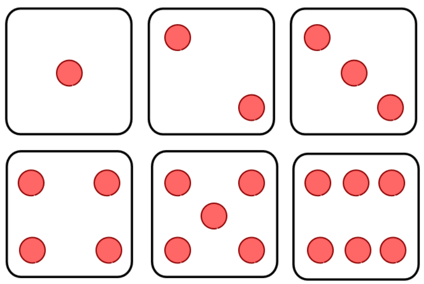
Lecture 6: Probabilistic Classifiers (Naive Bayes)

Naive Bayes Algorithm

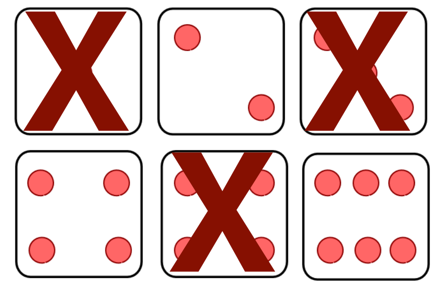
but before that…

Probability Theory - Review

Review: Probability Theory

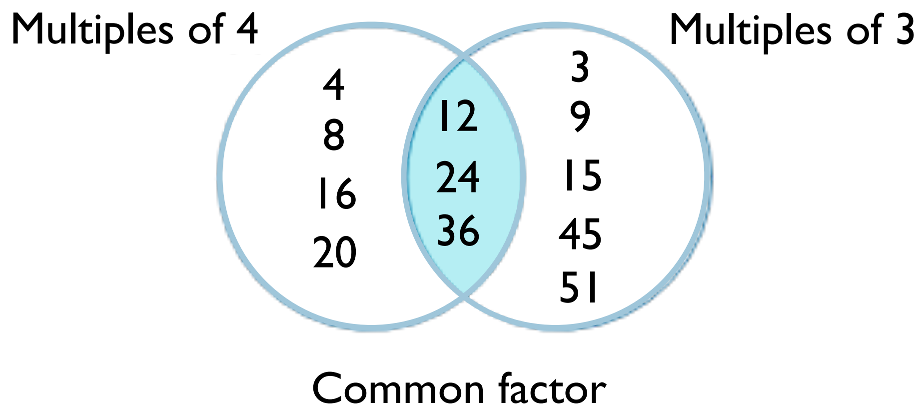


Die: probability of rolling a 2 **= 1/6**



Die: probability of rolling a 2
given that it's a special die
that only produces even numbers **= 1/3**

Review: Probability Theory

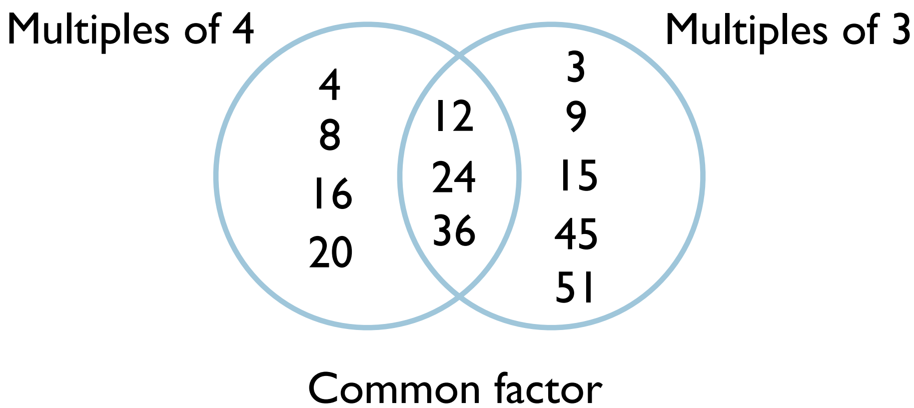


Pr(Multiple of 4) = 7/12

Pr(Multiple of 3) = 8/12

Pr(Multiple of 4 and Multiple of 3) = 3/12

Review: Probability Theory



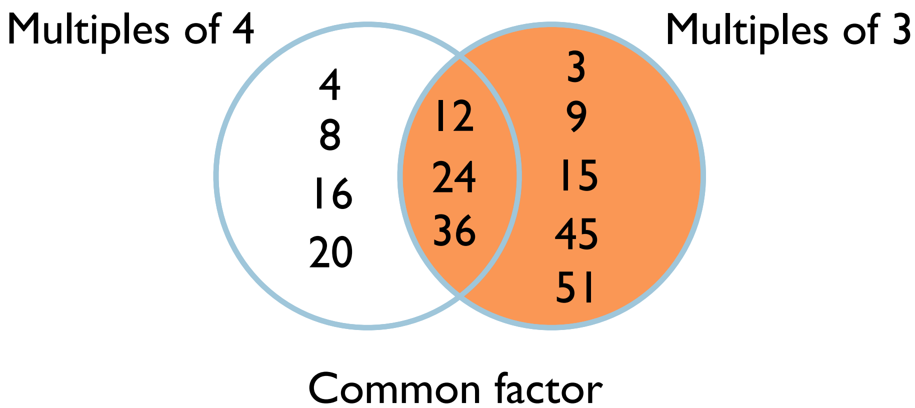
Pr(Multiple of 4) = 7/12

Pr(Multiple of 3) = 8/12

Pr(Multiple of 4 and Multiple of 3) = 3/12

Pr(Multiple of 4 | Multiple of 3)

Review: Probability Theory



Pr(Multiple of 4) = 7/12

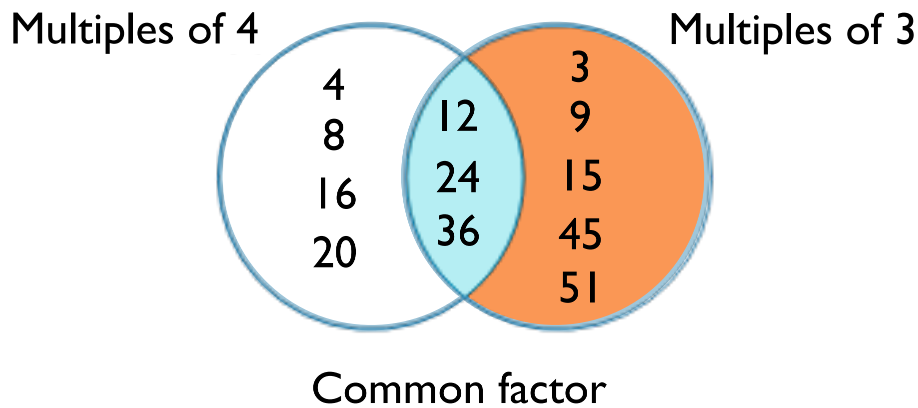
Pr(Multiple of 3) = 8/12

Pr(Multiple of 4 and Multiple of 3) = 3/12

Pr(Multiple of 4 | Multiple of 3)

Slide from Bruno Castro da Silva

Review: Probability Theory



Pr(Multiple of 4) = 7/12

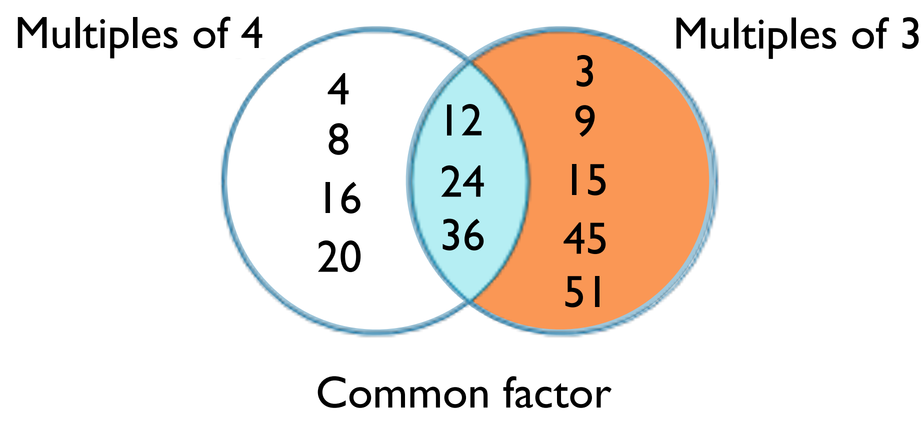
Pr(Multiple of 3) = 8/12

Pr(Multiple of 4 and Multiple of 3) = 3/12

$$\Pr(\text{Multiple of 4} \mid \text{Multiple of 3}) = \frac{\Pr(\text{Multiple of 4 and Multiple of 3})}{}$$

Slide from Bruno Castro da Silva

Review: Probability Theory



$$\Pr(\text{Multiple of 4}) = 7/12$$

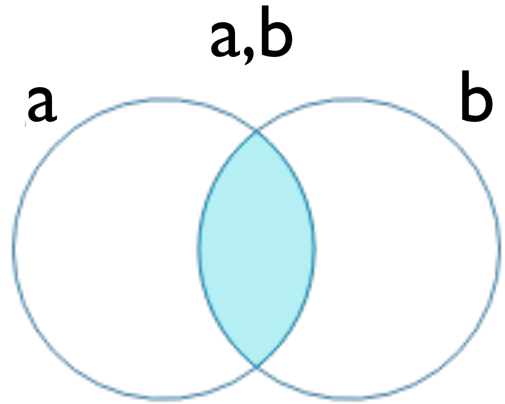
$$\Pr(\text{Multiple of 3}) = 8/12$$

$$\Pr(\text{Multiple of 4 and Multiple of 3}) = 3/12$$

$$\Pr(\text{Multiple of 4} \mid \text{Multiple of 3}) = \frac{\Pr(\text{Multiple of 4 and Multiple of 3})}{\Pr(\text{Multiple of 3})}$$

Slide from Bruno Castro da Silva

Review: Probability Theory

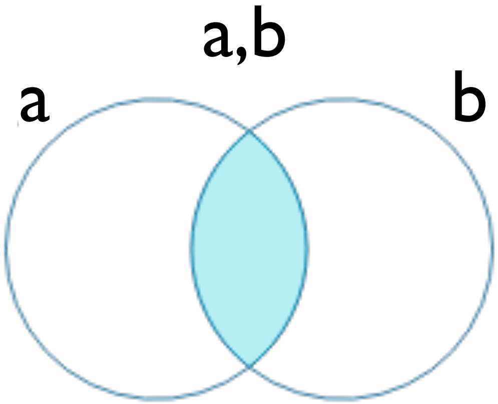


$$\Pr(\text{Multiple of 4} \mid \text{Multiple of 3}) = \frac{\Pr(\text{Multiple of 4 and Multiple of 3})}{\Pr(\text{Multiple of 3})}$$

$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)}$$

Slide from Bruno Castro da Silva

Review: Probability Theory

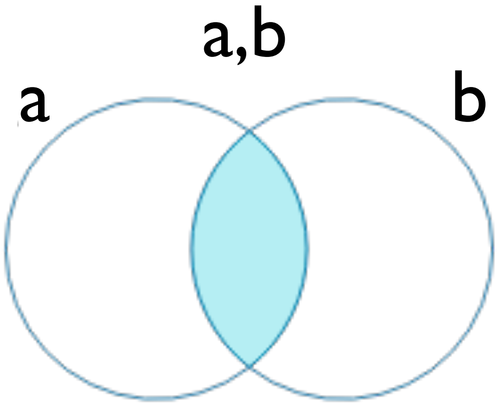


$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)} \qquad \Pr(b \mid a) = \frac{\Pr(b, a)}{\Pr(a)}$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(a, b) \qquad \Pr(b \mid a)\ \Pr(a) = \Pr(b, a)$$

$$\Pr(a, b) = \Pr(b, a)$$

Review: Probability Theory



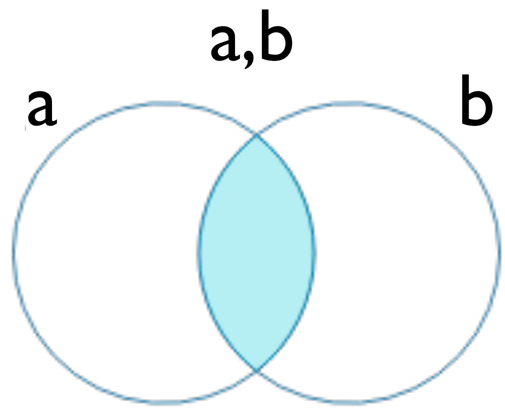
$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)} \qquad \Pr(b \mid a) = \frac{\Pr(b, a)}{\Pr(a)}$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(a, b) \qquad \Pr(b \mid a)\ \Pr(a) = \Pr(b, a)$$

$$\Pr(a, b) = \Pr(b, a)$$

Slide from Bruno Castro da Silva

Review: Probability Theory



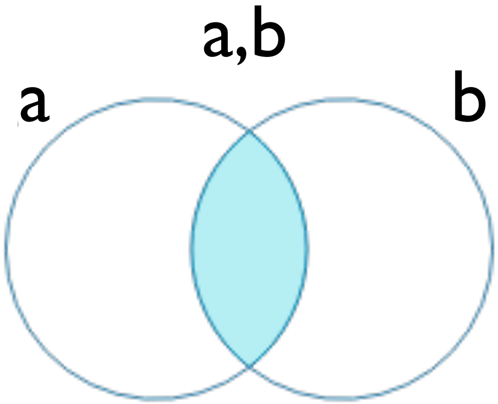
$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)} \qquad \Pr(b \mid a) = \frac{\Pr(b, a)}{\Pr(a)}$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(a, b) \qquad \Pr(b \mid a)\ \Pr(a) = \Pr(a, b)$$

$$\Pr(a, b) = \Pr(b, a)$$

Slide from Bruno Castro da Silva

Review: Probability Theory

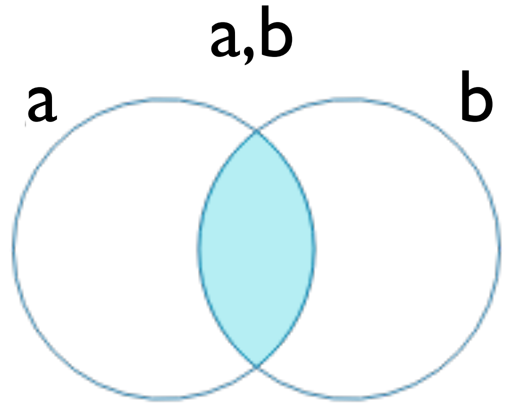


$$\Pr(a \mid b) = \frac{\Pr(a,b)}{\Pr(b)} \qquad \Pr(b \mid a) = \frac{\Pr(b,a)}{\Pr(a)}$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(a,b)$$
$$\Pr(b \mid a)\ \Pr(a) = \Pr(a,b)$$

