CS145: Probability & Computing

Lecture 3: Conditioning, Independence, Bayes' Rule, Bayesian Classification Algorithm

Instructors: Eli Upfal and Alessio Mazzetto

Brown University Computer Science

Figure credits:
Bertsekas & Tsitsiklis, ***Introduction to Probability****, 2008*
Pitman, ***Probability****, 1999*

CS145: Lecture 3 Outline

- Conditional Probability and Independence
- Bayes' Rule
- Bayesian Classification Algorithm

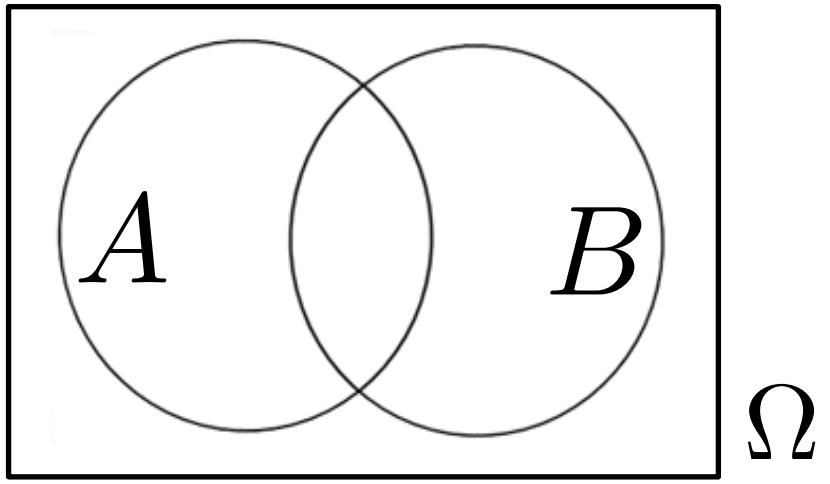
Conditional Probability

- $\mathbf{P}(A \mid B) =$ probability of A, given that B occurred
  - B is our new universe

- **Definition:** Assuming $\mathbf{P}(B) \neq 0$,

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

$\mathbf{P}(A \mid B)$ undefined if $\mathbf{P}(B) = 0$



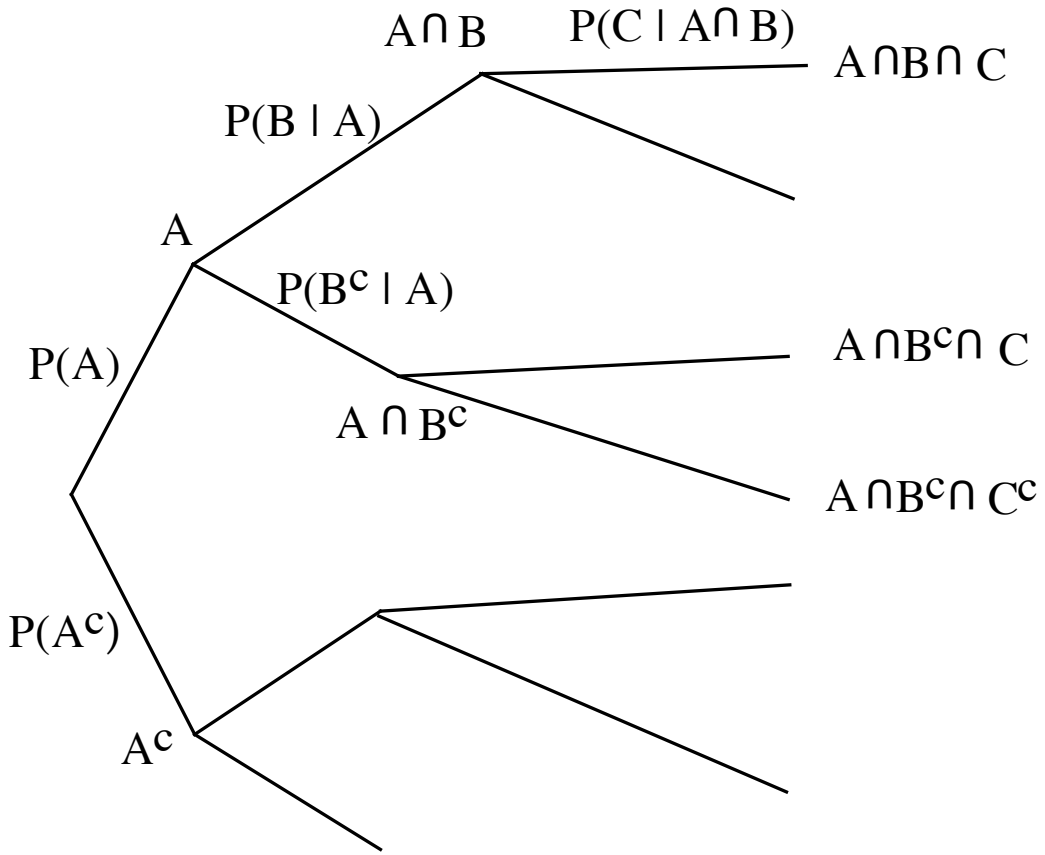
- Under discrete uniform law, where all outcomes equally likely:

$$\mathbf{P}(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B} = \frac{|A \cap B|}{|B|}$$

Multiplication Rule

$$\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A) \cdot \mathbf{P}(B \mid A) \cdot \mathbf{P}(C \mid A \cap B)$$

$$= P(A) \cdot \frac{P(A \cap B)}{P(A)} \cdot \frac{P(A \cap B \cap C)}{P(A \cap B)}$$



Multiplication Rule

Assuming that all of the conditioning events have positive probability, we have

$$\mathbf{P}\big(\cap_{i=1}^{n} A_i\big) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbf{P}\big(A_n \mid \cap_{i=1}^{n-1} A_i\big).$$

Example: Two-Sided Cards

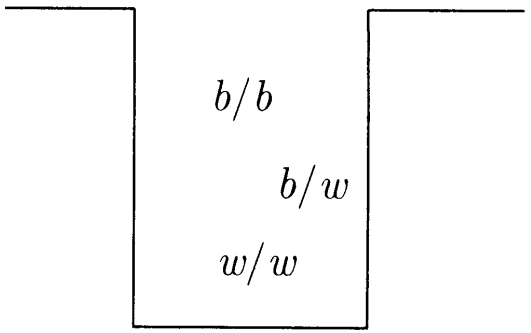
A hat contains three cards.

One card is black on both sides.

One card is white on both sides.

One card is black on one side and white on the other.

The cards are mixed up in the hat. Then a single card is drawn and placed on a table. If the visible side of the card is black, what is the chance that the other side is white?



Label the faces of the cards:

b_1 and b_2 for the black–black card;

w_1 and w_2 for the white–white card;

b_3 and w_3 for the black–white card.

$$\{\text{black on top}\} = \{b_1, b_2, b_3\}$$

$$\{\text{white on bottom}\} = \{b_3, w_1, w_2\}$$

$$P(\text{white on bottom}|\text{black on top}) = \frac{\#(\text{white on bottom and black on top})}{\#(\text{black on top})} = \frac{1}{3}$$

Observing the top face of the card provides information about the color of the bottom face.

