CS 112: Modeling Uncertainty in Information Systems

Prof. Jenn Wortman Vaughan

April 25, 2012

Lecture 8

Reminders & Announcements

- The midterm review sessions will be held on Friday after all, led by guest lecturer Eunice Chen
- Because of the scheduling confusion, you are welcome to attend either section on Friday (check the class website for the section locations)
- There will be no graded in-class exercise in the Friday sections this week, just review for the exam

Reminders & Announcements

- The course midterm is one week from today in class
- The exam will cover all of Chapters 1 and 2, plus 4.2 (so everything through the middle of today's lecture)
- One double-sided sheet of hand-written notes is allowed
- You may not use any other notes, books, **calculators**, cell phones, etc.
- Best way to prep is to practice problems from the book

Reminders & Announcements

- Homework 2 is due on Monday
- Homework 1 will be returned in section on Friday
- Solutions to homework 1 are available on Piazza

Today

A bit more about joint PMFs

- Finishing up our sample mean example
- Covariance and correlation

Information theory

- Measuring information

Independence

- X and Y are independent if and only if for all k and l, the events $\{X=k\}$ and $\{Y=l\}$ are independent.

- In other words, X and Y are independent if and only if for all k and l,

$$p_{X,Y}(k, l) = p_X(k)\, p_Y(l)$$

- Can be extended to more than two random variables in the natural way

Why is independence useful?

If $X_1, X_2, \ldots, X_n$ are all independent, then

$$E[X_1 X_2 \ldots X_n] = E[X_1]E[X_2]\ldots E[X_n]$$

Why is independence useful?

If $X_1, X_2, \ldots, X_n$ are all independent, then

$$E[X_1 X_2 \ldots X_n] = E[X_1]E[X_2]\ldots E[X_n]$$

(but it is possible for this to hold even if $X_1, X_2, \ldots, X_n$ are not independent… see the example on Piazza)

Why is independence useful?

If $X_1, X_2, \ldots, X_n$ are all independent, then

$$\mathrm{var}(X_1 + X_2 + \ldots + X_n) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \ldots + \mathrm{var}(X_n)$$

The Sample Mean

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

How accurate is our estimate as a function of n?

The Sample Mean

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

How accurate is our estimate as a function of n?

We will examine this problem in more detail later in the course…

Covariance and Correlation

Covariance

Covariance is a notion that measures the strength and direction of the relationship between two variables.

Covariance

Covariance is a notion that measures the strength and direction of the relationship between two variables.

Experiment: Choose a random UCLA student.

- Let X be his height and Y be his weight. These have positive covariance or are positively correlated.

Covariance

Covariance is a notion that measures the strength and direction of the relationship between two variables.

Experiment: Choose a random UCLA student.

- Let X be his height and Y be his weight. These have positive covariance or are positively correlated.
- Let X be the number of courses he is taking and Y be the average number of hours he sleeps each night. These have negative covariance or are negatively correlated.

Covariance

Covariance is a notion that measures the strength and direction of the relationship between two variables.

Experiment: Choose a random UCLA student.

- Let X be his height and Y be his weight. These have positive covariance or are positively correlated.
- Let X be the number of courses he is taking and Y be the average number of hours he sleeps each night. These have negative covariance or are negatively correlated.
- Let X be his height and Y be the last digit of his UID. These have (roughly) zero covariance or are uncorrelated.

Covariance

Formally, the covariance of two random variables X and Y is defined as

$$\mathrm{cov}(X, Y) = E[\ (X - E[X])\ (Y - E[Y])\ ]$$

Covariance

Formally, the covariance of two random variables X and Y is defined as

$$\begin{aligned}\mathrm{cov}(X, Y) &= E[\,(X - E[X])\,(Y - E[Y])\,] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Covariance

Formally, the covariance of two random variables X and Y is defined as

$$\mathrm{cov}(X, Y) = E[\,(X - E[X])\,(Y - E[Y])\,]$$
$$= E[XY] - E[X]E[Y]$$

Convince yourself at home that these are equal!!

Covariance

Formally, the covariance of two random variables X and Y is defined as

$$\mathrm{cov}(X, Y) = E[\ (X - E[X])\ (Y - E[Y])\ ]$$
$$= E[XY] - E[X]E[Y]$$

- What is cov(X, X)?