Slide from Bruno Castro da Silva

Review: Probability Theory



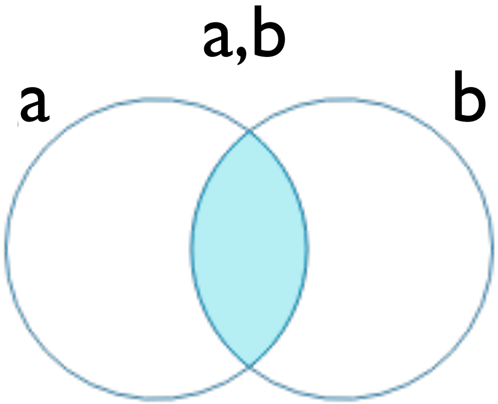
$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)} \qquad \Pr(b \mid a) = \frac{\Pr(b, a)}{\Pr(a)}$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(a, b)$$

$$\Pr(b \mid a)\ \Pr(a) = \Pr(a, b)$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(b \mid a)\ \Pr(a)$$

Review: Probability Theory



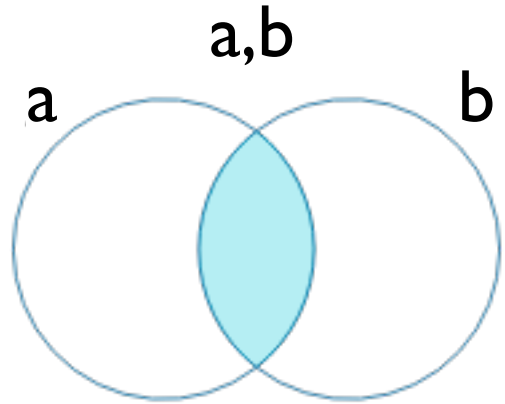
$$\Pr(a \mid b) = \frac{\Pr(a,b)}{\Pr(b)} \qquad \Pr(b \mid a) = \frac{\Pr(b,a)}{\Pr(a)}$$

$$\Pr(a \mid b)\ \Pr(b) = \Pr(b \mid a)\ \Pr(a)$$

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\ \Pr(a)}{\Pr(b)}$$

Slide from Bruno Castro da Silva

Review: Probability Theory



$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)}$$

$$\Pr(b \mid a) = \frac{\Pr(b, a)}{\Pr(a)}$$

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\ \Pr(a)}{\Pr(b)}$$

Bayes Theorem

Slide from Bruno Castro da Silva

Review: Bayes Theorem

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\ \Pr(a)}{\Pr(b)}$$

- Only 1% of the population has cancer (**C**)
- 0.2% of the population is 65 years old (**A**)
- Considering the group of people who have cancer, 0.5% are 65 years old

- **If we don't know anything about John, what is the probability that he has cancer?** **= 1%**
- **If we don't know anything about John, what is the probability that he does not have cancer?** **= 99%**
- **If we know that John is 65 years old, what is the probability that he has cancer?** **= 2.5%**
- **If we know that John is 65 years old, what is the probability that he does not have cancer?** **= 97.5%**

$$\Pr(C \mid A) = \frac{\Pr(A \mid C)\ \Pr(C)}{\Pr(A)} = \frac{(0.5/100)\ \times\ (1/100)}{(0.2/100)} = \frac{(0.5 \times 1)/100}{0.2} = 2.5\ \%$$

Review: Bayes Theorem

$$\Pr(a \mid b) = \frac{\Pr(b \mid a) \Pr(a)}{\Pr(b)}$$

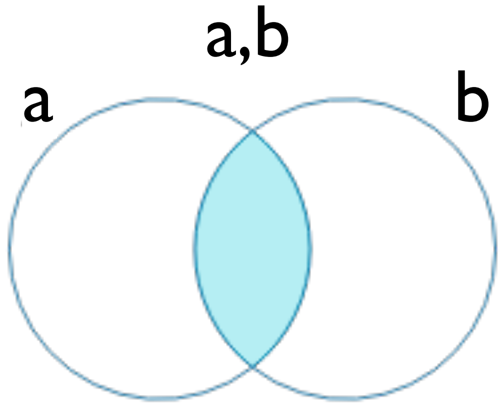
- Only 1% of the population has cancer (**C**)
- 0.2% of the population is 65 years old (**A**)
- Considering the group of people who have cancer, 0.5% are 65 years old

- If we don't know anything about John, what is the probability that he has cancer? = 1%
- If we don't know anything about John, what is the probability that he does not have cancer? = 99%
- If we know that John is 65 years old, what is the probability that he has cancer? = 2.5%
- If we know that John is 65 years old, what is the probability that he does not have cancer? = 97.5%

After we found out John's age (evidence),
we adjusted our prior belief (*a priori* probability of cancer) → prob. cancer = 1%
and obtained an updated belief (*a posteriori* probability) → prob. cancer = 2.5%

Review: Probability Theory

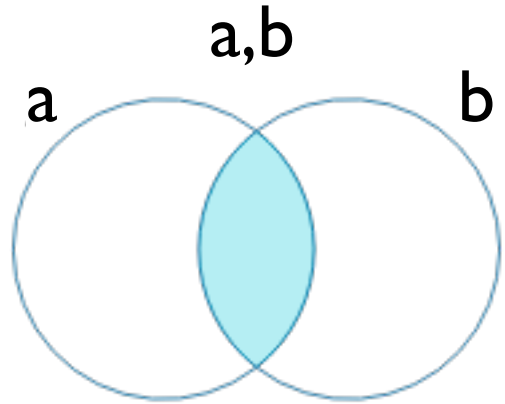
$$\Pr(a \mid b)\ \Pr(b) \;=\; \Pr(a, b)$$



$\Pr(a, b, c)$?

Review: Probability Theory

$$\Pr(a \mid b)\ \Pr(b) = \Pr(a, b)$$

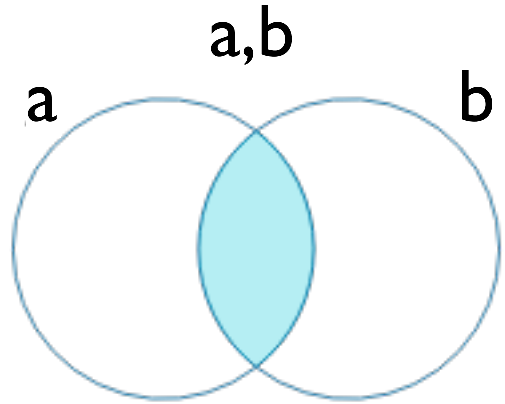


$\Pr(a, b, c)$?

$\Pr(a, \bullet)$ where $\bullet$ is (b,c)

Slide from Bruno Castro da Silva

Review: Probability Theory

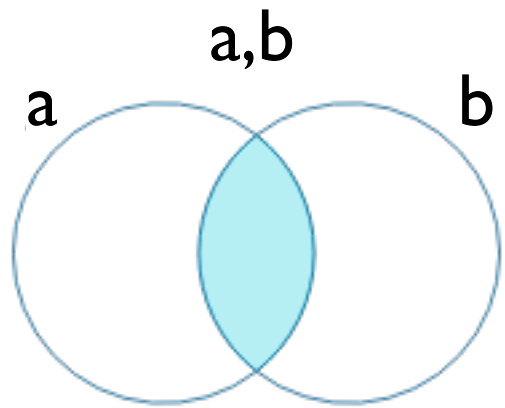


$\Pr(a, b, c)$ **?**

$\Pr(a, \bullet)$ where $\bullet$ is (b,c)

Recall that → $\Pr(a, b) = \Pr(a \mid b) \; \Pr(b)$

Review: Probability Theory



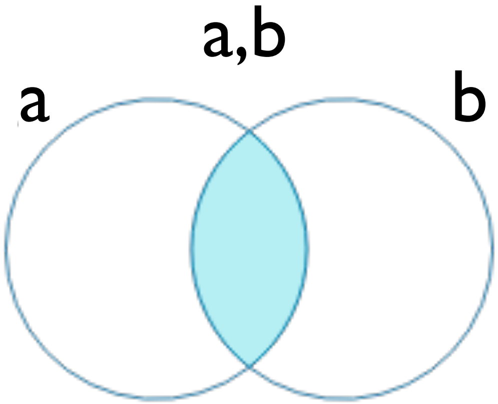
$\Pr(a, b, c)$?

$\Pr(a, \bullet\,)$ where $\bullet$ is (b,c)

Recall that → $\Pr(a, b) = \Pr(a \mid b)\ \Pr(b)$

$\Pr(a, \bullet\,) = \Pr(a \mid \bullet\,)\ \Pr(\,\bullet\,)$

Review: Probability Theory



$\Pr(a, b, c)$ **?**

$\Pr(a, \bullet\,)$ where $\bullet$ is (b,c)

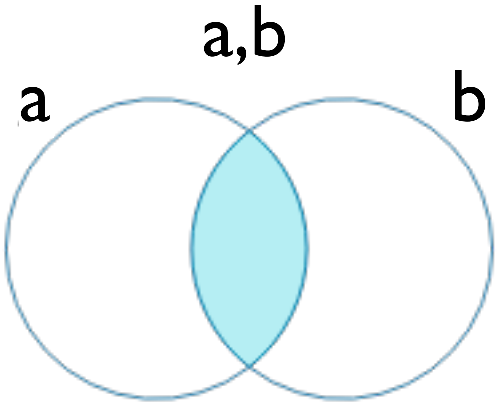
Recall that → $\Pr(a, b) = \Pr(a \mid b)\ \Pr(b)$

$\Pr(a, \bullet\,) = \Pr(a \mid \bullet\,)\ \Pr(\,\bullet\,)$

Replacing def. of $\bullet$ → $\Pr(a, b, c) = \Pr(a \mid b, c)\ \Pr(b, c)$

Slide from Bruno Castro da Silva

Review: Probability Theory



$\Pr(a, b, c)$?