Example: Roll of a 6-Sided Die

- The probability of "the outcome of a die roll is *even*" is $\frac{3}{6}$
- The probability of the event "the outcome is ≤ 4" is $\frac{4}{6}$.

$$P(\text{die even} \mid \text{die } \leq 4) = \frac{2}{4} = \frac{1}{2} = P(\text{die even})$$

Observing that the die roll was at most 4
does not provide information about whether it was even.

Independence

Two equivalent definitions of events independence:

- **"Defn:"** $\mathbf{P}(B \mid A) = \mathbf{P}(B)$
  - "occurrence of A provides no information about B's occurrence"

- Recall that $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B \mid A)$
- **Defn:** $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$
- Symmetric with respect to A and B
  - applies even if $\mathbf{P}(A) = 0$
  - implies $\mathbf{P}(A \mid B) = \mathbf{P}(A)$

Set of events $\{A_1, \ldots, A_n\}$ are **Mutually Independent** if

$$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i).$$

Sources of Independence

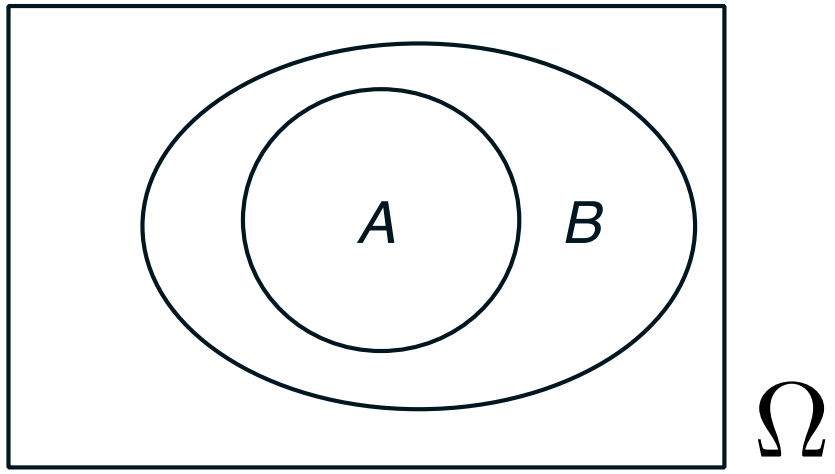
- Event associated with "independent" physical processes are independent.
- But independent events do not have to be related to independent physical processes.
- Example:
  - The probability of "the outcome of a die roll is *even*" is $\frac{3}{6}$
  - The probability of the event "the outcome is ≤ 4" is $\frac{4}{6}$.
  - The probability of "an even outcome ≤ 4" is

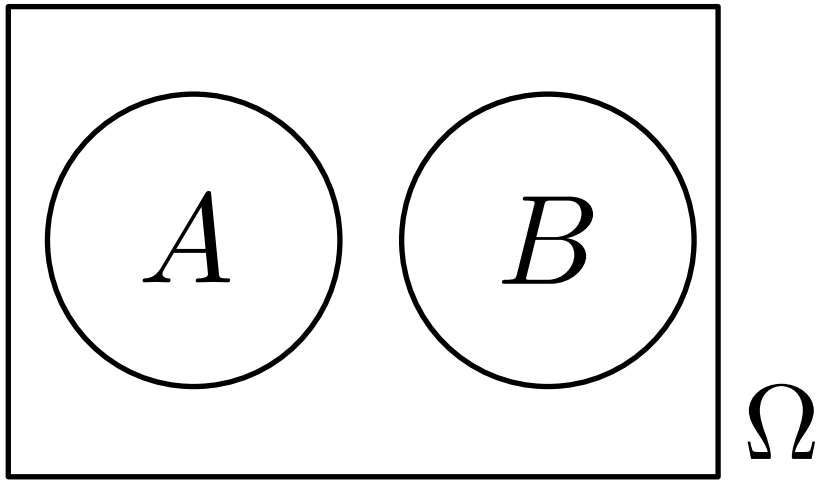
$$\frac{2}{6} = \frac{12}{36} = \frac{3}{6} \cdot \frac{4}{6}$$

$\Rightarrow$ the two events are independent.

- The "intuition" here is that there are the same number of odd and even outcomes that are ≤ 4. Thus, the "information" that the outcome is ≤ 4 does not "help" in deciding if it is odd or even.

Events that are NOT Independent





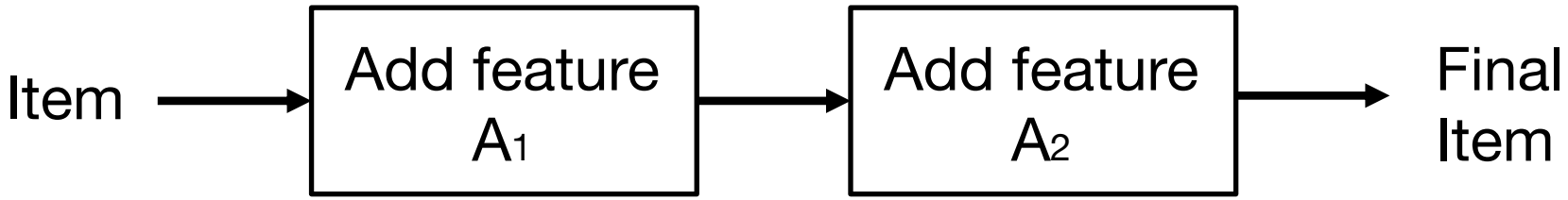
- Assume events are non-degenerate: $0 < P(A) < 1, 0 < P(B) < 1$
- Nested events are not independent:

$$\text{If } A \subset B, \quad P(B \mid A) = 1 \neq P(B).$$

- Mutually exclusive events are not independent:

$$\text{If } A \cap B = \emptyset, \quad P(A \cap B) = 0 \neq P(A)P(B).$$

Example: Assembly line

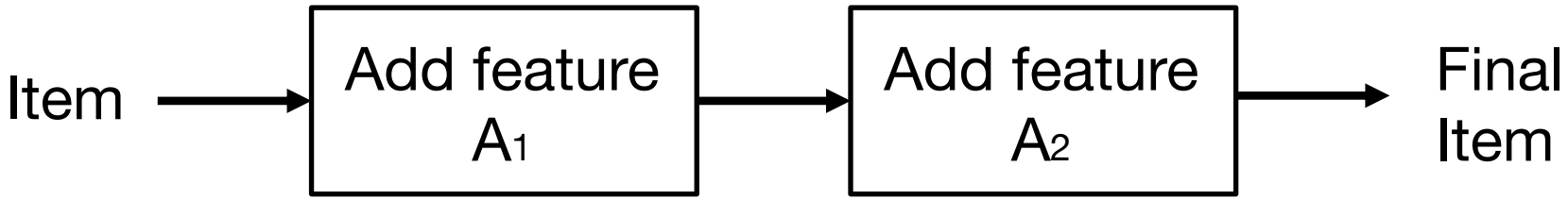


Let A_i be the event that the final item has a defect on feature i.

