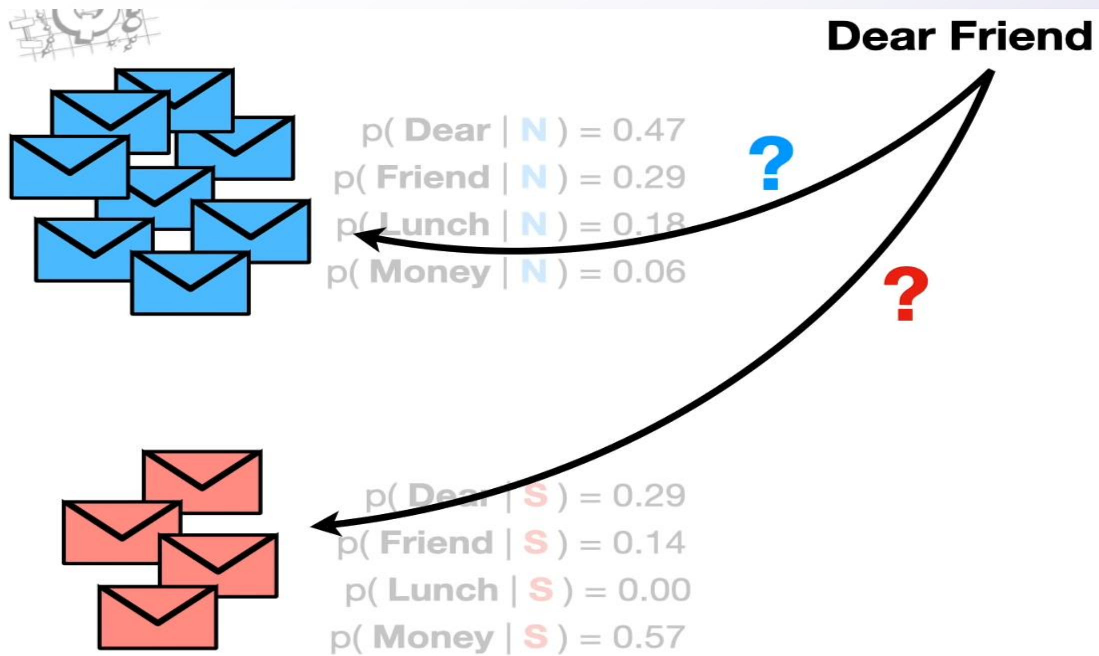
p(**Dear** | **S**) = 0.29



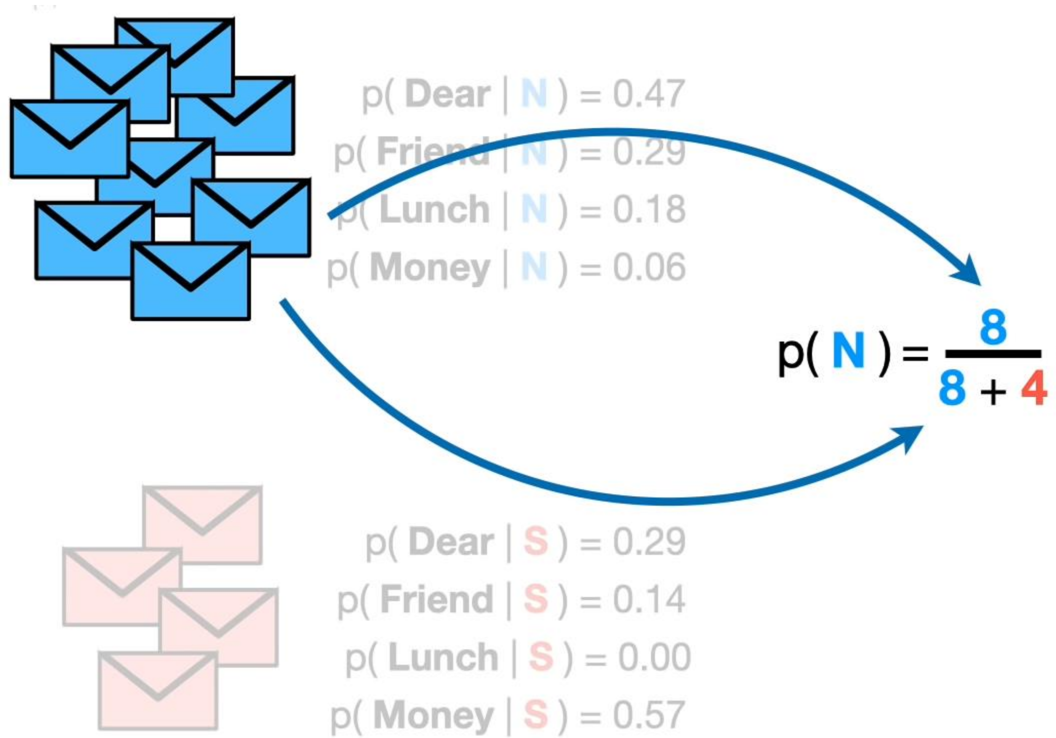
p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

Because we have calculated the probabilities of discrete, individual words, and not the probability of something continuous, like weight or height, these **Probabilities** are also called **Likelihoods**.