$\Pr(a, \bullet\,)$ where $\bullet$ is (b,c)

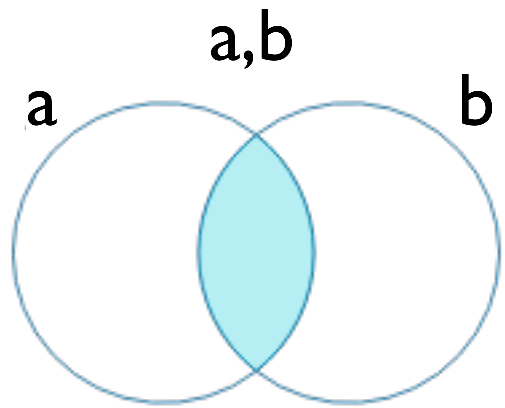
Recall that → $\Pr(a, b) = \Pr(a \mid b)\ \Pr(b)$

$\Pr(a, \bullet\,) = \Pr(a \mid \bullet\,)\ \Pr(\,\bullet\,)$

Replacing def. of • → $\Pr(a, b, c) = \Pr(a \mid b, c)\ \Pr(b, c)$

Slide from Bruno Castro da Silva

Review: Probability Theory



$\Pr(a, b, c)$ **?**

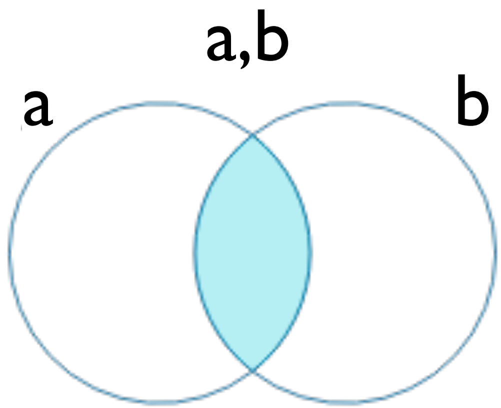
$\Pr(a, \bullet\,)$ where $\bullet$ is (b,c)

$$\Pr(a, b) = \Pr(a \mid b)\ \Pr(b)$$

$$\Pr(a, \bullet\,) = \Pr(a \mid \bullet\,)\ \Pr(\,\bullet\,)$$

$$\Pr(a, b, c) = \Pr(a \mid b, c)\ \boxed{\Pr(b, c)}$$

Review: Probability Theory



$\Pr(a, b, c)$?

$\Pr(a, \bullet)$ where $\bullet$ is (b,c)

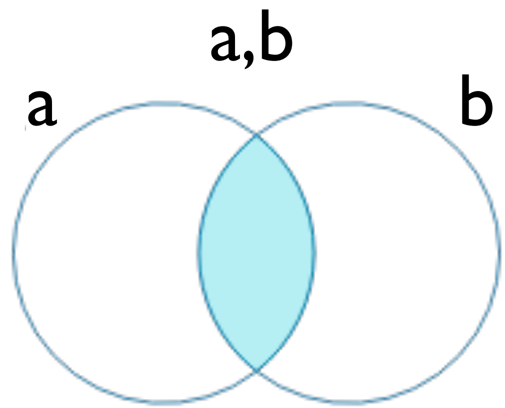
$$\Pr(a, b) = \Pr(a \mid b)\ \Pr(b)$$

$$\Pr(a, \bullet) = \Pr(a \mid \bullet)\ \Pr(\bullet)$$

$$\Pr(a, b, c) = \Pr(a \mid b, c)\ \Pr(b \mid c)\ \Pr(c)$$

Slide from Bruno Castro da Silva

Review: Probability Theory



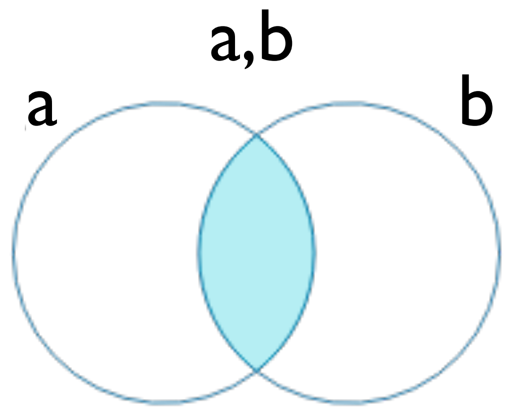
$$\Pr(a, b, c) = \Pr(a \mid b, c) \; \Pr(b \mid c) \; \Pr(c)$$

Similarly…

$$\Pr(a, b, c, d) \;=\; \Pr(a \mid b, c, d) \; \Pr(b \mid c, d) \; \Pr(c \mid d) \; \Pr(d)$$

Chain Rule

Review: Probability Theory



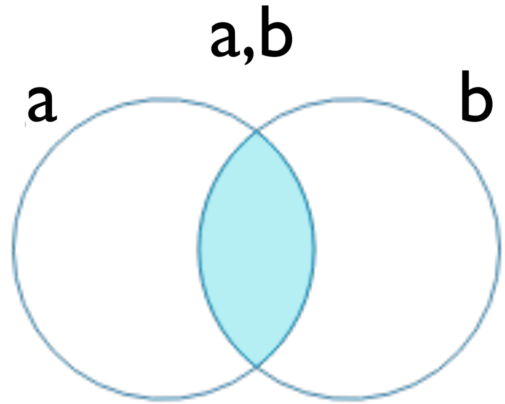
$\Pr(a, b \mid c) \;=\; \Pr(a \mid c)\ \Pr(b \mid c)$ *[conditional independence]*

a and b are conditionally independent given c

if we know that c happened, a does not influence b (and vice-versa)

Slide from Bruno Castro da Silva

Review: Probability Theory

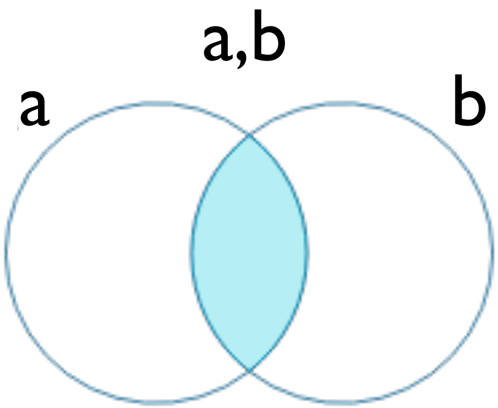


$$\Pr(a, b \mid c) \;=\; \Pr(a \mid c)\;\Pr(b \mid c)$$ *[conditional independence]*

$$\Pr(a \mid b) \;=\; \frac{\Pr(a, b)}{\Pr(b)}$$ *← recall this identity*

let us add a new event, c, to the LHS…

Review: Probability Theory

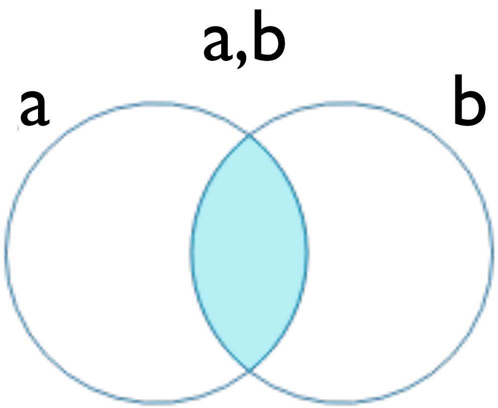


$\Pr(a, b \mid c) \;=\; \Pr(a \mid c)\;\Pr(b \mid c)$ *[conditional independence]*

$$\Pr(a \mid b) \;=\; \frac{\Pr(a, b)}{\Pr(b)}$$ ← *recall this identity*

$\Pr(a \mid b, c)$ *let us add a new event, c, to the LHS…*

Review: Probability Theory



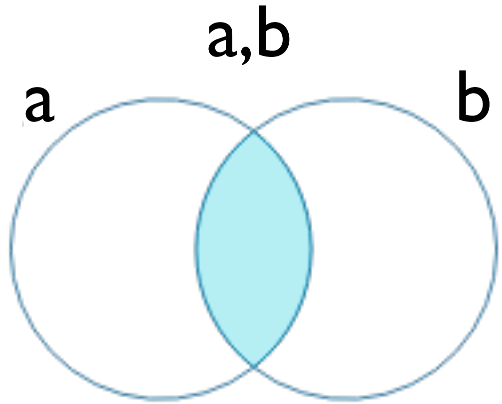
$\Pr(a, b \mid c) \;=\; \Pr(a \mid c)\;\Pr(b \mid c)$ *[conditional independence]*

$$\Pr(a \mid b) \;=\; \frac{\Pr(a, b)}{\Pr(b)}$$ ← *recall this identity*

$$\Pr(a \mid b, c) = \frac{\Pr(a, b \mid c)}{\Pr(b \mid c)}$$ *let us add a new event,* c*, to the LHS…*

Slide from Bruno Castro da Silva

Review: Probability Theory

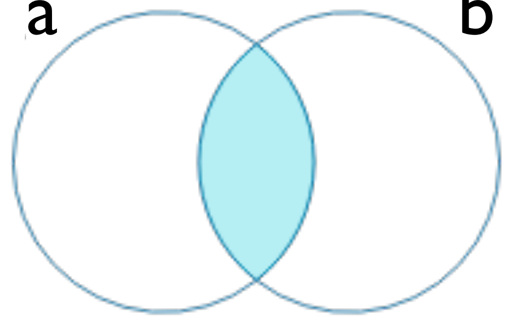


$\Pr(a, b \mid c) \;=\; \Pr(a \mid c)\ \Pr(b \mid c)$ *[conditional independence]*

$\Pr(a \mid b, c) = \dfrac{\boxed{\Pr(a, b \mid c)}}{\Pr(b \mid c)}$ *let us add a new event, c, to the LHS…*

Slide from Bruno Castro da Silva

Review: Probability Theory

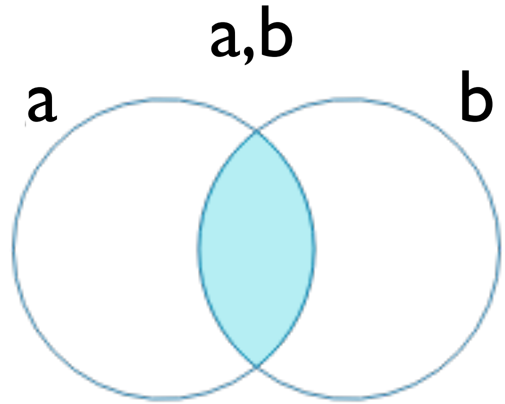


$\Pr(a, b \mid c) = \Pr(a \mid c)\ \Pr(b \mid c)$ *[conditional independence]*

$$\Pr(a \mid b, c) = \frac{\Pr(a, b \mid c)}{\Pr(b \mid c)}$$ *let us add a new event, c, to the LHS…*

$$= \frac{\Pr(a \mid c)\ \Pr(b \mid c)}{\Pr(b \mid c)}$$ *[applying def. of independence]*

Review: Probability Theory



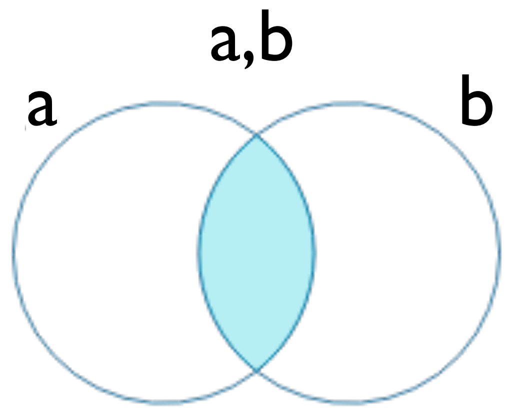
$$\Pr(a, b \mid c) = \Pr(a \mid c)\ \Pr(b \mid c)$$ *[conditional independence]*

$$\Pr(a \mid b, c) = \frac{\Pr(a, b \mid c)}{\Pr(b \mid c)}$$ *let us add a new event, c, to the LHS…*

$$= \frac{\Pr(a \mid c)\ \cancel{\Pr(b \mid c)}}{\cancel{\Pr(b \mid c)}}$$

Slide from Bruno Castro da Silva

Review: Probability Theory



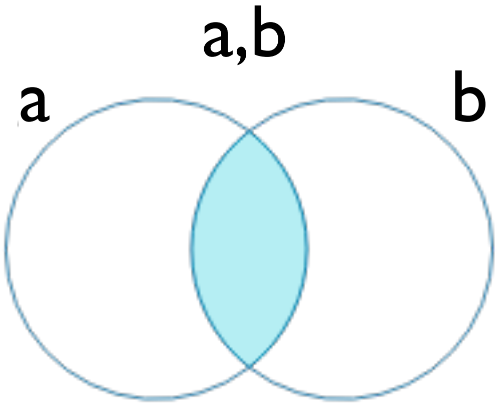
$$\Pr(a, b \mid c) = \Pr(a \mid c)\ \Pr(b \mid c)$$ *[conditional independence]*

$$\Pr(a \mid b, c) = \frac{\Pr(a, b \mid c)}{\Pr(b \mid c)}$$ *let us add a new event, c, to the LHS…*

$$= \frac{\Pr(a \mid c)\ \cancel{\Pr(b \mid c)}}{\cancel{\Pr(b \mid c)}} = \Pr(a \mid c)$$