$P(A_1) = 0.8, \quad P(A_2) = 0.7$

We assume that A_1 and A_2 are **independent events**.

Example: Assembly line



Let A_i be the event that the final item has a defect on feature i.

$P(A_1) = 0.3, \quad P(A_2) = 0.2$

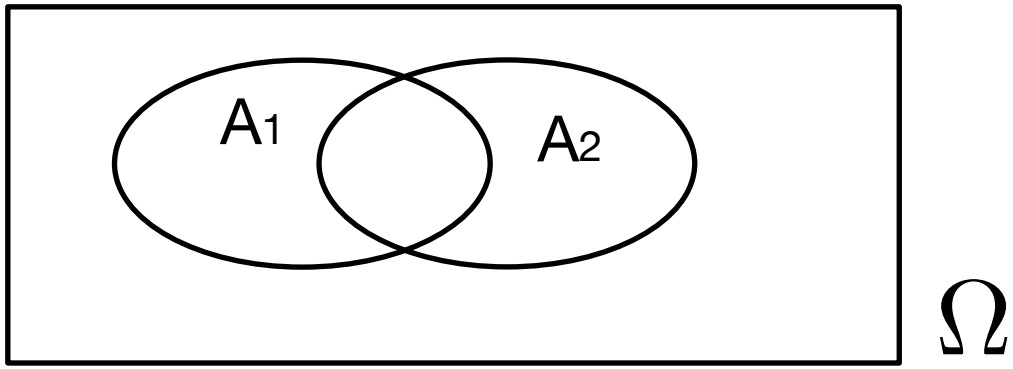
We assume that A_1 and A_2 are **independent events**.

$W = \{\text{no defect}\}$

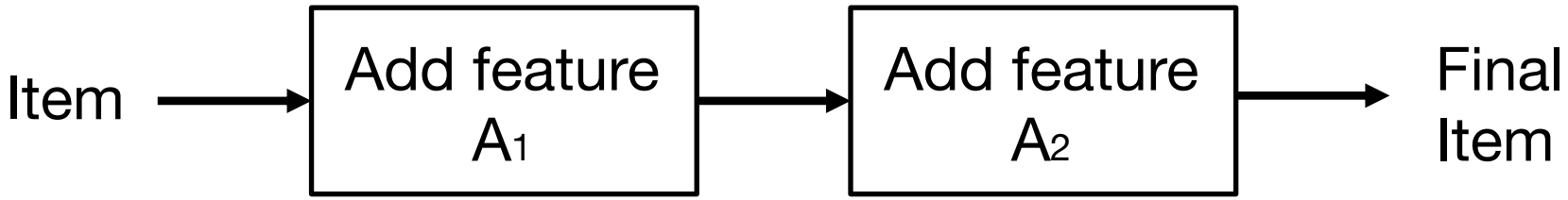
$W = A_1^c \cap A_2^c = \Omega \setminus (A_1 \cup A_2)$

$P(W) = 1 - P(A_1 \cup A_2)$

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$
$$= P(A_1) + P(A_2) - P(A_1) \cdot P(A_2)$$



Example: Assembly line



Let A_i be the event that the final item has a defect on feature i.

$P(A_1) = 0.3, \quad P(A_2) = 0.2$

We assume that A_1 and A_2 are **independent events**.

$W = \{\text{no defect}\}$

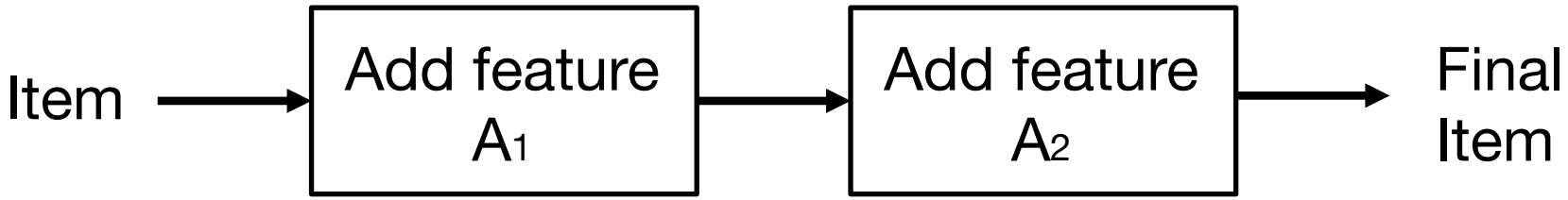
$W = A_1^c \cap A_2^c = \Omega \setminus (A_1 \cup A_2)$

$P(W) = 1 - P(A_1 \cup A_2)$

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1) \cdot P(A_2) \end{aligned}$$

$$\begin{aligned} P(W) &= 1 - P(A_1) - P(A_2) \cdot (1 - P(A_1)) \\ &= (1 - P(A_1)) \cdot (1 - P(A_2)) \\ &= P(A_1^c) \cdot P(A_2^c) = 0.56 \end{aligned}$$

Example: Assembly line



Let A_i be the event that the final item has a defect on feature i.

$P(A_1) = 0.3, \quad P(A_2) = 0.2$

We assume that A_1 and A_2 are **independent events**.

$W = \{\text{no defect}\}$

$W = A_1^c \cap A_2^c = \Omega \setminus (A_1 \cup A_2)$

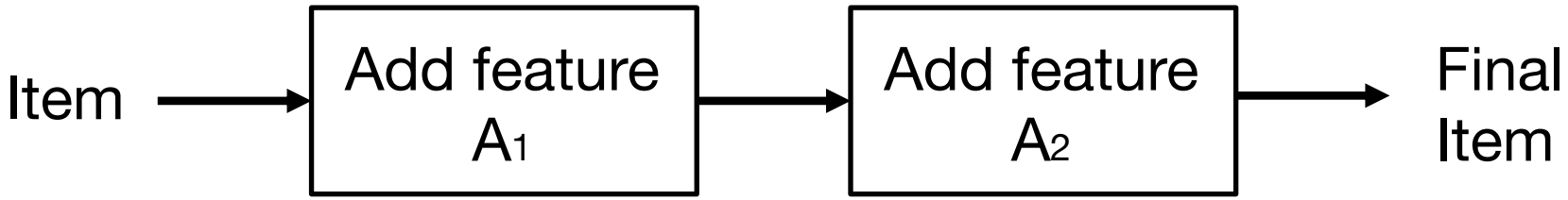
$P(W) = 1 - P(A_1 \cup A_2)$

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$
$$= P(A_1) + P(A_2) - P(A_1) \cdot P(A_2)$$

$$P(W) = 1 - P(A_1) - P(A_2) \cdot (1 - P(A_1))$$
$$= (1 - P(A_1)) \cdot (1 - P(A_2))$$
$$= P(A_1^c) \cdot P(A_2^c) = 0.56$$

Complements maintain independence

Example: Assembly line



Let A_i be the event that the final item has a defect on feature i.