Covariance

Formally, the covariance of two random variables X and Y is defined as

$$
\begin{aligned}
\mathrm{cov}(X, Y) &= E[\,(X - E[X])\,(Y - E[Y])\,] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

- What is cov(X, X)?
- Suppose X and Y are independent. What is cov(X, Y)?

Covariance

Formally, the covariance of two random variables X and Y is defined as

$$\begin{aligned} \mathrm{cov}(X, Y) &= E[\ (X - E[X])\ (Y - E[Y])\ ] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- What is cov(X, X)?
- Suppose X and Y are independent. What is cov(X, Y)?
- Suppose cov(X, Y) = 0. Are X and Y independent?

Covariance

- Let X be a random variable taking on values 1 and -1 with equal probability, and let $Y = -X$. What is $\mathrm{cov}(X,Y)$?

Covariance

- Let X be a random variable taking on values 1 and -1 with equal probability, and let Y = -X. What is cov(X,Y)?

- Let X be a random variable taking on values 100 and -100 with equal probability, and let Y = -X. What is cov(X,Y)?

Covariance

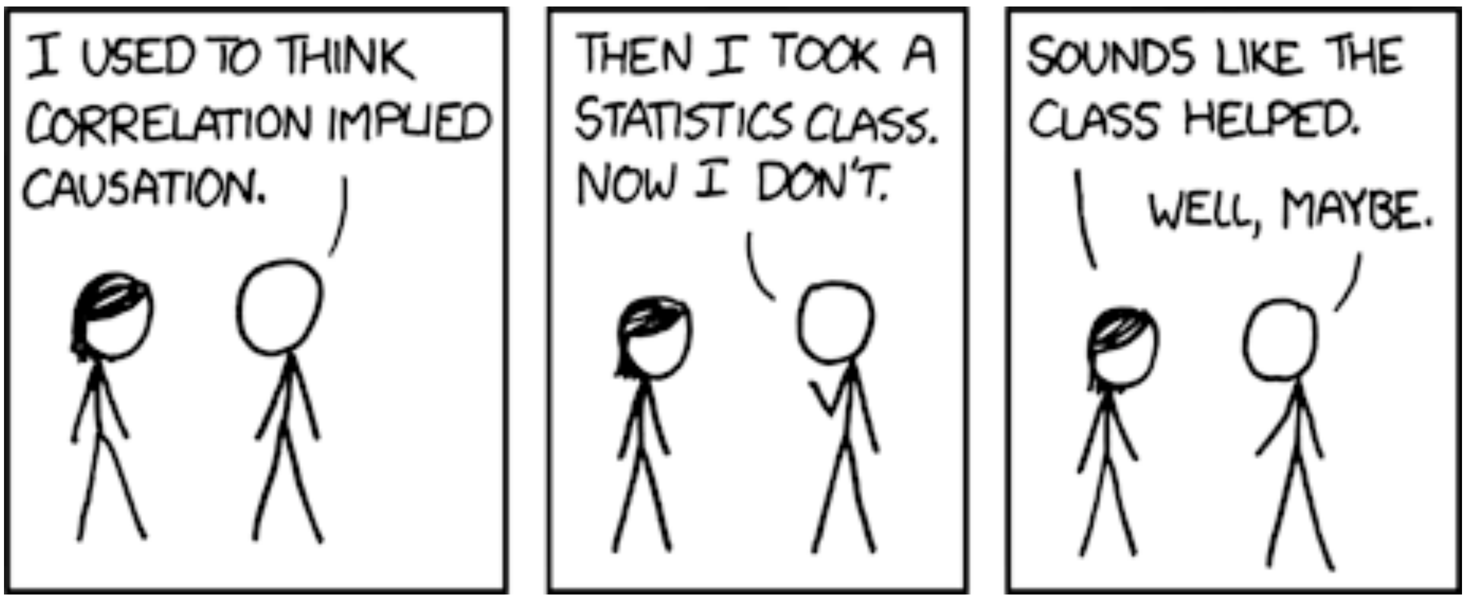
- Let X be a random variable taking on values 1 and -1 with equal probability, and let Y = -X. What is cov(X,Y)?

- Let X be a random variable taking on values 100 and -100 with equal probability, and let Y = -X. What is cov(X,Y)?