Slide from Bruno Castro da Silva

Review: Probability Theory



$$\Pr(a, b \mid c) = \Pr(a \mid c)\ \Pr(b \mid c)$$ *[conditional independence]*

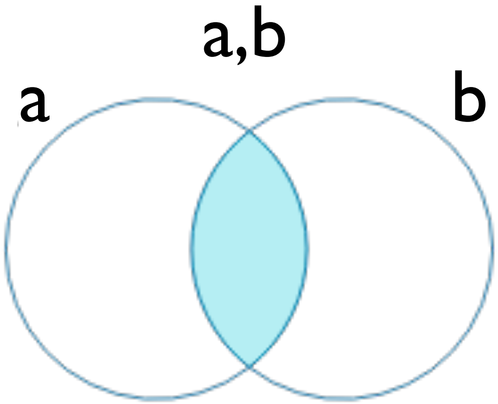
$$\Pr(a \mid b, c) = \Pr(a \mid c)$$

That is, if a and b are independent given c

we can "delete" b from the conditional part of the equation

Slide from Bruno Castro da Silva

Review: Probability Theory



$\Pr(a, \bullet \mid c) \; = \; \Pr(a \mid c) \; \Pr(\bullet \mid c)$ *[conditional independence]*

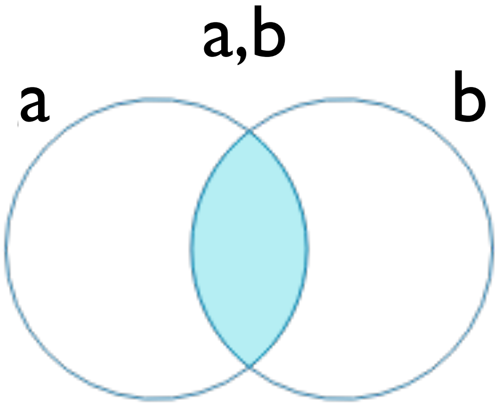
$\Pr(a \mid b, c) \; = \; \Pr(a \mid c)$

That is, if a and b are independent given c

we can "delete" b from the conditional part of the equation

Slide from Bruno Castro da Silva

Review: Probability Theory



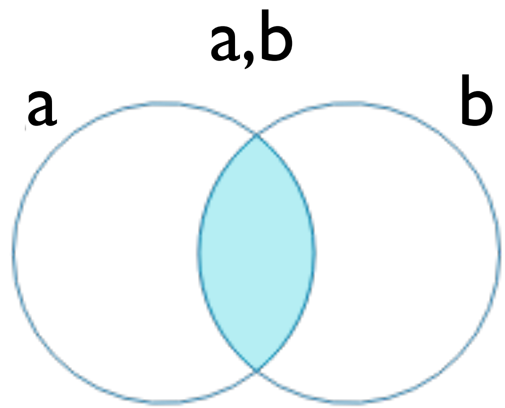
$\Pr(a, \bullet \mid c) \;=\; \Pr(a \mid c)\ \Pr(\bullet \mid c)$ *[conditional independence]*

$\Pr(a \mid \bullet, c) \;=\; \Pr(a \mid c)$

That is, if a and b are independent given c

we can "delete" b from the conditional part of the equation

Review: Probability Theory



$\Pr(a, \bullet \mid c) = \Pr(a \mid c) \Pr(\bullet \mid c)$ *[conditional independence]*

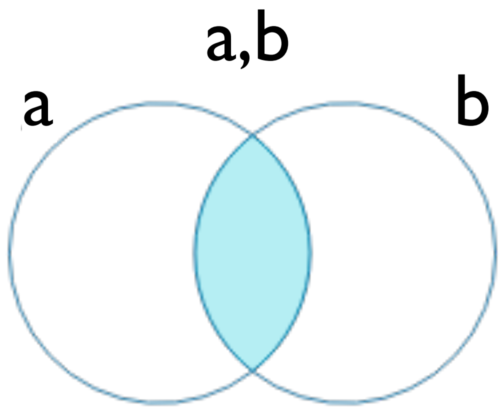
$\Pr(a \mid \bullet, c) = \Pr(a \mid c)$

That is, if a and <u>any</u> • are independent given c

we can "delete" • from the conditional part of the equation

Slide from Bruno Castro da Silva

Review: Probability Theory



$$\Pr(a \mid \bullet, c) = \Pr(a \mid c)$$

That is, if a and <u>any</u> • are independent given c

we can "delete" • from the conditional part of the equation

Pr(go to Bluewall | sunny in Japan, hungry) = Pr(go to Bluewall | hungry)

Can "delete" sunny in Japan *because I am conditioning on what matters* (hungry)

Slide from Bruno Castro da Silva

Review: Probability Theory

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\ \Pr(a)}{\Pr(b)}$$

Bayes Theorem

$$\Pr(a, b, c, d) = \Pr(a \mid b, c, d)\ \Pr(b \mid c, d)\ \Pr(c \mid d)\ \Pr(d)$$

Chain Rule

$$\Pr(a \mid \bullet, c) = \Pr(a \mid c)$$

That is, if a and <u>any</u> • are independent given c

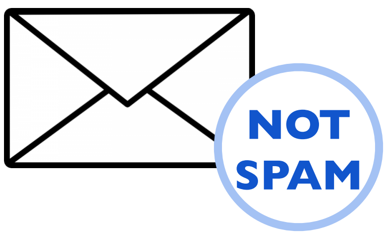
we can "delete" • from the conditional part of the equation

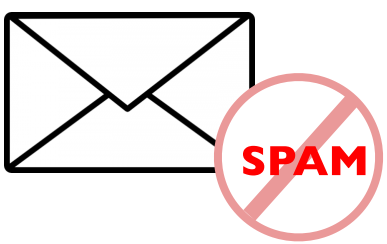
Slide from Bruno Castro da Silva

Naive Bayes Algorithm

Most Likely Class/Label

- Intuition





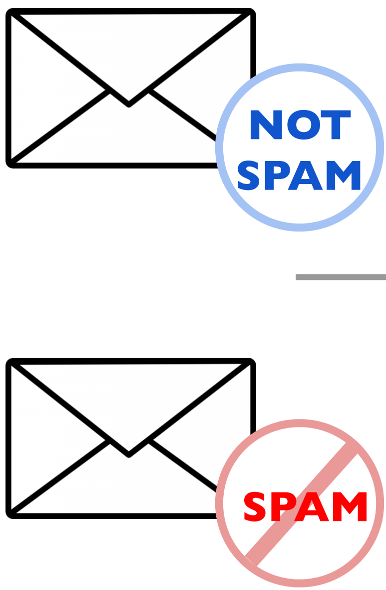
e-mail	Class
1	SPAM
2	SPAM
3	NOT SPAM
4	SPAM
5	SPAM
6	SPAM
7	SPAM
8	NOT SPAM
9	NOT SPAM
10	SPAM

7/10 e-mails:
SPAM

"Given a new email, is it more likely to be a SPAM or NOT SPAM?"

Most Likely Class/Label

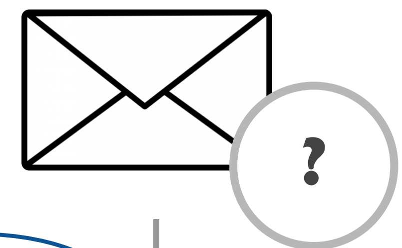
- Intuition

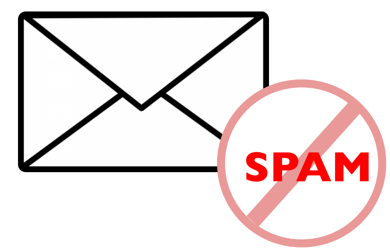


e-mail	Class
1	SPAM
2	SPAM
3	NOT SPAM
4	SPAM
5	SPAM
6	SPAM
7	SPAM
8	NOT SPAM
9	NOT SPAM
10	SPAM

7/10 e-mails: SPAM

New email



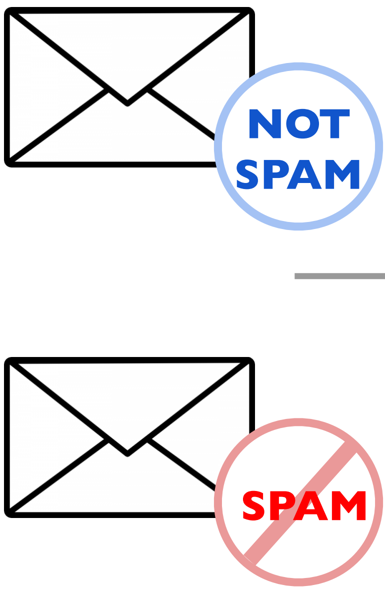


Probability(SPAM) = 0.7

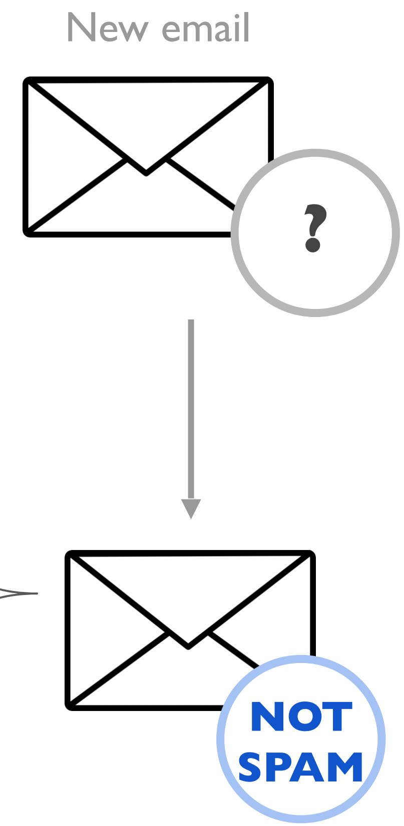
Slide from Bruno Castro da Silva

Most Likely Class/Label

- Intuition



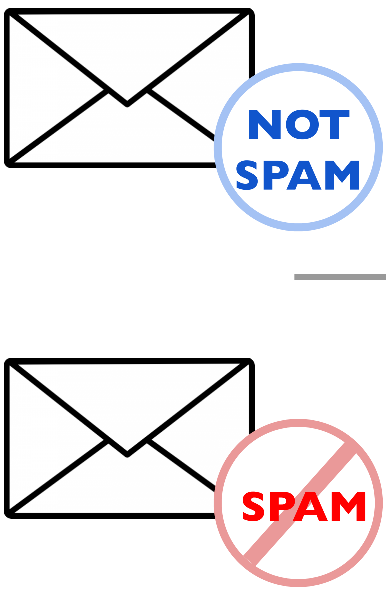
e-mail	Contents	Class
1	"password expired"	SPAM
2	"send password"	SPAM
3	"review conference"	NOT SPAM
4	"password review"	SPAM
5	"review account"	SPAM
6	"account password"	SPAM
7	"send account"	SPAM
8	"conference paper"	NOT SPAM
9	"send paper"	NOT SPAM
10	"expired account"	SPAM



“Given a new email whose contents are "*review conference paper*", it is more likely to be a SPAM or NOT SPAM?

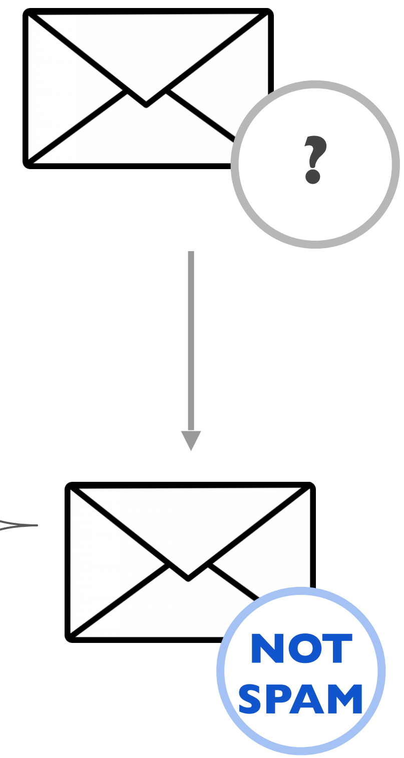
Most Likely Class/Label

- Intuition



e-mail	Contents	Class
1	"password expired"	SPAM
2	"send password"	SPAM
3	"review conference"	NOT SPAM
4	"password review"	SPAM
5	"review account"	SPAM
6	"account password"	SPAM
7	"send account"	SPAM
8	"conference paper"	NOT SPAM
9	"send paper"	NOT SPAM
10	"expired account"	SPAM

New email



What is the most likely class of the email?