$P(A_1) = 0.3, \quad P(A_2) = 0.2$

We assume that A_1 and A_2 are **independent events**.

$W = \{\text{defect only on feature } A_2\}$

$W = A_1^c \cap A_2$

A_1 independent with $A_2 \Rightarrow A_1^c$ independent with A_2

$P(W) = P(A_1^c \cap A_2) = P(A_1^c) \cdot P(A_2) = (1 - P(A_1))P(A_2) = 0.7 \cdot 0.2 = 0.14$

Serial versus Parallel Systems

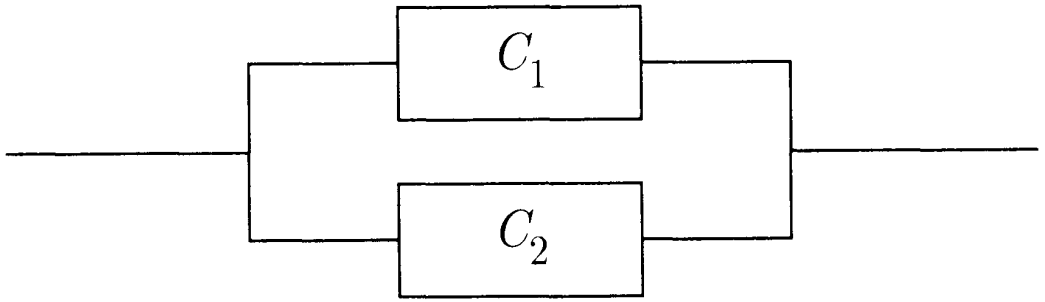
Let W_i be the event that component C_i works without failure

$$\{\text{system works}\} = W_1 \cap W_2$$

Assume component failures are independent events, and that

$$P(W_1) = 0.9 \text{ and } P(W_2) = 0.8$$

$$P(\text{system works}) = P(W_1 W_2) = P(W_1)P(W_2) = 0.9 \times 0.8 = 0.72$$

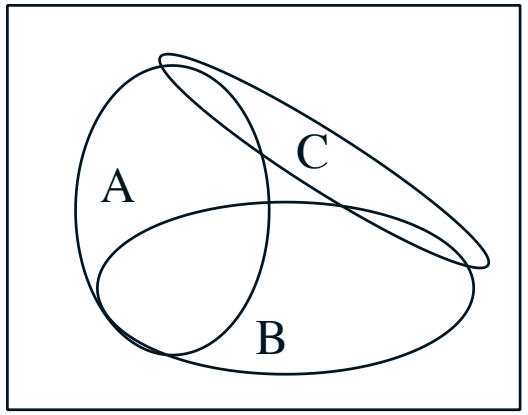


$$\{\text{system works}\} = W_1 \cup W_2$$

$$P(\text{system works}) = 1 - (0.1)(0.2) = 0.98$$

Conditioning may Affect Independence

- Conditional independence, given C, is defined as independence under probability law $\mathbf{P}(\cdot \mid C)$
- Assume A and B are independent



- If we are told that C occurred, are A and B independent?

Definition of conditional independence:

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$$

For this example:

$$P(A \cap B) = P(A)P(B)$$

$$P(A \cap B \mid C) = 0 \neq P(A \mid C)P(B \mid C)$$

Conditioning may Affect Independence

A - dice outcome is even
B - dice outcome is <= 4

Definition of conditional independence:

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$$

$$P(A \cap B) = P(A)P(B) = \frac{1}{2}\frac{4}{6} = \frac{2}{3}$$

C= dice outcome >1

$$P(A \cap B \mid C) = P(A \cap B \cap C)/P(C) = \frac{2/6}{5/6} = \frac{2}{5}$$

$$\neq \; P(A \mid C)\, P(B \mid C) = \frac{3}{5}\frac{3}{5} = \frac{9}{25}$$

Example: Conditioning & Independence

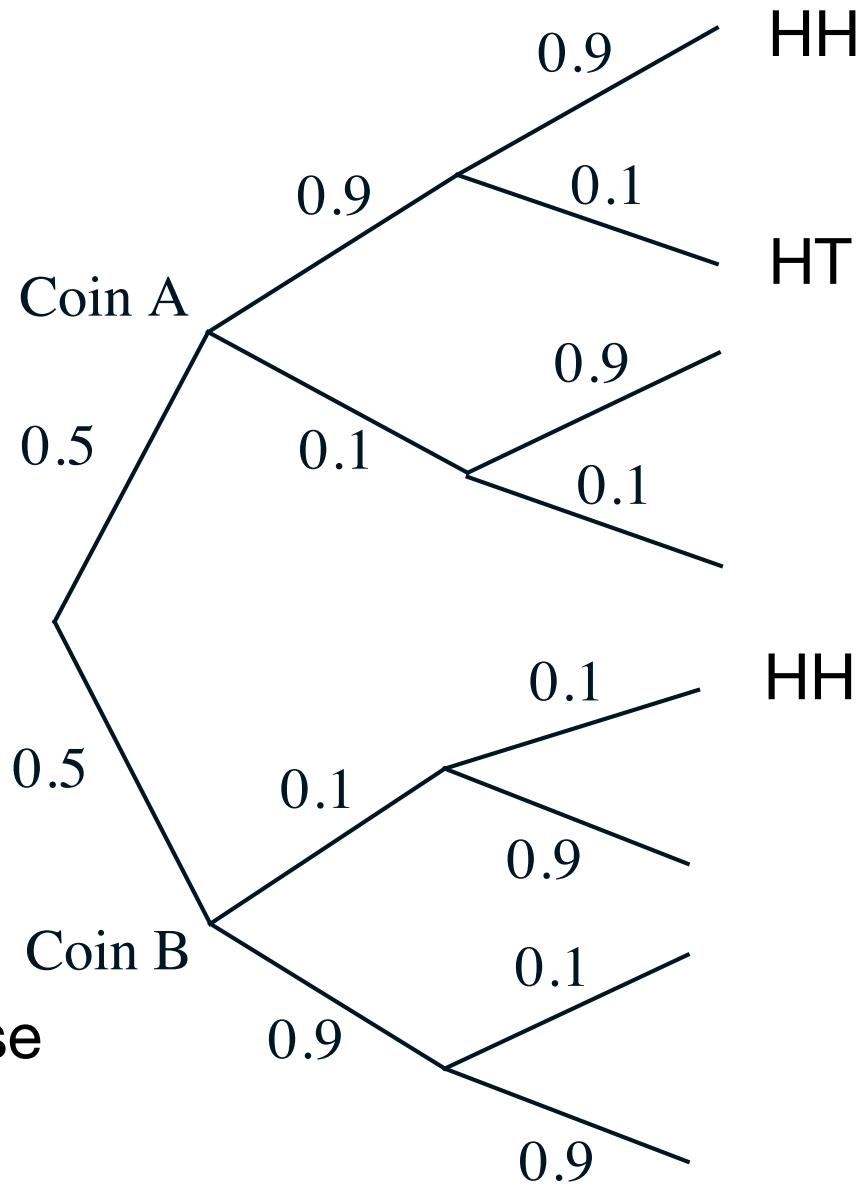
- Two unfair coins, A and B:
  $\mathbf{P}(H \mid \text{coin } A) = 0.9$, $\mathbf{P}(H \mid \text{coin } B) = 0.1$
  choose either coin with equal probability

- Once we know it is coin A, are tosses independent?

