CMPSCI 240: "Reasoning Under Uncertainty"

Lecture 9

Prof. Hanna Wallach
wallach@cs.umass.edu

February 21, 2012

Reminders

- Check the course website: `http://www.cs.umass.edu/~wallach/courses/s12/cmpsci240/`
- Third homework is due on Friday

Information Theory

Communication

- Communication: representing and transmitting information

Communication

- Communication: representing and transmitting information
- "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." — Claude Shannon

Communication

- Communication: representing and transmitting information
- "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." — Claude Shannon
- Information theory: the math behind this

Uses of Information Theory

- Measuring information content

Uses of Information Theory

- Measuring information content
- Compressing data (e.g., images, video, text, etc.)

Uses of Information Theory

- Measuring information content
- Compressing data (e.g., images, video, text, etc.)
- Communicating without errors over noisy, imperfect communication channels (e.g., phone lines, disk drives, etc.)

Information Content

Probability and Information: Intuition

- Probability and information content are inversely related

Probability and Information: Intuition

- Probability and information content are inversely related
- Less probable events have a greater information content

Probability and Information: Intuition

- Probability and information content are inversely related
- Less probable events have a greater information content
- The more certain you are that some event will occur, the less information you gain by knowing that it did indeed occur

Probability and Information: Intuition

- Probability and information content are inversely related
- Less probable events have a greater information content
- The more certain you are that some event will occur, the less information you gain by knowing that it did indeed occur
- e.g., sun rising vs. professor making a TurBacHenDuckEn

Desirable Properties of Information Content

- If $P(A) = 1$ and $P(A^c) = 0$, the information gained by learning that the outcome x is in A should be zero and the information gained by learning that $x \in A^c$ should be infinite

Desirable Properties of Information Content

- If $P(A) = 1$ and $P(A^c) = 0$, the information gained by learning that the outcome x is in A should be zero and the information gained by learning that $x \in A^c$ should be infinite
- If $P(A) = P(A^c) = 1 / 2$, we are maximally uncertain about the outcome and we should gain one unit of information by learning that the event is in A and similarly for A^c

Asking Informative Questions

- If $A_1, \ldots, A_n$ have probabilities $P(A_1), \ldots, P(A_n)$ and partition Ω, how would you identify the event containing outcome x using as few yes/no questions as possible?

Asking Informative Questions

- If $A_1, \ldots, A_n$ have probabilities $P(A_1), \ldots, P(A_n)$ and partition Ω, how would you identify the event containing outcome x using as few yes/no questions as possible?
- e.g., suppose $n = 4$ and $P(A_i) = 1 \,/\, 4$ for $i = 1, \ldots, 4$

Asking Informative Questions

- If $A_1, \ldots, A_n$ have probabilities $P(A_1), \ldots, P(A_n)$ and partition Ω, how would you identify the event containing outcome x using as few yes/no questions as possible?
- e.g., suppose $n = 4$ and $P(A_i) = 1 / 4$ for $i = 1, \ldots, 4$
- e.g., suppose $n = 4$ again but now we have $P(A_1) = 1 / 2$, $P(A_2) = 1 / 4$, $P(A_3) = 1 / 8$, and $P(A_4) = 1 / 8$

Asking Informative Questions

- Ask the questions that you will gain the most information from regardless of the answer obtained, i.e., those questions whose answers you have the most uncertainty about

Asking Informative Questions

- Ask the questions that you will gain the most information from regardless of the answer obtained, i.e., those questions whose answers you have the most uncertainty about
- Ask the questions that will eliminate the most remaining possibilities regardless of the answers obtained, i.e., those questions that divide the remaining probability equally

Asking Informative Questions

- Think of the equiprobable yes/no questions as defining a (weighted, binary) tree with events $A_1, \ldots, A_n$ as leaves

Asking Informative Questions

- Think of the equiprobable yes/no questions as defining a (weighted, binary) tree with events $A_1, \ldots, A_n$ as leaves
- The depth of A_i is equal to the number of equiprobable yes/no questions required to uniquely determine that $x \in A_i$

Asking Informative Questions

- Think of the equiprobable yes/no questions as defining a (weighted, binary) tree with events $A_1, \ldots, A_n$ as leaves
- The depth of A_i is equal to the number of equiprobable yes/no questions required to uniquely determine that $x \in A_i$
- Each yes/no question divides the remaining probability in half, so we know that if event A_i is at depth k, its probability $P(A_i)$ must be equal to $(1\,/\,2)^k = (2^{-1})^k = 2^{-k}$

Information Content

- Information content of A_i: number of equiprobable yes/no questions required to uniquely determine that $x \in A_i$