Slide from Bruno Castro da Silva

Naive Bayes Classifier

- Probabilistic classifier
- Computes the probability of each class/label based on input attributes of an instance

$$\left.\begin{array}{c} \Pr(y_1 \mid \mathbf{x}) \\ \Pr(y_2 \mid \mathbf{x}) \\ \dots \\ \Pr(y_N \mid \mathbf{x}) \end{array}\right\} \text{max}$$

$$\Pr(y_i \mid \mathbf{x}) = \Pr(y_i \mid x_1, \ \dots, x_n) \qquad C = \underset{y_i \in Y}{\arg\max} \ \Pr(y_i \mid x_1, \ \dots, x_n)$$

Naive Bayes Classifier

- Probabilistic classifier
- Computes the probability of each class/label based on input attributes of an instance

$$\left.\begin{array}{c} \Pr(y_1 \mid \mathbf{x}) \\ \Pr(y_2 \mid \mathbf{x}) \\ \dots \\ \Pr(y_N \mid \mathbf{x}) \end{array}\right\} \text{max}$$

$$\Pr(y_i \mid \mathbf{x}) = \Pr(y_i \mid x_1, \ldots, x_n)$$

$$C = \underset{y_i \in Y}{\arg\max}\ \Pr(y_i \mid x_1, \ldots, x_n)$$

Bayes Theorem

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam

$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

Naive Bayes Classifier

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

How to compute these probabilities based on a training set?

y_1 = spam
y_2 = not_spam

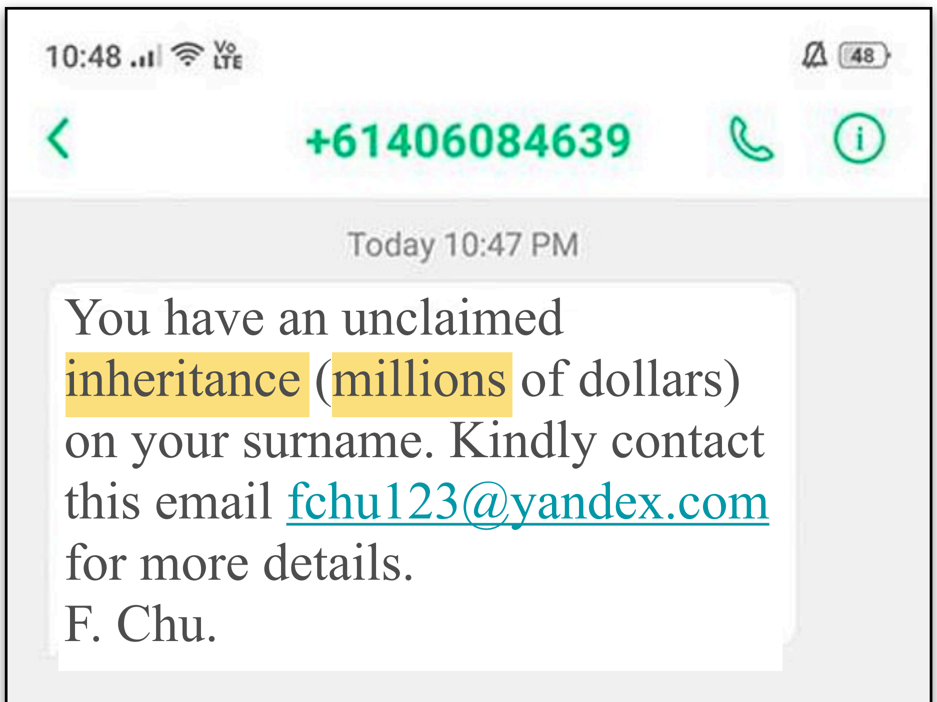
$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

Naive Bayes Classifier

y_1 = **spam**
y_2 = **not_spam**

$\mathbf{x} = (x_1, \ldots x_n)$ = **words in email**

$$\Pr(y_i \,|\, \mathbf{x}) = \frac{\Pr(y_i) \;\; \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$



Naive Bayes Classifier

y_1 = spam
y_2 = not_spam
$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

$$\Pr(y = \textbf{not\_spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{not\_spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{not\_spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

Which class has the highest probability, given the contents of the email?

Naive Bayes Classifier

y_1 = **spam**
y_2 = **not_spam**
$\mathbf{x} = (x_1, \ldots x_n)$ = **words in email**

$$\Pr(y_i \,|\, \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \,|\, x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

$$\Pr(y = \textbf{not\_spam} \,|\, x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{not\_spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{not\_spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

IF $\Pr(\textbf{spam} \,|\, \textbf{contents}) > \Pr(\textbf{not\_spam} \,|\, \textbf{contents})$

Naive Bayes Classifier

y_1 = **spam**
y_2 = **not_spam**
$\mathbf{x} = (x_1, \ldots x_n)$ = **words in email**

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

$$\Pr(y = \textbf{not\_spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{not\_spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{not\_spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

IF $\Pr(\textbf{spam} \mid \textbf{contents}) < \Pr(\textbf{not\_spam} \mid \textbf{contents})$ →

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam

$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

same denominator in both quantities being compared
→ can be ignored

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

$$\Pr(y = \textbf{not\_spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) = \frac{\Pr(y = \textbf{not\_spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{not\_spam})}{\Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions})}$$

Which class has the highest probability, given the contents of the email?

Naive Bayes Classifier

y_1 = **spam**
y_2 = **not_spam**
$\mathbf{x} = (x_1, \ldots x_n)$ = **words in email**

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{spam})$$

$$\Pr(y = \textbf{not\_spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{not\_spam})\ \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \mid y = \textbf{not\_spam})$$

Which class has the highest probability, given the contents of the email?

Naive Bayes Classifier

y_1 = **spam**

y_2 = **not_spam**

$\mathbf{x} = (x_1, \dots x_n)$ = **words in email**

$$\Pr(y_i \,|\, \mathbf{x}) = \frac{\Pr(y_i) \;\; \Pr(x_1, \dots, x_n \mid y_i)}{\Pr(x_1, \dots, x_n)}$$

$$\Pr(y = \textbf{spam} \,|\, x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{spam}) \;\; \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \;|\, y = \textbf{spam})$$

$$\Pr(y = \textbf{not\_spam} \,|\, x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{not\_spam}) \;\; \Pr(x_1 = \textbf{inheritance}, x_2 = \textbf{millions} \;|\, y = \textbf{not\_spam})$$

Which class has the highest probability, given the contents of the email?

$$\underset{y_i \in Y}{\arg\max} \;\; \Pr(y_i \mid x_1, x_2, \;\dots, x_n)$$

Naive Bayes Classifier

$$\arg\max_{y_i \in Y} \Pr(y_i \mid x_1, x_2, \ldots, x_n)$$

$$\arg\max_{y_i \in Y} \frac{\boxed{\Pr(x_1, x_2, \ldots, x_n \mid y_i)}\,\Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$\Pr(x_1, x_2, \ldots, x_n \mid y_i)$ **?**

Remember the chain rule:

$$\Pr(a, b, c, d) = \Pr(a \mid b, c, d)\ \Pr(b \mid c, d)\ \Pr(c \mid d)\ \Pr(d)$$

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) =$$

$$\Pr(x_1 \mid x_2, x_3, \ldots, x_n, y_i)\ \Pr(x_2 \mid x_3, x_4, \ldots, x_n, y_i)\ \ldots\ \Pr(x_{n-1} \mid x_n, y_i)\ \Pr(x_n \mid y_i)$$

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \boxed{\Pr(y_i)}}{\Pr(x_1, x_2, \ldots, x_n)}$$

Among all emails, the percentage of spams

Easy/fast to count

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) =$$

$$\Pr(x_1 \mid x_2, x_3, \ldots, x_n, y_i) \Pr(x_2 \mid x_3, x_4, \ldots, x_n, y_i) \ldots \Pr(x_{n-1} \mid x_n, y_i) \Pr(x_n \mid y_i)$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

Among all emails, the percentage of spams

Easy/fast to count

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) =$$

$$\Pr(x_1 \mid x_2, x_3, \ldots, x_n, y_i) \Pr(x_2 \mid x_3, x_4, \ldots, x_n, y_i) \ldots \Pr(x_{n-1} \mid x_n, y_i) \Pr(x_n \mid y_i)$$

(example)

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\arg\max_{y_i \in Y} \frac{\boxed{\Pr(x_1, x_2, \ldots, x_n \mid y_i)}\ \boxed{\Pr(y_i)}}{\Pr(x_1, x_2, \ldots, x_n)}$$

Among all emails, the percentage of spams

Easy/fast to count

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) =$$

$$\Pr(x_1 \mid x_2, x_3, \ldots, x_n, y_i)\ \Pr(x_2 \mid x_3, x_4, \ldots, x_n, y_i)\ \ldots\ \Pr(x_{n-1} \mid x_n, y_i)\ \Pr(x_n \mid y_i)$$

$$\Pr(x_1 = \text{inheritance}, x_2 = \text{millions} \mid y_i = \text{spam}) =$$ *[using the Chain Rule…]*

$$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions}, y_i = \text{spam})\ \boxed{\Pr(x_2 = \text{millions} \mid y_i = \text{spam})}$$

Among the spams, the percentage that contain the word *millions*

Easy/fast to count

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

Among all emails, the percentage of spams

Easy/fast to count

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) =$$

$$\Pr(x_1 \mid x_2, x_3, \ldots, x_n, y_i) \Pr(x_2 \mid x_3, x_4, \ldots, x_n, y_i) \ldots \Pr(x_{n-1} \mid x_n, y_i) \Pr(x_n \mid y_i)$$

$\Pr(x_1 = \text{inheritance}, x_2 = \text{millions} \mid y_i = \text{spam}) =$ *[using the Chain Rule…]*

$$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions}, y_i = \text{spam}) \Pr(x_2 = \text{millions} \mid y_i = \text{spam})$$

Among the spams that contain the word *millions*, the percentage that also contain the word *inheritance*

Hard/slow to count

Among the spams, the percentage that contain the word *millions*

Easy/fast to count

Naive Bayes Classifier

$\Pr(x_1 = \text{inheritance}, x_2 = \text{millions} \mid y_i = \text{spam}) =$ *[using the Chain Rule…]*

$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions}, y_i = \text{spam})$ $\Pr(x_2 = \text{millions} \mid y_i = \text{spam})$

Among the spams that contain the word *millions*, the percentage that also contain the word *inheritance*

<u>Hard/slow to count</u>

Naive Bayes assumption:
inheritance* and *millions* are conditionally independent given *spam

---

That is: if I know the email is a *spam*,
knowing that it contains the word *inheritance*
does not tell me anything
about how likely it is for *millions* to also show up

Naive Bayes Classifier

$\Pr(x_1 = \text{inheritance},\ x_2 = \text{millions} \mid y_i = \text{spam}) \ =$ *[using the Chain Rule…]*

$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions},\ y_i = \text{spam})\ \Pr(x_2 = \text{millions} \mid y_i = \text{spam})$

Among the spams that contain the word *millions*,
the percentage that also contain the word *inheritance*

Hard/slow to count

Naive Bayes assumption:
inheritance* and *millions* are conditionally independent given *spam

---

The probability of *inheritance* and *millions* are independent given that it's a *spam*

The probability of *inheritance* showing up depends only on whether it is a spam but it does not depend on any other words

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$\Pr(x_1 = \text{inheritance}, x_2 = \text{millions} \mid y_i = \text{spam}) =$ *[using the Chain Rule…]*

$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions}, y_i = \text{spam})$ $\Pr(x_2 = \text{millions} \mid y_i = \text{spam})$

Among the spams that contain the word *millions*, the percentage that also contain the word *inheritance*

Hard/slow to count

Naive Bayes assumption:
inheritance* and *millions* are conditionally independent given *spam

---

True? Not really…. these words show up together all the time!

$$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions}, y_i = \text{spam}) = \Pr(x_1 = \text{inheritance} \mid y_i = \text{spam})$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$$\Pr(x_1 = \text{inheritance} \mid x_2 = \text{millions}, y_i = \text{spam}) = \Pr(x_1 = \text{inheritance} \mid y_i = \text{spam})$$

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

More generally …

Naive Bayes assumption:
x_i **and all other words** $x_{i+1}, \ldots, x_n$ **are conditionally independent given the class** y_i

Naive Bayes Classifier

$$\arg\max_{y_i \in Y} \frac{\Pr(x_1, x_2, \ \ldots, x_n \mid y_i) \ \Pr(y_i)}{\Pr(x_1, x_2, \ \ldots, x_n)}$$