  Yes, by definition.

- If we do not know which coin it is, are tosses independent?

  No, consider probability that second toss is heads given first.



H in the first toss increases the probability that we chose the first coin and therefore of H is the second toss

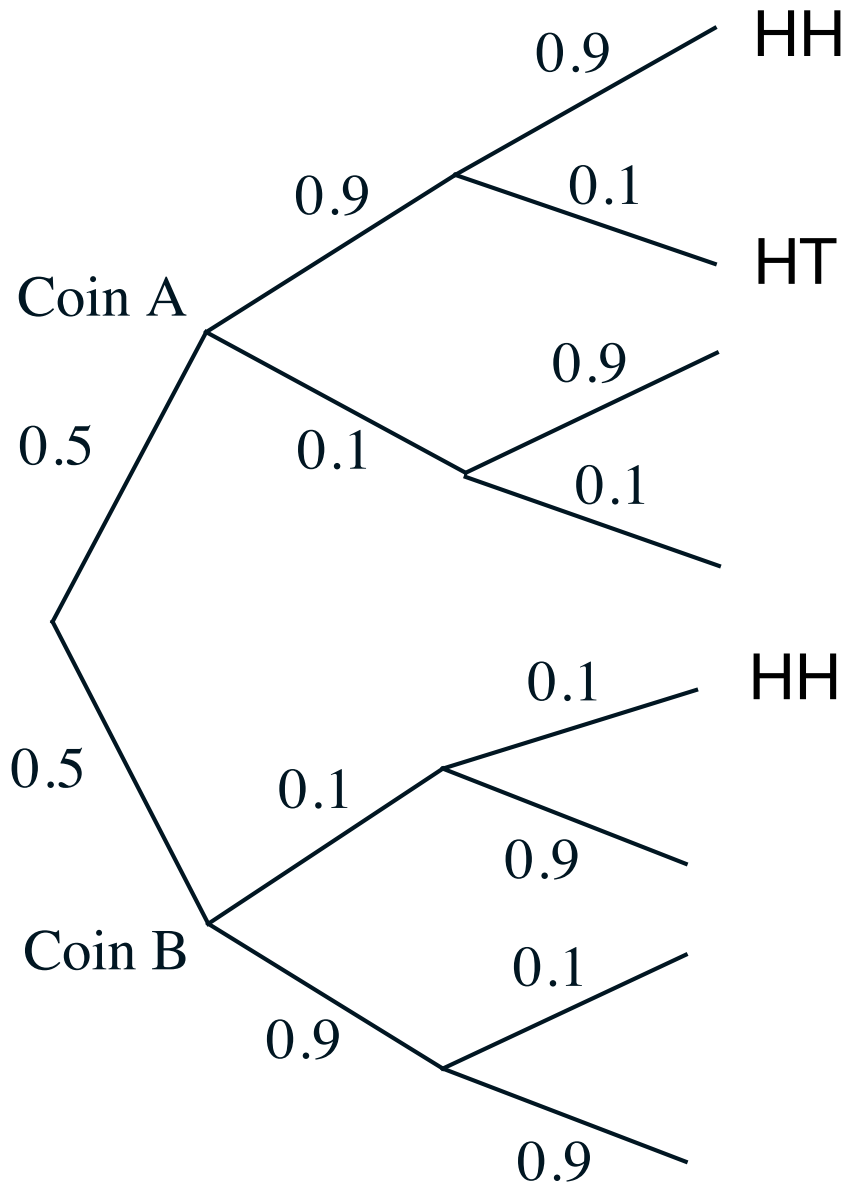
Example: Conditioning & Independence

- Two unfair coins, A and B:
  $\mathbf{P}(H \mid \text{coin } A) = 0.9$, $\mathbf{P}(H \mid \text{coin } B) = 0.1$
  choose either coin with equal probability
  - Once we know it is coin A, are tosses independent? *Yes, by definition.*
  - If we do not know which coin it is, are tosses independent?

No, consider probability that second toss is head given the first.

$H1$ - H in the first toss, $H2$ - H in the second toss.

$$P(H2 \mid H1) = \frac{P(H1 \cap H2)}{P(H1)} = \frac{0.5(0.9)^2 + 0.5(0.1)^2}{0.5 \cdot 0.9 + 0.5 \cdot 0.1} = 0.82$$



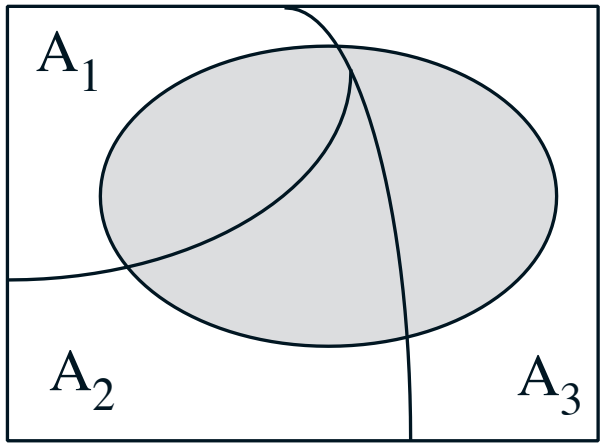
CS145: Lecture 3 Outline

- Conditional Probability and Independence
- Bayes' Rule
- Bayesian Classification Algorithm

T. Bayes.

Total Probability Theorem

- Divide and conquer
- Partition of sample space into A_1, A_2, A_3
- Have $\mathbf{P}(B \mid A_i)$, for every i



$$
\begin{aligned}
P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\
P(A \cap B) &= P(B)P(A \mid B) \\
\\
B &= (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \\
P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) \\
P(B) &= P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + P(A_3)P(B \mid A_3)
\end{aligned}
$$

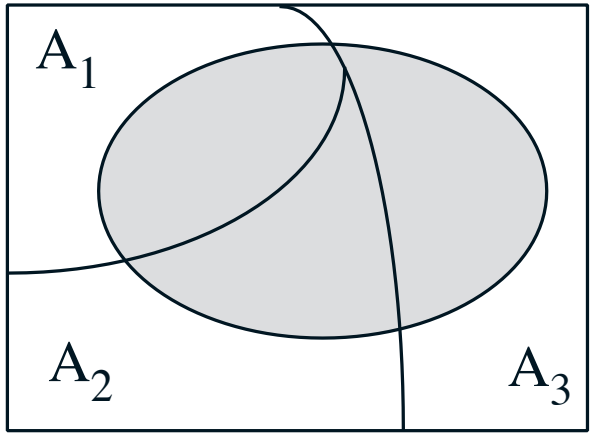
- One way of computing $\mathbf{P}(B)$:

$$
\begin{aligned}
\mathbf{P}(B) = \quad & \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) \\
+ \quad & \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) \\
+ \quad & \mathbf{P}(A_3)\mathbf{P}(B \mid A_3)
\end{aligned}
$$

Bayes' Rule

- "Prior" probabilities $\mathbf{P}(A_i)$
  - initial "beliefs"
- We know $\mathbf{P}(B \mid A_i)$ for each i
- Wish to compute $\mathbf{P}(A_i \mid B)$
  - revise "beliefs", given that B occurred

T. Bayes.



$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\sum_j \mathbf{P}(A_j)\mathbf{P}(B \mid A_j)}$$

Example: Witness reliability

A witness testifies that they saw a robber escaping with a yellow taxi

How reliable is this information?