- The correlation coefficient is a normalized version of covariance that is always between -1 and 1:

$$\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

Correlation does not imply causation



Correlation does not imply causation

X = number of ice cream cones sold in a day

Y = number of deaths from drowning in a day

X and Y are correlated… therefore ice cream causes drowning? Or drowning causes increased ice cream sales?

Correlation does not imply causation

X = number of ice cream cones sold in a day

Y = number of deaths from drowning in a day

X and Y are correlated… therefore ice cream causes drowning? Or drowning causes increased ice cream sales?

More likely explanation: Both X and Y tend to be higher on summer days.

Correlation does not imply causation

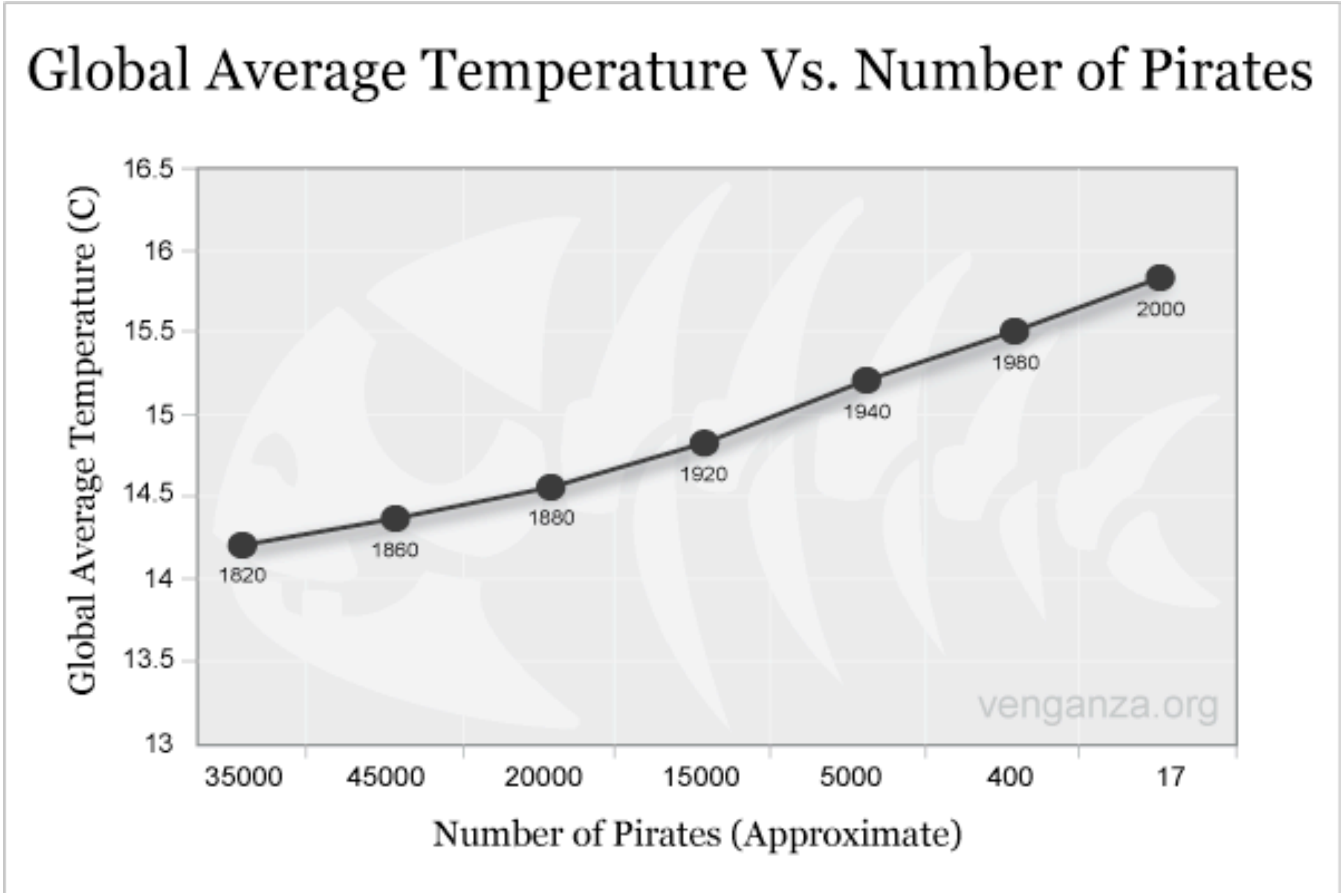
- X might cause Y
- Y might cause X
- Z might cause X and Y
- Some combination of these