Information Content

- Information content of A_i: number of equiprobable yes/no questions required to uniquely determine that $x \in A_i$
- The information content $I(A_i)$ of event A_i is therefore

$$I(A_i) = \log_2 \frac{1}{P(A_i)}$$

Information Content

- Information content of A_i: number of equiprobable yes/no questions required to uniquely determine that $x \in A_i$
- The information content $I(A_i)$ of event A_i is therefore

$$I(A_i) = \log_2 \frac{1}{P(A_i)}$$

- e.g., $I(A_1) = 1$, $I(A_2) = 2$, $I(A_3) = 3$, $I(A_4) = 3$

Properties of Information Content

- Probability and information content are inversely related

Properties of Information Content

- Probability and information content are inversely related
- Use of log means information content is additive

Properties of Information Content

- Probability and information content are inversely related
- Use of log means information content is additive

$$
\begin{aligned}
I(A \cap B) &= \log_2 \frac{1}{P(A \cap B)} \\
&= \log_2 \frac{1}{P(A)\, P(B \mid A)} = I(A) + I(B \mid A)
\end{aligned}
$$

Units of Information Content

- For any event A_i, the corresponding sequence of yes/no answers is a unique bit string that identifies A_i

Units of Information Content

- For any event A_i, the corresponding sequence of yes/no answers is a unique bit string that identifies A_i
- e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$

Units of Information Content

- For any event A_i, the corresponding sequence of yes/no answers is a unique bit string that identifies A_i
- e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$
- More probable events are identified via fewer equiprobable yes/no questions (i.e., shorter bit strings) while less probable events require more questions (i.e., longer bit strings)

Units of Information Content

- For any event A_i, the corresponding sequence of yes/no answers is a unique bit string that identifies A_i
- e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$
- More probable events are identified via fewer equiprobable yes/no questions (i.e., shorter bit strings) while less probable events require more questions (i.e., longer bit strings)
- Information content is therefore measured in bits

Examples of Information Content

- e.g., $P(A_1) = P(A_2) = 1\,/\,2$
- e.g., $P(A_1) = P(A_2) = P(A_3) = P(A_4) = 1\,/\,4$
- e.g., $P(A_1) = 1\,/\,2$, $P(A_2) = 3\,/\,8$, $P(A_3) = P(A_4) = 1\,/\,16$
- e.g., $P(A_1) = 1$, $P(A_2) = P(A_3) = P(A_4) = 0$

Entropy

- Entropy: average information content of a set of n disjoint, mutually exclusive events $A_1, \ldots, A_n$ that partition Ω

$$H(A_1, \ldots, A_n) = \sum_{i=1}^{n} P(A_i) \log_2 \frac{1}{P(A_i)}$$

Entropy

- Entropy: average information content of a set of n disjoint, mutually exclusive events $A_1, \ldots, A_n$ that partition Ω

$$H(A_1, \ldots, A_n) = \sum_{i=1}^{n} P(A_i) \log_2 \frac{1}{P(A_i)}$$

- Average number of equiprobable yes/no questions required to uniquely determine that outcome x is in any event A_i

Entropy

- Entropy: average information content of a set of n disjoint, mutually exclusive events $A_1, \ldots, A_n$ that partition Ω

$$H(A_1, \ldots, A_n) = \sum_{i=1}^{n} P(A_i) \log_2 \frac{1}{P(A_i)}$$

- Average number of equiprobable yes/no questions required to uniquely determine that outcome x is in any event A_i
- Measure of uncertainty of the entire set of events

Examples of Entropy

- e.g., $P(A_1) = P(A_2) = 1\,/\,2$
- e.g., $P(A_1) = P(A_2) = P(A_3) = P(A_4) = 1\,/\,4$
- e.g., $P(A_1) = 1\,/\,2$, $P(A_2) = 3\,/\,8$, $P(A_3) = P(A_4) = 1\,/\,16$
- e.g., $P(A_1) = 1$, $P(A_2) = P(A_3) = 0$

Information Rate

- Suppose we have some encoding of events $A_1, \ldots, A_n$ such that each event A_i is represented using $L(A_i)$ bits

Information Rate

- Suppose we have some encoding of events $A_1, \ldots, A_n$ such that each event A_i is represented using $L(A_i)$ bits
- Information rate: average number of bits per event

$$R(A_1, \ldots, A_n) = \sum_{i=1}^{n} P(A_i)\, L(A_i)$$

Information Rate

- Suppose we have some encoding of events $A_1, \ldots, A_n$ such that each event A_i is represented using $L(A_i)$ bits
- Information rate: average number of bits per event

$$R(A_1, \ldots, A_n) = \sum_{i=1}^{n} P(A_i)\, L(A_i)$$

- Entropy $H(A_1, \ldots, A_n)$ is the best achievable (lowest possible) information rate if events must be uniquely encoded

Examples of Information Rate

- e.g., suppose $P(A_1) = 1\,/\,2$, $P(A_2) = 1\,/\,4$, $P(A_3) = P(A_4) = 1\,/\,8$, what is the information rate of $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 10$?