- Taxi are either red or yellow
- A witness is accurate 80% of the times (independent of the color)

Example: Witness reliability

A witness testifies that they saw a robber escaping with a yellow taxi

How reliable is this information?

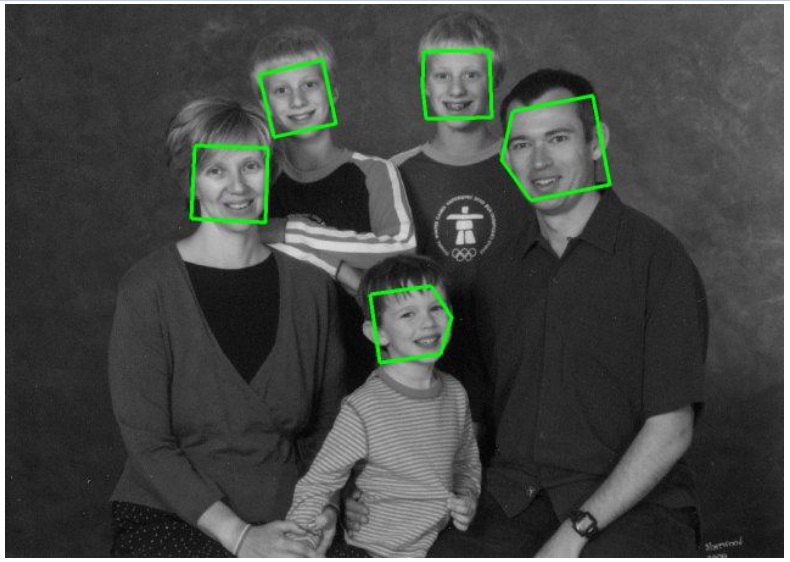
- Taxi are either red or yellow
- A witness is accurate 80% of the times (independent of the color)
- 70% of taxi are red and 30% of taxi are yellow

$$T = \{\text{taxi of the robber is yellow}\}, \quad W = \{\text{witness says yellow}\}$$

$$P(T|W) = P(W|T)\frac{P(T)}{P(W)} = \frac{P(W|T) \cdot P(T)}{P(W|T)P(T) + P(W|T^c)P(T^c)}$$

$$= \frac{0.8 \cdot 0.3}{0.8 \cdot 0.3 + 0.7 \cdot 0.2} \simeq 0.63$$

Example: Face Detection

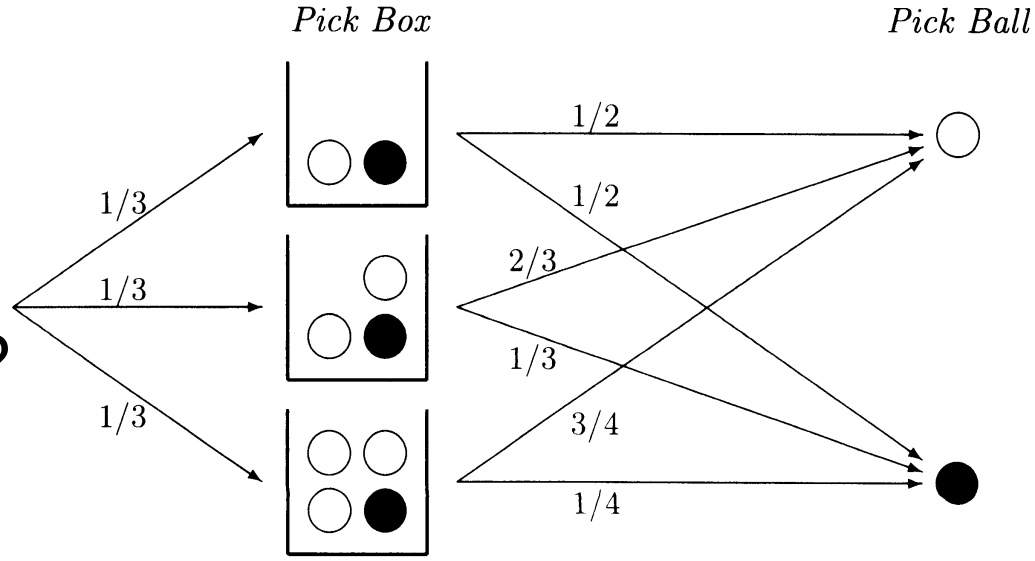


CS145: Lecture 3 Outline

- Conditional Probability and Independence
- Bayes' Rule
- Bayesian Classification Algorithm

Boxes and Balls

- Three boxes, box *i* contains *i* white balls and one black ball
- I pick one of the boxes at random, then randomly draw one of its balls
- If I show you that the ball I drew was white, what box would you guess it came from?
- With what probability is your guess correct?



$$P(\text{Box } i|\text{white}) = \frac{P(\text{Box } i \text{ and white})}{P(\text{white})} \qquad (i = 1, 2, 3)$$

$$P(\text{Box } i \text{ and white}) = P(\text{Box } i)P(\text{white}|\text{Box } i) = \frac{1}{3} \times \frac{i}{i+1} \quad (i = 1, 2, 3)$$

$$P(\text{white}) = \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{3}{4} = \frac{23}{36}$$