Correlation does not imply causation

- X might cause Y
- Y might cause X
- Z might cause X and Y
- Some combination of these

- Or maybe the observed “relationship” is a coincidence or very complex and indirect

Correlation does not imply causation



Information Theory

Information Theory

- Information theory is the math behind communication, i.e., how we represent and transmit information

Information Theory

- Information theory is the math behind communication, i.e., how we represent and transmit information

- Uses of information theory:
  - Measuring information content
  - Compressing data (text, videos, images, etc.)
  - Communicating without errors over noisy communication channels (phone lines, disk drives, etc.)

Information Theory

- Information theory is the math behind communication, i.e., how we represent and transmit information

- Uses of information theory:
  - Measuring information content
  - Compressing data (text, videos, images, etc.)
  - Communicating without errors over noisy communication channels (phone lines, disk drives, etc.)

- We'll start by discussing how to measure information content, and then see how those ideas apply to compression

Probability & Information

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

Probability & Information

- Probability and information are inversely related

- Less probable events have greater information content
    - *The sun came up this morning.*

Probability & Information

- Probability and information are inversely related

- Less probable events have greater information content
  - *The sun came up this morning.*
  - *There will be a problem about the total expectation theorem on the midterm.*

Probability & Information

- Probability and information are inversely related
- Less probable events have greater information content
  - *The sun came up this morning.*
  - *There will be a problem about the total expectation theorem on the midterm.*
  - *The answer to problem 2 is 1024.*

Desirable Properties of Information

- Intuitively, if P(A) = 1, the information gained from learning A should be zero, while the information gained from learning A^c should be infinite

Desirable Properties of Information

- Intuitively, if $P(A) = 1$, the information gained from learning A should be zero, while the information gained from learning A^c should be infinite

- If A and B are independent, then we would like it to be the case that $I(A \cap B) = I(A) + I(B)$

Desirable Properties of Information

- Intuitively, if P(A) = 1, the information gained from learning A should be zero, while the information gained from learning A^c should be infinite

- If A and B are independent, then we would like it to be the case that $I(A \cap B) = I(A) + I(B)$

- These criteria can be satisfied by defining

$$I(A) = \log_2\left(\frac{1}{P(A)}\right)$$

Information Content

- Suppose we flip a fair coin. Let A be the event that the coin is heads. What is I(A)? What is $I(A^c)$?

Information Content

- Suppose we flip a fair coin. Let A be the event that the coin is heads. What is I(A)? What is $I(A^c)$?

- Suppose we generate a sequence of k bits uniformly at random. What is the information content of the particular sequence that we generate?

Information Content

- Suppose we flip a fair coin. Let A be the event that the coin is heads. What is I(A)? What is I(A^c)?

- Suppose we generate a sequence of *k* bits uniformly at random. What is the information content of the particular sequence that we generate?

- Based on this intuition, a unit of information is often referred to as one bit

Asking Informative Questions

Consider four mutually exclusive and exhaustive events, called A, B, C, and D. Suppose I know which one of these events contains the true outcome.

How many yes/no questions would you have to ask me to identify that event?

Asking Informative Questions

Consider four mutually exclusive and exhaustive events, called A, B, C, and D. Suppose I know which one of these events contains the true outcome.

How many yes/no questions would you have to ask me to identify that event?

Correction: There was a small typo in what was written on the board here during class! The question at the root node should have asked if the outcome is in $A \cup B$*, not* $A \cap B$*.*

Asking Informative Questions

Suppose $P(A) = 1/2$, $P(B) = 1/4$, and $P(C) = P(D) = 1/8$.

How many yes/no questions would you have to ask me to identify the event containing the outcome on expectation?

Asking Informative Questions

Suppose $P(A) = 1/2$, $P(B) = 1/4$, and $P(C) = P(D) = 1/8$.

How many yes/no questions would you have to ask me to identify the event containing the outcome on expectation?

Information of an event ≈ number of questions that we need to ask to determine that the outcome is in that event

Entropy

- The average information content of a set of events $A_1,\ldots,A_n$ that partition Ω is called the entropy

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

- Entropy ≈ number of questions that we need to ask to determine on expectation