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57



And we want to decide if is a **normal message** or **spam**.



$$p(\ N\ ) = \frac{8}{8 + 4} = 0.67$$

For example, since **8** of the **12** messages are **normal messages**, our initial guess will be **0.67**.

TERMINOLOGY ALERT!!!!

The initial guess that we observe a **Normal** messages is called a **Prior Probability**.

Dear Friend

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

$\boxed{p(N) = 0.67}$

Now we just plug in the values that we worked out earlier and do the math…

$$p(\mathbf{N}) \times p(\mathbf{Dear} \mid \mathbf{N}) \times p(\mathbf{Friend} \mid \mathbf{N})$$

Dear Friend

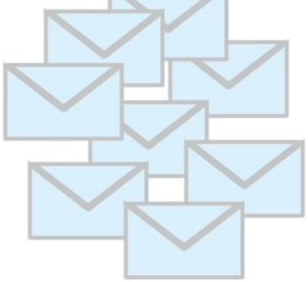
p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**N**) = 0.67

However, technically, it is *proportional* to the probability that the message is **normal**, given that it says **Dear Friend**.

$$0.67 \times 0.47 \times 0.29 = 0.09 \propto p(\mathbf{N} \mid \mathbf{Dear\ Friend})$$

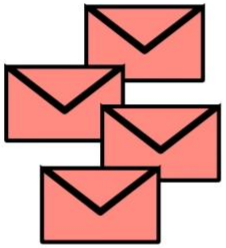
So let's put that under the **spam** so we don't forget it.



p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

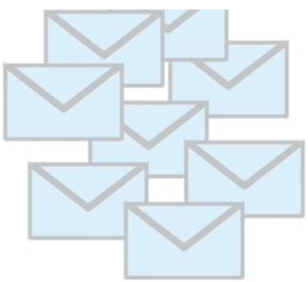
p(**N**) = 0.67

$$p(\mathbf{S}) = \frac{4}{4 + 8} = 0.33$$



p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

p(**S**) = 0.33



p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**N**) = 0.67

…and the probability that the word **Friend** occurs in **spam**.

p(**S**) x p(**Dear** | **S**) x p(**Friend** | **S**)

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00

$$0.33 \times 0.29 \times 0.14 \boxed{= 0.01}$$

Dear Friend

Then we did the math and decided that **Dear Friend** was a **normal message** because 0.09 > 0.01.

$p(\textbf{Dear} \mid N) = 0.47$

$p(\textbf{Friend} \mid N) = 0.29$

$p(\textbf{Lunch} \mid N) = 0.18$

$p(\textbf{Money} \mid N) = 0.06$

$p(N) = 0.67$

$$p(N) \times p(\textbf{Dear} \mid N) \times p(\textbf{Friend} \mid N) \boxed{= 0.09}$$

$$p(S) \times p(\textbf{Dear} \mid S) \times p(\textbf{Friend} \mid S) = 0.01$$

$p(\textbf{Dear} \mid S) = 0.29$

$p(\textbf{Friend} \mid S) = 0.14$

$p(\textbf{Lunch} \mid S) = 0.00$

$p(\textbf{Money} \mid S) = 0.57$

$p(S) = 0.33$

Bayesian Classifier – zero frequency problem

- What if a descriptor value doesn't occur with every class value

  P(Humidity=High|yes)=0

- Remedy: add 1 to the count for every descriptor-class combination (Laplace Estimator)

Outlook	No	Yes
Sunny	3+1	2+1
Overcast	0+1	4+1
Rainy	2+1	3+1

Temp.	No	Yes
Hot	2+1	2+1
Mild	2+1	4+1
Cool	1+1	3+1

Humidity	No	Yes
High	4+1	3+1
Normal	1+1	6+1

Windy	No	Yes
False	2+1	6+1
True	3+1	3+1

Bayesian Classifier – Missing values

- **Training**: instance is not included in frequency count for attribute value-class combination
- **Testing**: attribute will be omitted from calculation

Example:

Outlook	Temp.	Humidity	Windy	Play
?	cool	high	true	?

