CS145: Probability & Computing
Lecture 3: Conditioning, Independence, Bayesian Classification Algorithm

Instructor: Eli Upfal
Brown University Computer Science

Figure credits:
Bertsekas & Tsitsiklis, Introduction to Probability, 2008
Pitman, Probability, 1999
CS145: Lecture 3 Outline

- Conditional Probability and Independence
- Bayesian Classification Algorithm
Conditional Probability

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

- **Definition:** Assuming \( P(B) \neq 0 \),
  \[
  P(A \mid B) = \frac{P(A \cap B)}{P(B)}
  \]
  \( P(A \mid B) \) undefined if \( P(B) = 0 \)

- Under discrete uniform law, where all outcomes equally likely:
  \[
  P(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B} = \frac{|A \cap B|}{|B|}
  \]
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 \( \leq 4 \)" 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 of Two Events**

- **“Defn:”** \( P(B \mid A) = P(B) \)
  - “occurrence of \( A \) provides no information about \( B \)'s occurrence”

- Recall that \( P(A \cap B) = P(A) \cdot P(B \mid A) \)

- **Defn:** \( P(A \cap B) = P(A) \cdot P(B) \)

- Symmetric with respect to \( A \) and \( B \)
  - applies even if \( P(A) = 0 \)
  - implies \( P(A \mid B) = P(A) \)
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 \( \leq 4 \)" is \( \frac{4}{6} \).
  • The probability of "an even outcome \( \leq 4 \)" 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 \( \leq 4 \). Thus, the "information" that the outcome is \( \leq 4 \) 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:
  If $A \subset B$, $P(B \mid A) = 1 \neq P(B)$.
- Mutually exclusive events are not independent:
  If $A \cap B = \emptyset$, $P(A \cap B) = 0 \neq P(A)P(B)$. 

\[ \begin{align*}
\Omega & \quad \Omega \\
A & \subset B \\
\Omega & \quad \Omega \\
A & \quad B
\end{align*} \]
Serial versus Parallel Systems

Assume component failures are independent events, and that 

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

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

\[
\{\text{system works}\} = W_1 \cup W_2
\]

\[
P(\text{system works}) = 1 - (0.1)(0.2) = 0.98
\]
CS145: Lecture 3 Outline

- Conditional Probability and Independence
- 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?

\[
P(\text{Box } i|\text{white}) = \frac{P(\text{Box } i \text{ and white})}{P(\text{white})} \quad (i = 1, 2, 3)
\]

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

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

<table>
<thead>
<tr>
<th>(i)</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(\text{Box } i</td>
<td>\text{white}))</td>
<td>6/23</td>
<td>8/23</td>
</tr>
</tbody>
</table>
Total Probability and Bayes’ Theorem

- Partition of sample space into $A_1, A_2, A_3$
- Have $\mathbb{P}(B | A_i)$, for every $i$
  
  ![Diagram showing partition of sample space into $A_1, A_2, A_3$.]

- One way of computing $\mathbb{P}(B)$:
  
  $$
  \mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B | A_1) + \mathbb{P}(A_2)\mathbb{P}(B | A_2) + \mathbb{P}(A_3)\mathbb{P}(B | A_3)
  $$

- “Prior” probabilities $\mathbb{P}(A_i)$
  - initial “beliefs”
- Wish to compute $\mathbb{P}(A_i | B)$
  - revise “beliefs”, given that $B$ occurred

  $$
  \mathbb{P}(A_i | B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}
  $$

  $$
  = \frac{\mathbb{P}(A_i)\mathbb{P}(B | A_i)}{\mathbb{P}(B)}
  $$

  $$
  = \frac{\mathbb{P}(A_i)\mathbb{P}(B | A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B | A_j)}
  $$
Classification Problems

- Which of the 10 digits did a person write by hand?
- Which of the basic ASL phonemes did a person sign?
- Is an email spam or not spam (ham)?
- Is this image taken in an indoor or 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

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

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

### Bayesian Classifier

- If I observe a blip \( B \), predict \( A \) if and only if \( P(A \mid B) > P(A^c \mid B) \)
  
  (or, \( P(A \mid B) > 0.5 \))

- If I observe no blip \( B^c \), predict \( A \) if and only if \( P(A \mid B^c) > P(A^c \mid B^c) \)
  
  (or, \( P(A \mid B^c) > 0.5 \))
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:

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

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

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

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

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

The test accuracy of our classifier is then

$$\text{accuracy} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(\hat{A}_i = A_i) \quad \text{error-rate} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(\hat{A}_i \neq A_i)$$
**Training: What are the probabilities?**

<table>
<thead>
<tr>
<th>Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Data (size N)</strong></td>
<td><strong>Test Data (size M)</strong></td>
</tr>
</tbody>
</table>

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

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

\[
\begin{align*}
P(A) &= \\
P(A^c) &= \\
P(B \mid A) &= \\
P(B \mid A^c) &= 
\end{align*}
\]

Separate training & test data (model fitting & model checking)

\[N = N_{11} + N_{10} + N_{01} + N_{00}\]