Naive Bayes assumption:
x_i and all other words $x_{i+1}, \ \ldots, x_n$ are conditionally independent given the class y_i

$$\Pr(x_i \mid x_{i+1}, \ \ldots, x_n, y_i) \ = \ \Pr(x_i \mid y_i)$$

$$\Pr(x_1, x_2, \ \ldots, x_n \mid y_i) \ =$$

$$\Pr(x_1 \mid x_2, x_3, \ \ldots, x_n, y_i) \ \Pr(x_2 \mid x_3, x_4, \ \ldots, x_n, y_i) \ \ldots \ \Pr(x_{n-1} \mid x_n, y_i) \ \Pr(x_n \mid y_i)$$

$$= \ \Pr(x_1 \mid y_i) \ \Pr(x_2 \mid y_i) \ \ldots \ \Pr(x_{n-1} \mid y_i) \ \Pr(x_n \mid y_i)$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \; \frac{\Pr(x_1,\, x_2,\; \dots,\, x_n \mid y_i)\; \Pr(y_i)}{\Pr(x_1,\, x_2,\; \dots,\, x_n)}$$

$$\Pr(x_1,\, x_2,\; \dots,\, x_n \mid y_i) \;=\; \Pr(x_1 \mid y_i)\; \Pr(x_2 \mid y_i)\; \dots\; \Pr(x_{n-1} \mid y_i)\; \Pr(x_n \mid y_i)$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\boxed{\Pr(x_1, x_2, \ \ldots, x_n \mid y_i)} \Pr(y_i)}{\Pr(x_1, x_2, \ \ldots, x_n)}$$

$$\boxed{\Pr(x_1, x_2, \ \ldots, x_n \mid y_i)} = \Pr(x_1 \mid y_i) \ \Pr(x_2 \mid y_i) \ \ldots \ \Pr(x_{n-1} \mid y_i) \ \Pr(x_n \mid y_i)$$

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \frac{\boxed{\Pr(x_1, x_2, \ldots, x_n \mid y_i)} \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$$\boxed{\Pr(x_1, x_2, \ldots, x_n \mid y_i)} = \boxed{\Pr(x_1 \mid y_i) \Pr(x_2 \mid y_i) \ldots \Pr(x_{n-1} \mid y_i)} \Pr(x_n \mid y_i)$$

$$\underset{y_i \in Y}{\arg\max} \frac{\boxed{\Pr(x_1 \mid y_i) \Pr(x_2 \mid y_i) \ldots \Pr(x_{n-1} \mid y_i) \Pr(x_n \mid y_i)} \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \; \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \; \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) = \Pr(x_1 \mid y_i) \; \Pr(x_2 \mid y_i) \; \ldots \; \Pr(x_{n-1} \mid y_i) \; \Pr(x_n \mid y_i)$$

$$\underset{y_i \in Y}{\arg\max} \; \frac{\boxed{\Pr(x_1 \mid y_i) \; \Pr(x_2 \mid y_i) \; \ldots \Pr(x_{n-1} \mid y_i) \; \Pr(x_n \mid y_i)} \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$$\underset{y_i \in Y}{\arg\max} \; \frac{\Pr(y_i) \boxed{\prod_{k=1}^{n} \Pr(x_k \mid y_i)}}{\Pr(x_1, x_2, \ldots, x_n)}$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \; \frac{\Pr(x_1, x_2, \ldots, x_n \mid y_i) \; \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$$\Pr(x_1, x_2, \ldots, x_n \mid y_i) = \Pr(x_1 \mid y_i) \; \Pr(x_2 \mid y_i) \; \ldots \; \Pr(x_{n-1} \mid y_i) \; \Pr(x_n \mid y_i)$$

$$\underset{y_i \in Y}{\arg\max} \; \frac{\Pr(x_1 \mid y_i) \; \Pr(x_2 \mid y_i) \; \ldots \Pr(x_{n-1} \mid y_i) \; \Pr(x_n \mid y_i) \; \Pr(y_i)}{\Pr(x_1, x_2, \ldots, x_n)}$$

$$\underset{y_i \in Y}{\arg\max} \; \frac{\Pr(y_i) \; \prod_{k=1}^{n} \Pr(x_k \mid y_i)}{\boxed{\Pr(x_1, x_2, \ldots, x_n)}}$$

Doesn't depend on the variable being maximized (y_i)

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \ \frac{\Pr(x_1,\, x_2,\ \ldots,\, x_n \mid y_i)\ \Pr(y_i)}{\Pr(x_1,\, x_2,\ \ldots,\, x_n)}$$

$$\Pr(x_1,\, x_2,\ \ldots,\, x_n \mid y_i) \ = \ \Pr(x_1 \mid y_i)\ \Pr(x_2 \mid y_i)\ \ldots\ \Pr(x_{n-1} \mid y_i)\ \Pr(x_n \mid y_i)$$

$$\underset{y_i \in Y}{\arg\max} \ \frac{\Pr(x_1 \mid y_i)\ \Pr(x_2 \mid y_i)\ \ldots \Pr(x_{n-1} \mid y_i)\ \Pr(x_n \mid y_i)\ \Pr(y_i)}{\Pr(x_1,\, x_2,\ \ldots,\, x_n)}$$

$$\underset{y_i \in Y}{\arg\max}\ \Pr(y_i) \prod_{k=1}^{n} \Pr(x_k \mid y_i)$$

Slide from Bruno Castro da Silva

Naive Bayes Classifier

$$\underset{y_i \in Y}{\arg\max} \; \Pr(y_i) \prod_{k=1}^{n} \Pr(x_k \mid y_i)$$

Predicted class

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam

$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{spam}) \times \Pr(x_1 = \textbf{inheritance} \mid y = \textbf{spam})\ \Pr(x_2 = \textbf{millions} \mid y = \textbf{spam})$$

$$\Pr(y = \textbf{not\_spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{not\_spam}) \times \Pr(x_1 = \textbf{inheritance} \mid y = \textbf{not\_spam})\ \Pr(x_2 = \textbf{millions} \mid y = \textbf{not\_spam})$$

Which class has the highest probability, given the contents of the email?

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam

$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \boxed{\Pr(y = \textbf{spam})} \times \Pr(x_1 = \textbf{inheritance} \mid y = \textbf{spam})\ \Pr(x_2 = \textbf{millions} \mid y = \textbf{spam})$$

$$\boxed{\frac{\textbf{spams}}{\textbf{emails}}} \times \frac{\textbf{spams containing inheritance}}{\textbf{spams}} \times \frac{\textbf{spams containing millions}}{\textbf{spams}}$$

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam
$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{spam}) \times \boxed{\Pr(x_1 = \textbf{inheritance} \mid y = \textbf{spam})}\ \Pr(x_2 = \textbf{millions} \mid y = \textbf{spam})$$

$$\frac{\textbf{spams}}{\textbf{emails}} \times \boxed{\frac{\textbf{spams containing inheritance}}{\textbf{spams}}} \times \frac{\textbf{spams containing millions}}{\textbf{spams}}$$

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam
$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i) \ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto \Pr(y = \textbf{spam}) \times \Pr(x_1 = \textbf{inheritance} \mid y = \textbf{spam}) \boxed{\Pr(x_2 = \textbf{millions} \mid y = \textbf{spam})}$$

$$\frac{\textbf{spams}}{\textbf{emails}} \times \frac{\textbf{spams containing inheritance}}{\textbf{spams}} \times \boxed{\frac{\textbf{spams containing millions}}{\textbf{spams}}}$$

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam
$\mathbf{x} = (x_1, \dots x_n)$ = words in email

$$\Pr(y_i \mid \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \dots, x_n \mid y_i)}{\Pr(x_1, \dots, x_n)}$$

$$\Pr(y = \textbf{spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto$$

$$\frac{\textbf{spams}}{\textbf{emails}} \times \frac{\textbf{spams containing inheritance}}{\textbf{spams}} \times \frac{\textbf{spams containing millions}}{\textbf{spams}}$$

$$\Pr(y = \textbf{not\_spam} \mid x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \propto$$

$$\frac{\textbf{not\_spams}}{\textbf{emails}} \times \frac{\textbf{not\_spams containing inheritance}}{\textbf{not\_spams}} \times \frac{\textbf{not\_spams containing millions}}{\textbf{not\_spams}}$$

in practice

Naive Bayes Classifier

y_1 = spam
y_2 = not_spam
$\mathbf{x} = (x_1, \ldots x_n)$ = words in email

$$\Pr(y_i \,|\, \mathbf{x}) = \frac{\Pr(y_i)\ \Pr(x_1, \ldots, x_n \mid y_i)}{\Pr(x_1, \ldots, x_n)}$$

$$\left.\begin{array}{l} \Pr(y = \textbf{spam} \,|\, x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \\[2em] \Pr(y = \textbf{not\_spam} \,|\, x_1 = \textbf{inheritance}, x_2 = \textbf{millions}) \end{array}\right\} \text{max}$$

in practice

Naive Bayes Classifier - Example

$$y := \underset{y_i \in Y}{\arg\max} \ \Pr(y_i) \prod_{k=1}^{n} \Pr(x_k \mid y_i)$$

Instance #	***Color***	***Type***	***Made_In***	**Class/Label: Likely to be stolen?**
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

- We wish to predict: will a red SUV made in the U.S. be stolen?

Naive Bayes Classifier - Example

$$y := \underset{y_i \in Y}{\arg\max} \; \Pr(y_i) \prod_{k=1}^{n} \Pr(x_k \mid y_i)$$

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	Yes
2	Red	Convertible	USA	No
3	Red	Convertible	USA	Yes
4	Yellow	Convertible	USA	No
5	Yellow	Convertible	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	USA	No
9	Red	Convertible	Imported	No
10	Red	Convertible	Imported	Yes

will a
red SUV made in the U.S.
be stolen?

Pr(Stolen=yes | color=Red, type=SUV, made_in=USA)

Pr(Stolen=no | color=Red, type=SUV, made_in=USA)

Naive Bayes Classifier - Example

Instance #	Color	Type	Made_In	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	Yes
2	Red	Convertible	USA	No
3	Red	Convertible	USA	Yes
4	Yellow	Convertible	USA	No
5	Yellow	Convertible	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	USA	No
9	Red	Convertible	Imported	No
10	Red	Convertible	Imported	Yes

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA})$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10} \qquad \Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{3}{5} \qquad \Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA})$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

$$\Pr(\text{Stolen=No}) = \frac{5}{10} \qquad \Pr(\text{Made\_In = USA} \mid \text{Stolen=No}) = \frac{3}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=No}) = \frac{2}{5} \qquad \Pr(\text{Type = SUV} \mid \text{Stolen=No}) = \frac{2}{5}$$

Naive Bayes Classifier - Example

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	Yes
2	Red	Convertible	USA	No
3	Red	Convertible	USA	Yes
4	Yellow	Convertible	USA	No
5	Yellow	Convertible	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	USA	No
9	Red	Convertible	Imported	No
10	Red	Convertible	Imported	Yes

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} = \frac{30}{10 \times 5^3}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10} \qquad \Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{3}{5} \qquad \Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = \frac{60}{10 \times 5^3}$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

$$\Pr(\text{Stolen=No}) = \frac{5}{10} \qquad \Pr(\text{Made\_In = USA} \mid \text{Stolen=No}) = \frac{3}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=No}) = \frac{2}{5} \qquad \Pr(\text{Type = SUV} \mid \text{Stolen=No}) = \frac{2}{5}$$

Naive Bayes Classifier - Problem #1

Instance #	*Color*	*Type*	*Made_In*	**Class/Label: Likely to be stolen?**
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA})$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10}$$

$$\Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \text{?}$$

$$\Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

What would happen if there were no Red examples in the training set?

Naive Bayes Classifier - Problem #1

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA})$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10}$$

$$\Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{\mathbf{0}}{5}$$

$$\Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

What would happen if there were no Red examples in the training set?

Naive Bayes Classifier - Problem #1

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{\mathbf{0}}{5} \times \frac{2}{5} \times \frac{1}{5} = \mathbf{0}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10}$$

$$\Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{\mathbf{0}}{5}$$

$$\Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

What would happen if there were no Red examples in the training set?

Naive Bayes Classifier - Problem #1

Instance #	Color	Type	Made_In	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{\mathbf{0}}{5} \times \frac{2}{5} \times \frac{1}{5} = \mathbf{0}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10}$$

$$\Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{\mathbf{0}}{5}$$

$$\Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

What would happen if there were no Red examples in the training set?

Would always estimate zero probability!