Examples of Information Rate

- e.g., suppose $P(A_1) = 1\,/\,2$, $P(A_2) = 1\,/\,4$, $P(A_3) = P(A_4) = 1\,/\,8$, what is the information rate of $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 10$?
- e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$?

Examples of Information Rate

- e.g., suppose $P(A_1) = 1\,/\,2$, $P(A_2) = 1\,/\,4$, $P(A_3) = P(A_4) = 1\,/\,8$, what is the information rate of $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 10$?
- e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$?
- e.g., $A_1 = 01$, $A_2 = 000$, $A_3 = 001$, $A_4 = 1$?

Representing (Encoding) Events

- Fixed length codes: use same number of bits to encode each event, e.g., $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 00$

Representing (Encoding) Events

- Fixed length codes: use same number of bits to encode each event, e.g., $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 00$
- Optimal for events with equal probabilities

Representing (Encoding) Events

- Fixed length codes: use same number of bits to encode each event, e.g., $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 00$
- Optimal for events with equal probabilities
- Variable length codes: use different number of bits to encode each event, e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$

Representing (Encoding) Events

- Fixed length codes: use same number of bits to encode each event, e.g., $A_1 = 11$, $A_2 = 10$, $A_3 = 01$, $A_4 = 00$
- Optimal for events with equal probabilities

- Variable length codes: use different number of bits to encode each event, e.g., $A_1 = 1$, $A_2 = 01$, $A_3 = 001$, $A_4 = 000$
- Optimal for events with unequal probabilities

Communicating Perfectly

Transmitting Information

- Goal: transmit one of n messages, encoded in binary

Transmitting Information

- Goal: transmit one of n messages, encoded in binary
- Want to make sure the correct message is received even if there are transmission errors (e.g., static, disk failure, ...)

Transmitting Information

- Goal: transmit one of n messages, encoded in binary
- Want to make sure the correct message is received even if there are transmission errors (e.g., static, disk failure, ...)
- Probability of a single bit being flipped is p

Transmitting Information

- Goal: transmit one of n messages, encoded in binary
- Want to make sure the correct message is received even if there are transmission errors (e.g., static, disk failure, ...)

- Probability of a single bit being flipped is p
- Error probability: overall probability of there being an undetected error when using some encoding scheme

[Got to here in class...]

Examples of Error Probability

- e.g., 8 events represented as 000, 001, 010, . . . 111, probability of a single bit flip is $1/10$, what is the error probability?

Encoding with Redundancy

- Can use additional bits when encoding messages to ensure that they are "protected" against errors in transmission

Encoding with Redundancy

- Can use additional bits when encoding messages to ensure that they are "protected" against errors in transmission
- Error detecting codes vs. error correcting codes

Encoding with Redundancy

- Can use additional bits when encoding messages to ensure that they are "protected" against errors in transmission
- Error detecting codes vs. error correcting codes
- Fundamental trade-off: want encoding schemes that minimize both the error probability and the information rate

Error-Detecting Codes: Parity Check

- Append a parity bit to each binary string such that every binary string always contains an even number of ones

Error-Detecting Codes: Parity Check

- Append a parity bit to each binary string such that every binary string always contains an even number of ones
- e.g., 0000, 0011, 0101, . . . , 1100, 1111

Error-Detecting Codes: Parity Check

- Append a parity bit to each binary string such that every binary string always contains an even number of ones
- e.g., 0000, 0011, 0101, . . . , 1100, 1111
- Can detect error if an odd number of bits get flipped

Error-Detecting Codes: Parity Check

- Append a parity bit to each binary string such that every binary string always contains an even number of ones
- e.g., 0000, 0011, 0101, . . . , 1100, 1111
- Can detect error if an odd number of bits get flipped
- Cannot detect error if an even number of bits get flipped

Examples of Parity Check Codes

- e.g., 8 events represented as 0000, 0011, ..., 1111, probability of a single bit flip is $1\,/\,10$, what is the error probability?

For Next Time

- Check the course website: `http://www.cs.umass.edu/~wallach/courses/s12/cmpsci240/`
- Third homework is due on Friday
- IMPORTANT: check you can log into the EdLab in preparation for next week's homework