i	1	2	3
$P(\text{Box } i\|\text{white})$	6/23	8/23	9/23

Classification Problems

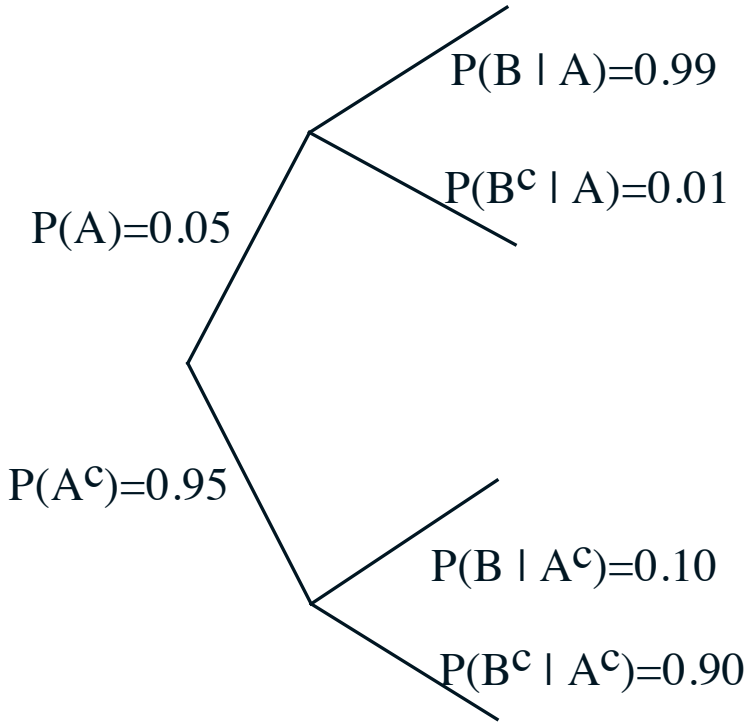
Athitsos et al., CVPR 2004 & PAMI 2008

- Which of the 10 digits did a person write by hand?
- Is an email spam or not spam (ham)?
- Is this image taken in an indoor our outdoor environment?
- Is a pedestrian visible from a self-driving car's camera?
- What language is a webpage or document written in?
- How many stars would a user rate a movie that they've never seen?

Models Based on Conditional Probabilities

Event A: An airplane is flying above

Event B: A "blip" appears on radar



Bayesian Classifier

- If I observe a blip *B*, predict *A* if and only if

$$P(A \mid B) > P(A^c \mid B)$$

$$(\text{or, } P(A \mid B) > 0.5)$$

- If I observe no blip B^c, predict *A* if and only if

$$P(A \mid B^c) > P(A^c \mid B^c)$$

$$(\text{or, } P(A \mid B^c) > 0.5)$$

Blip: $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

No Blip: $$P(A \mid B^c) = \frac{P(B^c \mid A)P(A)}{P(B^c)}$$

A Simplified Classification Rule

- If I observe *B*, I will predict *A* is true if and only if:

$$P(A \mid B) > P(A^c \mid B)$$

- By Bayes' rule, this is equivalent to checking:

$$\frac{P(B \mid A)P(A)}{P(B)} > \frac{P(B \mid A^c)P(A^c)}{P(B)}$$

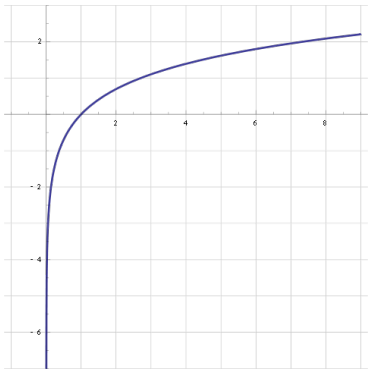
- Because *P(B) > 0*, I can ignore the denominator, and check:

$$P(B \mid A)P(A) > P(B \mid A^c)P(A^c)$$

- Because the *logarithm* is monotonic:

$$\log P(B \mid A) + \log P(A) > \log P(B \mid A^c) + \log P(A^c)$$

More numerically robust when probabilities small.



Testing: How good is my classifier?

- Suppose I have a dataset of *M* labeled test examples:

$$(A_i, B_i), i = 1, \ldots, M \qquad A_i \in \{0, 1\}$$

- For each test example, the classifier makes a prediction about A_i given the information from B_i:

$$\text{Predict } \hat{A}_i = 1 \text{ if}$$

$$\log P(B_i \mid A_i = 1) + \log P(A_i = 1) > \log P(B_i \mid A_i = 0) + \log P(A_i = 0)$$

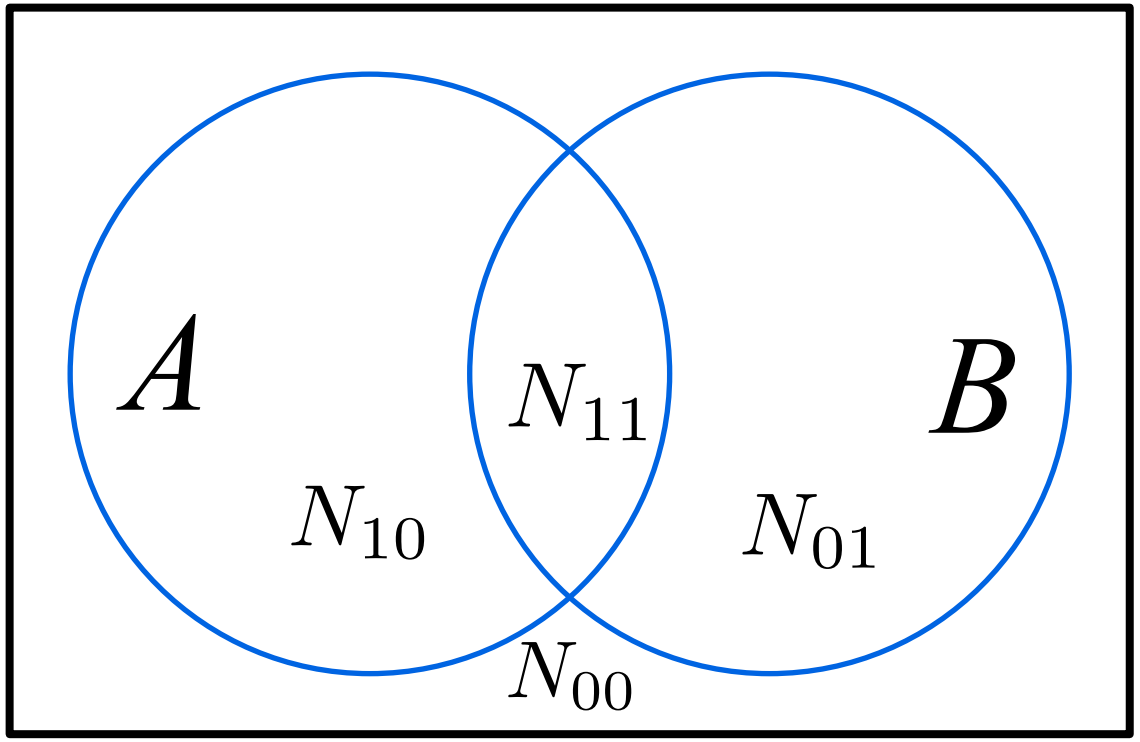
$$\text{Otherwise, predict } \hat{A}_i = 0.$$

- The test accuracy of our classifier is then

$$\text{accuracy} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(\hat{A}_i = A_i) \qquad \text{error-rate} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(\hat{A}_i \neq A_i)$$

Training: What are the probabilities?

$$\log P(B \mid A) + \log P(A) > \log P(B \mid A^c) + \log P(A^c)$$



$$N = N_{11} + N_{10} + N_{01} + N_{00}$$

A simple way to estimate probabilities is to *count frequencies of training events*:

$$P(A) = (N_{10}+N_{11})/N$$
$$P(A^c) = (N_{00}+N_{01})/N$$
$$P(B \mid A) = N_{11}/(N_{11}+N_{10})$$
$$P(B \mid A^c) = N_{01}/(N_{00}+N_{10})$$