P(**yes**|X) = $\boxed{2/9}$ x 3/9 x 3/9 x 3/9 x 9/14 = 0.0238

P(**no**|X) = $\boxed{3/5}$ x 1/5 x 4/5 x 3/5 x 5/14 = 0.0343

P(**yes**|X) = 0.0238/(0.0238+0.0343) = 0.41

P(**no**|X) = 0.0343/(0.0238+0.0343) = 0.59

Laplace Correction

Avoid getting 1 or 0 as an answer

- $R_j(1,1) = \dfrac{\#(x_j^j = 1 \land y^j = 1) + 1}{\#(y^j = 1) + 2}$

- $R_j(0,1) = 1 - R_j(1,1)$

- $R_j(1,0) = \dfrac{\#(x_j^j = 1 \land y^j = 0) + 1}{\#(y^j = 0) + 2}$

- $R_j(0,0) = 1 - R_j(1,0)$

Example with Correction

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 2/7$ $R_1(0,1) = 5/7$
$R_1(1,0) = 6/7$ $\mathbf{R_1(0,0) = 1/7}$

$R_2(1,1) = 2/7$ $R_2(0,1) = 5/7$
$R_2(1,0) = 3/7$ $R_2(0,0) = 4/7$

$R_3(1,1) = 5/7$ $R_3(0,1) = 2/7$
$R_3(1,0) = 2/7$ $R_3(0,0) = 5/7$

$R_4(1,1) = 3/7$ $R_4(0,1) = 4/7$
$R_4(1,0) = 5/7$ $R_4(0,0) = 2/7$

Prediction with correction

$R_1(1,1) = 2/7$	$R_1(0,1) = 5/7$
$R_1(1,0) = 6/7$	$R_1(0,0) = 1/7$
$R_2(1,1) = 2/7$	$R_2(0,1) = 5/7$
$R_2(1,0) = 3/7$	$R_2(0,0) = 4/7$
$R_3(1,1) = 5/7$	$R_3(0,1) = 2/7$
$R_3(1,0) = 2/7$	$R_3(0,0) = 5/7$
$R_4(1,1) = 3/7$	$R_4(0,1) = 4/7$
$R_4(1,0) = 5/7$	$R_4(0,0) = 2/7$

New $x = \langle 0,0,1,1 \rangle$

$S(1) = R_1(0,1) * R_2(0,1) * R_3(1,1) * R_4(1,1) = .156$

$S(0) = R_1(0,0) * R_2(0,0) * R_3(1,0) * R_4(1,0) = .017$

$S(1) > S(0)$, so predict class 1

Bayesian Classifier in Medicine

$$P(Cold \mid Cough) = \frac{P(Cold)\,P(Cough \mid Cold)}{P(Cough)}$$

Total number of patients = 1000

Total number of patients with Cold = 400
P(Cold) = 400/1000 = 0.4

Total number of patients with Cough = 600
P(Cough) = 600/1000 = 0.6

Total number of patients with Cough & Cold = 120
P(Cough|Cold) = 120/400 = 0.3

P(Cold|Cough) = (0.4 * 0.3) / 0.6 = 0.2

Bayesian Classifier in Medicine

$$P(Flu \mid Cough) = \frac{P(Flu)\,P(Cough \mid Flu)}{P(Cough)}$$

Total number of patients = 1000

Total number of Flu = 200
P(Flu) = 200/1000 = 0.2

Total number of Cough = 600
P(Cough) = 600/1000 = 0.6

Total number of patients with Cough & Flu = 200
P(Cough|Flu) = 200/200 = 1.0

P(Flu|Cough) = (0.2 * 1.0) / 0.6 = 0.33

Does patient have cancer or not?

- A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases. Furthermore, only 0.008 of the entire population has this disease.

1. What is the probability that this patient has cancer?
2. What is the probability that he does not have cancer?
3. What is the diagnosis?

$$\left.\begin{array}{r} hypothesis1: 'cancer' \\ hypothesis2: '\neg cancer' \end{array}\right\} \quad hypothesis\ space\ H$$

$$-\, data: '+'$$

$$1. P(cancer \mid +) = \frac{P(+ \mid cancer)P(cancer)}{P(+)} = \frac{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} = \ldots\ldots\ldots\ldots$$

$$P(+ \mid cancer) = 0.98$$

$$P(cancer) = 0.008$$

$$P(+) = P(+ \mid cancer)P(cancer) + P(+ \mid \neg cancer)P(\neg cancer)$$

$$= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$P(+ \mid \neg cancer) = 0.03$$

$$P(\neg cancer) = \ldots\ldots\ldots\ldots$$

$$2. P(\neg cancer \mid +) = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$3. Diagnosis ??$$

Bayesian Classifier - Discussion

- One of the most used model in AI and ML
- However adding too many redundant variables will cause problem
- Works with any data type (nominal and numerical)
- Show a clear statistical picture of the descriptor-class or input/output relationship