Naive Bayes Classifier - Problem #1

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	Yes
2	Red	Convertible	USA	No
3	Red	Convertible	USA	Yes
4	Yellow	Convertible	USA	No
5	Yellow	Convertible	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	USA	No
9	Red	Convertible	Imported	No
10	Red	Convertible	Imported	Yes

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \boxed{\frac{5}{10} \times \frac{\mathbf{0}}{5} \times \frac{2}{5} \times \frac{1}{5} = \mathbf{0}}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \frac{5}{10}$$

$$\Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \frac{2}{5}$$

$$\boxed{\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{\mathbf{0}}{5}}$$

$$\Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \frac{1}{5}$$

Possible solution: Laplace Smoothing

→ **adds (to numerator) 1**

→ **adds (to denominator) the number of possible values of the attribute**

e.g., |Color| = |{Red, Yellow}| = 2

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{\#\text{stolenRedCars} + 1}{\#\text{stolenCars} + 2} = \frac{0+1}{5+2} = \frac{1}{7}$$

Naive Bayes Classifier - Problem #1

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	Yes
2	Red	Convertible	USA	No
3	Red	Convertible	USA	Yes
4	Yellow	Convertible	USA	No
5	Yellow	Convertible	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	USA	No
9	Red	Convertible	Imported	No
10	Red	Convertible	Imported	Yes

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{\mathbf{0}}{5} \times \frac{2}{5} \times \frac{1}{5} = \mathbf{0}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=Yes}) = \cancel{\frac{5}{10}} \quad \frac{5+1}{10+2} = \frac{6}{12}$$

$$\Pr(\text{Made\_In = USA} \mid \text{Stolen=Yes}) = \cancel{\frac{2}{5}} \quad \frac{2+1}{5+2} = \frac{3}{7}$$

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \cancel{\frac{0}{5}} \quad \frac{0+1}{5+2} = \frac{1}{7}$$

$$\Pr(\text{Type = SUV} \mid \text{Stolen=Yes}) = \cancel{\frac{1}{5}} \quad \frac{1+1}{5+2} = \frac{2}{7}$$

Possible solution: Laplace Smoothing

→ **adds (to numerator) 1**

→ **adds (to denominator) the number of possible values of the attribute**

e.g., |Color| = |{Red, Yellow}| = 2

$$\Pr(\text{Color=Red} \mid \text{Stolen=Yes}) = \frac{\#\text{stolenRedCars} + 1}{\#\text{stolenCars} + 2} = \frac{0+1}{5+2} = \frac{1}{7}$$

Naive Bayes Classifier - Problem #2

Instance #	*Color*	*Type*	*Made_In*	**Class/Label: Likely to be stolen?**
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} = \frac{30}{10 \times 5^3}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = \frac{60}{10 \times 5^3}$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

Naive Bayes Classifier - Problem #2

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} = \frac{30}{10 \times 5^3}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = \frac{60}{10 \times 5^3}$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

Multiplying many small numbers between 0 and 1 (fractions)

Rapidly gets close to zero

Problems losing precision due to floating point representation

E.g.: $\left(\frac{1}{1000} \times \frac{35}{550} \times \frac{52}{550} \times \frac{27}{550} \times \frac{67}{550} \times \frac{181}{550}\right) = .00000000956$

Naive Bayes Classifier - Problem #2

Instance #	Color	Type	Made_In	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	Yes
2	Red	Convertible	USA	No
3	Red	Convertible	USA	Yes
4	Yellow	Convertible	USA	No
5	Yellow	Convertible	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	USA	No
9	Red	Convertible	Imported	No
10	Red	Convertible	Imported	Yes

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} = \frac{30}{10 \times 5^3}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = \frac{60}{10 \times 5^3}$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

Possible solution: to compare the logarithm of those quantities!

E.g.: $$\left(\frac{1}{1000} \times \frac{35}{550} \times \frac{52}{550} \times \frac{27}{550} \times \frac{67}{550} \times \frac{181}{550}\right) = .0000000956$$

Naive Bayes Classifier - Problem #2

Instance #	*Color*	*Type*	*Made_In*	**Class/Label: Likely to be stolen?**
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \log\left(\frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5}\right)$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \log\left(\frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5}\right)$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

Possible solution: to compare the logarithm of those quantities!

E.g.: $$\left(\frac{1}{1000} \times \frac{35}{550} \times \frac{52}{550} \times \frac{27}{550} \times \frac{67}{550} \times \frac{181}{550}\right) = .0000000956$$

Naive Bayes Classifier - Problem #2

Instance #	*Color*	*Type*	*Made_In*	**Class/Label: Likely to be stolen?**
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \log\left(\frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5}\right)$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \log\left(\frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5}\right)$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

Possible solution: to compare the logarithm of those quantities!

E.g.: $\log\left(\frac{1}{1000} \times \frac{35}{550} \times \frac{52}{550} \times \frac{27}{550} \times \frac{67}{550} \times \frac{181}{550}\right) = ?$

Naive Bayes Classifier - Problem #2

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \log\left(\frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5}\right)$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \log\left(\frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5}\right)$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

E.g.: $\log\left(\frac{1}{1000} \times \frac{35}{550} \times \frac{52}{550} \times \frac{27}{550} \times \frac{67}{550} \times \frac{181}{550}\right)$

$$= \log\left(\frac{1}{1000}\right) + \log\left(\frac{35}{550}\right) + \log\left(\frac{52}{550}\right) + \log\left(\frac{27}{550}\right) + \log\left(\frac{67}{550}\right) + \log\left(\frac{181}{550}\right)$$

Naive Bayes Classifier - Problem #2

Instance #	*Color*	*Type*	*Made_In*	Class/Label: Likely to be stolen?
1	Red	Convertible	USA	**Yes**
2	Red	Convertible	USA	**No**
3	Red	Convertible	USA	**Yes**
4	Yellow	Convertible	USA	**No**
5	Yellow	Convertible	Imported	**Yes**
6	Yellow	SUV	Imported	**No**
7	Yellow	SUV	Imported	**Yes**
8	Yellow	SUV	USA	**No**
9	Red	Convertible	Imported	**No**
10	Red	Convertible	Imported	**Yes**

will a
red SUV made in the U.S.
get stolen?

$$\Pr(\text{Stolen=yes} \mid \text{color=Red, type=SUV, made\_in=USA}) = \boxed{\log\left(\frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5}\right)}$$

$$= \Pr(\text{Stolen}) \times \Pr(\text{Red} \mid \text{Stolen}) \times \Pr(\text{USA} \mid \text{Stolen}) \times \Pr(\text{SUV} \mid \text{Stolen})$$

$$\Pr(\text{Stolen=no} \mid \text{color=Red, type=SUV, made\_in=USA}) = \boxed{\log\left(\frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5}\right)}$$

$$= \Pr(\text{Not\_Stolen}) \times \Pr(\text{Red} \mid \text{Not\_Stolen}) \times \Pr(\text{USA} \mid \text{Not\_Stolen}) \times \Pr(\text{SUV} \mid \text{Not\_Stolen})$$

which is larger?

$$= \log\left(\frac{1}{1000}\right) + \log\left(\frac{35}{550}\right) + \log\left(\frac{52}{550}\right) + \log\left(\frac{27}{550}\right) + \log\left(\frac{67}{550}\right) + \log\left(\frac{181}{550}\right)$$

~~Multiplying~~ Adding <u>many</u> small numbers between 0 and 1 (fractions)

Fewer problems with precision due to floating point operations

Naive Bayes with Numeric Attributes

- Recall the Naive Bayes "strategy" we discussed last time
  - it worked well in case each instance was described by categorial attributes
  - e.g., one possible word, one possible color of a car, etc.

$$\Pr(y_i \mid \mathbf{x}) = \Pr(y_i \mid x_1, \ldots, x_n) = \Pr(y_i) \prod_{k=1}^{n} \Pr(x_k \mid y_i)$$

Probability of SPAM given words "inheritance" and "millions"

We estimated this probability via counters:

- how many times does "inheritance" appear in SPAMs
- how many times does "millions" appear in SPAMs

- But what if the attributes x_k are not words (or, more generally speaking, categorical variables)?
  - what if the attributes are numeric/continuous?

$$\Pr(y_i = \text{Snow} \mid \text{Temp} = 20^{\circ}\text{F},\ \text{Humidity}=60\%)$$

Naive Bayes with Numeric Attributes

- But what if the attributes x_k are not words (or, more generally speaking, categorical variables)?
  - what if the attributes are numeric/continuous?

$$\Pr(y_i = \text{Snow} \mid \text{Temp} = 20°\text{F},\ \text{Humidity}=60\%)$$

$$= \Pr(y_i = \text{Snow})\ \Pr(\text{Temp} = 20°\text{F}, \mid y_i = \text{Snow})\ \Pr(\text{Humidity}=60\% \mid y_i = \text{Snow})$$

We can still estimate this.
Just count the percentage of instances
that have the label "Snow"

Temperature can be any real number from, say, -20°F to 110°F

- Can we keep counters for how often each possible temperature occurs?
- No! There's an infinite number of possible values…

Slide from Bruno Castro da Silva

Naive Bayes with Numeric Attributes

- But what if the attributes x_k are not words (or, more generally speaking, categorical variables)?
  - what if the attributes are numeric/continuous?

$$\Pr(y_i = \text{Snow} \mid \text{Temp} = 20^\circ\text{F},\ \text{Humidity=60\%})$$

$$= \Pr(y_i = \text{Snow})\ \ \Pr(\text{Temp} = 20^\circ\text{F}, \mid y_i = \text{Snow})\ \ \Pr(\text{Humidity=60\%} \mid y_i = \text{Snow})$$

- Two possible approaches:

1. Discretize the continuous variable
   - "Temperature" can be transformed into a categorial attribute
     - Discretized_temperature=Cold if Temperature is from -20°F to 40°F
     - Discretized_temperature=Mild if Temperature is from 40°F to 70°F
     - Discretized_temperature=Hot if Temperature is from 70°F to 110°F

2. Assume that the continuous variable comes from some distribution
   - e.g., that temperature values are distributed according to a Gaussian distribution
   - use the training examples to find the parameters of such a distribution
   - for example, the mean temperature and how much temperature varies (its standard deviation)

Slide from Bruno Castro da Silva

Naive Bayes with Numeric Attributes

- But what if the attributes x_k are not words (or, more generally speaking, categorical variables)?
  - what if the attributes are numeric/continuous?

$$\Pr(y_i = \text{Snow} \mid \text{Temp} = 20^\circ\text{F},\ \text{Humidity=60\%})$$

$$= \Pr(y_i = \text{Snow})\ \Pr(\text{Temp} = 20^\circ\text{F}, \mid y_i = \text{Snow})\ \Pr(\text{Humidity=60\%} \mid y_i = \text{Snow})$$

- Two possible approaches:

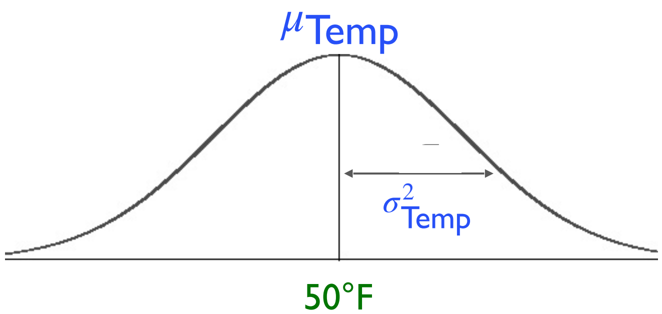
1. Discretize the continuous variable
   - "Temperature" can be transformed into a categorial attribute
     - Discretized_temperature=Cold if Temperature is from -20°F to 40°F
     - Discretized_temperature=Mild if Temperature is from 40°F to 70°F
     - Discretized_temperature=Hot if Temperature is from 70°F to 110°F

2. Assume that the continuous variable comes from some distribution
   - e.g., that temperature values are distributed according to a Gaussian distribution
   - use the training examples to find the parameters of such a distribution
   - for example, the mean temperatura and how much temperature varies (its standard deviation)

Naive Bayes with Numeric Attributes

- Let's assume that the continuous attribute, Temperature, is distributed according to a Gaussian

$$\text{Temp} \sim \mathcal{N}\left(\mu_{\text{Temp}}, \sigma^2_{\text{Temp}}\right)$$



Pr(Temp=53˚F) ?

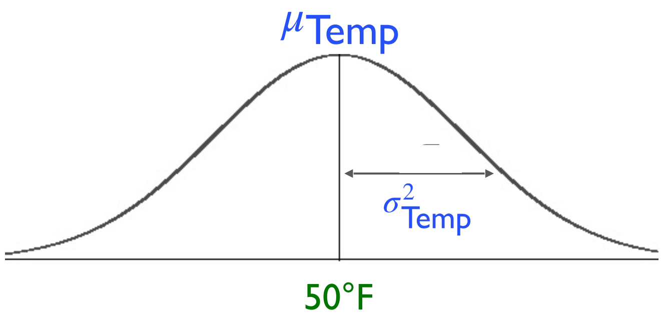
We can use the probability density function
of a Gaussian distribution
with mean μ and standard deviation σ

Slide from Bruno Castro da Silva

Naive Bayes with Numeric Attributes

- Let's assume that the continuous attribute, Temperature, is distributed according to a Gaussian

$$\text{Temp} \sim \mathcal{N}\left(\mu_{\text{Temp}}, \sigma^2_{\text{Temp}}\right)$$



Pr(Temp=53°F) **?**

We can use the probability density function
of a Gaussian distribution
with mean μ and standard deviation σ

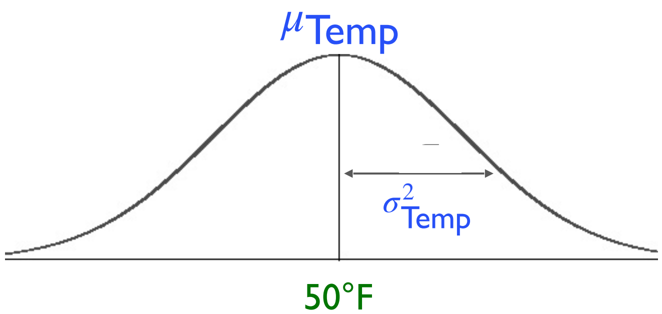
$$\Pr(\text{Temp}=x°\text{F}) = \boxed{f(x;\, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Slide from Bruno Castro da Silva

Naive Bayes with Numeric Attributes

- Let's assume that the continuous attribute, Temperature, is distributed according to a Gaussian

$$\text{Temp} \sim \mathcal{N}\left(\mu_{\text{Temp}}, \sigma^2_{\text{Temp}}\right)$$



$$\Pr(y_i = \text{Snow} \mid \text{Temp=53°F})$$

$$= \Pr(y_i = \text{Snow}) \Pr(\text{Temp=53°F} \mid y_i = \text{Snow})$$

Assume the temperature (when it is snowing) is modeled by a Gaussian distribution with:

- mean $\mu_{\text{Temp, Snow}}$
- standard deviation $\sigma_{\text{Temp, Snow}}$

$\Pr(\text{Temp=53°F})$?