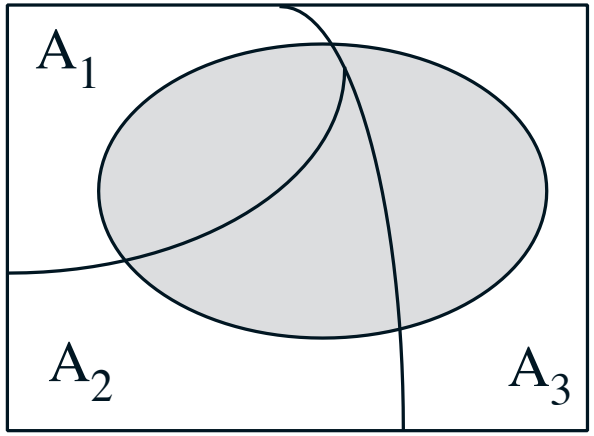
The “Naïve” Bayesian Classifier

- We often classify based on a set of observations.
- Example: classify the subject of a document based on a set of keywords
- We have a set of subjects $S=\{s_1, s_2,\ldots,s_m\}$ and a set of keyworks $W=\{w_1, w_2,\ldots,w_n\}$
- A document d is represented by a Boolean vector $b(d)=(b_1, b_2,\ldots,b_n)$ where $b_i=1$ if word w_i is in the document d.
- To apply the Bayesian classification method we need for every subject s and Boolean vector b an estimate of $P(b \mid s)$ - we need to estimate $2^n|S|$ probabilities – not practical.
- Instead – we assume that occurrences of keywords in a document are “independent” events, $P(b \mid s) = \prod_i P(b_i \mid s)$. In that case we need to estimate only $n|S|$ probabilities.
- While the assumption in naïve, it works very well in practice.

Review on Bayes Rule

Review on Bayes Rule

- "Prior" probabilities $\mathbf{P}(A_i)$
  - initial "beliefs"
- We know $\mathbf{P}(B \mid A_i)$ for each i
- Wish to compute $\mathbf{P}(A_i \mid B)$
  - revise "beliefs", given that B occurred

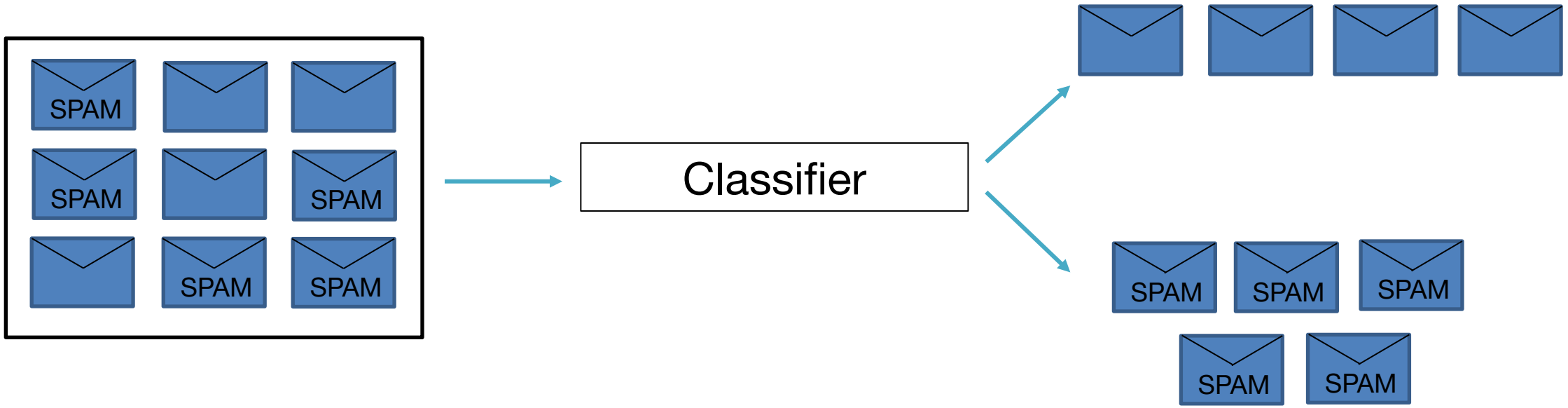


$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\sum_j \mathbf{P}(A_j)\mathbf{P}(B \mid A_j)}$$

Example: Spam Detection



Example: Spam Detection

- We want to build a spam detection system
- We have access to a huge dataset of emails that are labeled as 'spam' or 'not spam'
- We identify a set of keywords $W = \{w_1, \ldots, w_n\}$ that are discriminative (e.g., most likely to appear if email is spam)
- We describe an email through a Boolean vector $B = (b_1, \ldots, b_n)$ where $b_i = 1$ if and only if the keyword w_i appears in the document

Example: Spam Detection

$S = \{\text{email is spam}\}, \quad B = (b_1, \ldots, b_n)$

Bayesian Classifier: classify spam iff $\Pr(S|B) > \Pr(S^c|B)$

Bayes Rule

$$\Pr(S|B) > \Pr(S^c|B) \iff \Pr(B|S)\Pr(S) > \Pr(B|S^c)\Pr(S^c)$$

Example: Spam Detection

$S = \{\text{email is spam}\}, \quad B = (b_1, \ldots, b_n)$

Bayesian Classifier: classify spam iff $\Pr(S|B) > \Pr(S^c|B)$

Bayes Rule

$$\Pr(S|B) > \Pr(S^c|B) \iff \Pr(B|S)\boxed{\Pr(S)} > \Pr(B|S^c)\boxed{\Pr(S^c)}$$

Easy to estimate from dataset

Example: Spam Detection

$S = \{\text{email is spam}\}, \quad B = (b_1, \ldots, b_n)$

Bayesian Classifier: classify spam iff $\Pr(S|B) > \Pr(S^c|B)$

Bayes Rule

$$\Pr(S|B) > \Pr(S^c|B) \iff \Pr(B|S)\boxed{\Pr(S)} > \Pr(B|S^c)\boxed{\Pr(S^c)}$$

Easy to estimate from dataset

Naive Assumption: given the true classification, the words appear independently (conditional independence)

$$\Pr(B|S) = \prod_{i=1}^{n} \Pr(b_i|S)$$

Hard to estimate, 2^n possible keywords combination

We only need to estimate n probabilities (event that a keyword appears if the email is spam)

Example: Spam Detection

Probability appearing	w_1	w_2	…	…	…	w_n
Spam	0.7	0.4				0.2
Not Spam	0.1	0.7				0.6

$$B = (1, 0, \ldots, 1)$$

$$\Pr(B|S) = \prod_{i=1}^{n} \Pr(b_i|S) = \boxed{}$$

Example: Spam Detection

Probability appearing	w_1	w_2	…	…	…	w_n
Spam	0.7	0.4				0.2
Not Spam	0.1	0.7				0.6

$B = (1, 0, \ldots, 1)$

1 0 1

$$\Pr(B|S) = \prod_{i=1}^{n} \Pr(b_i|S) = 0.7 \cdot 0.6 \cdot \ldots \cdot 0.2$$