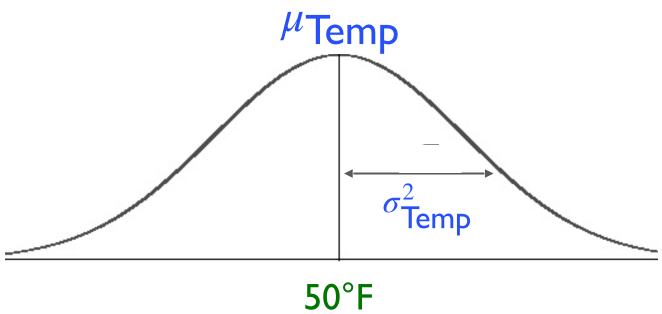
We can use the probability density function of a Gaussian distribution with mean μ and standard deviation σ

$$\Pr(\text{Temp=x°F}) = \boxed{f(x;\, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Naive Bayes with Numeric Attributes

- Let's assume that the continuous attribute, Temperature, is distributed according to a Gaussian

$$\text{Temp} \sim \mathcal{N}\left(\mu_{\text{Temp}}, \sigma^2_{\text{Temp}}\right)$$



$\Pr(y_i = \text{Snow} \mid \text{Temp=53°F})$

$= \Pr(y_i = \text{Snow}) \Pr(\text{Temp=53°F} \mid y_i = \text{Snow})$

Assume the temperature (when it is snowing) is modeled by a Gaussian distribution with:

- mean $\mu_{\text{Temp, Snow}}$
- standard deviation $\sigma_{\text{Temp, Snow}}$

$\Pr(\text{Temp=53°F} \mid y_i = \text{Snow})$

$= f\left(53;\ \mu_{\text{Temp, Snow}}, \sigma_{\text{Temp, Snow}}\right)$

Similarly,

$\Pr(\text{Temp=53°F} \mid y_i = \text{Sunny})$

$= f\left(53;\ \mu_{\text{Temp, Sunny}}, \sigma_{\text{Temp, Sunny}}\right)$

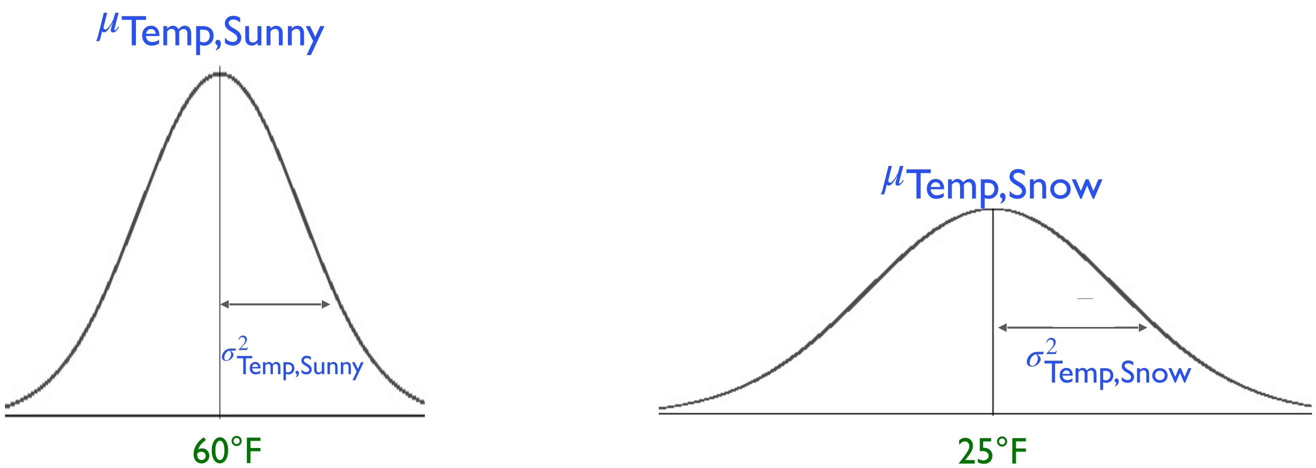
$\Pr(\text{Temp=53°F})$ **?**

We can use the probability density function of a Gaussian distribution with mean μ and standard deviation σ

$$\Pr(\text{Temp=x°F}) = f(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Naive Bayes with Numeric Attributes

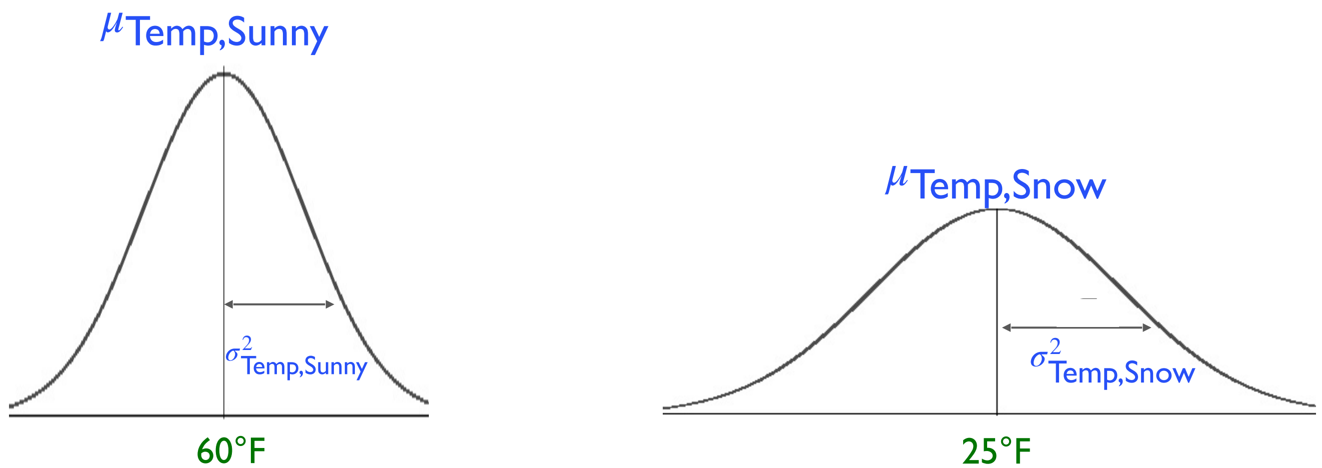
- Let's assume that the continuous attribute, Temperature, is distributed according to a Gaussian



Notice that the distribution of temperatures might be different depending on the class (i.e., whether it is Snowing or Sunny)

Naive Bayes with Numeric Attributes

- Let's assume that the continuous attribute, Temperature, is distributed according to a Gaussian



Notice that the distribution of temperatures might be different depending on the class (i.e., whether it is Snowing or Sunny)

So to be able to compute

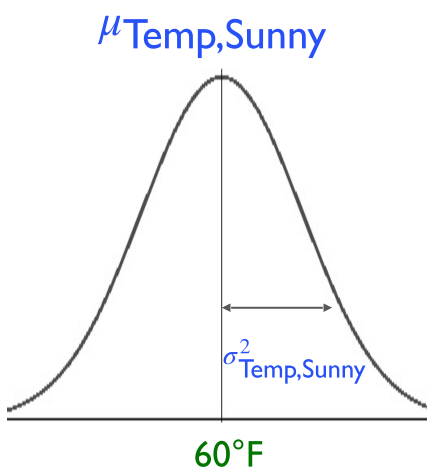
$\Pr(\text{Temp=53°F} \mid y_i = \text{Snow})$

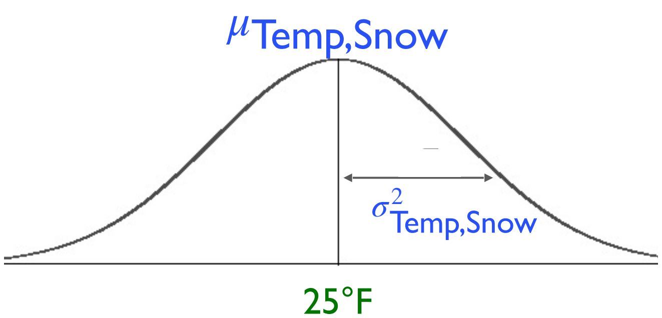
and

$\Pr(\text{Temp=53°F} \mid y_i = \text{Sunny})$

→ we need to estimate the parameters of both distributions → $\mu_{\text{Temp, Snow}}$ $\mu_{\text{Temp, Sunny}}$ $\sigma_{\text{Temp, Snow}}$ $\sigma_{\text{Temp, Sunny}}$

Naive Bayes with Numeric Attributes





we need to estimate
the parameters of
both distributions

→ $\mu_{Temp, Snow}$ $\mu_{Temp, Sunny}$
$\sigma_{Temp, Snow}$ $\sigma_{Temp, Sunny}$

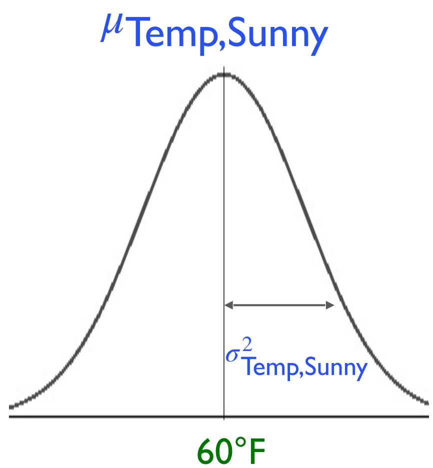
Instance	Temp	Class
#1	20	Snow
#2	57	Sunny
#3	63	Sunny
#4	15	Snow
#5	25	Snow

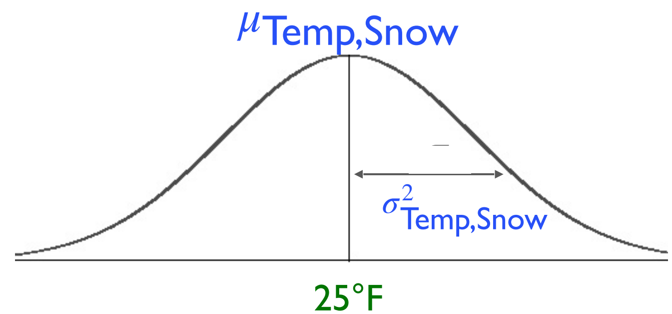
$\mu_{Temp, Snow}$ = (20+15+25)/3 = 20
$\sigma_{Temp, Snow}$ = 5

$\mu_{Temp, Sunny}$ = (57,63)/2 = 60
$\sigma_{Temp, Snow}$ = 4.24

Slide from Bruno Castro da Silva

Naive Bayes with Numeric Attributes





$\boldsymbol{\mu}_{\text{Temp, Sunny}} = (57,63)/2 = 60$

$\boldsymbol{\sigma}_{\text{Temp, Snow}} = 4.24$

$\boldsymbol{\mu}_{\text{Temp, Snow}} = (20+15+25)/3 = 20$

$\boldsymbol{\sigma}_{\text{Temp, Snow}} = 5$

$$\Pr(\text{Temp=61°F} \mid y_i = \text{Sunny}) = f(\,61;\, \boldsymbol{\mu}_{\text{Temp, Sunny}}\,,\, \boldsymbol{\sigma}_{\text{Temp, Sunny}}) = 0.091$$

$$\Pr(\text{Temp=18°F} \mid y_i = \text{Snow}) = f(\,18;\, \boldsymbol{\mu}_{\text{Temp, Snow}}\,,\, \boldsymbol{\sigma}_{\text{Temp, Snow}}) = 0.073$$

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