Introduction to Probability

CSU, Chico Math 314

2018-09-05

outline

outline

Probability of Events

```
url <- "https://roualdes.us/data/email.csv"

email <- read.csv(url)
```

Let's check out `spam` and `format`.

Probability of Events

Here are some columns from the dataset `email`

	spam	num_char	format	number
47	0	8.72	0	small
24	0	0.20	0	small
30	0	2.21	1	small
41	0	12.27	0	small
33	0	10.61	0	small
25	0	4.94	1	none

Table 1: Four rows of data from the email data set

Probability of Events

What is the probability a randomly chosen email is marked as spam, $P(spam)$?

```
mean(email[,"spam"])

## [1] 0.09359857
```

Probability of Events

What is the probability a randomly chosen email is written in HTML, $P(\mathit{format})$?

Probability of Events

What is the probability a randomly chosen email is written in HTML **or** it's spam, $P(?)$?

Probability of Events

What is the probability a randomly chosen email is not written in HTML **and** it's not spam, $P(?)$?

outline

Motivating Conditional Probability

Wouldn't it be great if there was a formal method to update probabilities given new information. Consider this scenario. You draw a card from a deck of 52 cards and then learn that the card you drew is a face card. What is the probability that this card is a king?

- There's 12 face cards.
- 4 of the 12 face cards are kings.
- Therefore, there's a $1/3$ probability that the card you drew is a king.

Motivating Conditional Probability

Without the learned knowledge that the card you drew is a face card, the probability of drawing a king is $4/52 = 0.08$. But after incorporating this new knowledge, the probability of drawing a king is $1/3 = 0.33$.

Conditional Probability

Conditional probability allows us to update probabilities given a new set of information; e.g. we learned that the card we drew was a face card.

conditional probability

For any two events A and B

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

so long as $P(A) > 0$.

Conditional Probability, example

Let A be the probability that we draw a face card and let B be the probability that we draw a king.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(\text{face card and king})}{P(\text{face card})} = \frac{4/52}{12/52}.$$

Conditional Probability, online interactive explanation

Visual Explanations Blog: conditional probability

Conditional Probability, *spam* and *format*

Find $P(format|spam)$.

Conditional Probability, *spam* and *format*

Find $P(spam|format)$.

outline

Independence, intuition

The intuition of independence is best seen mathematically via conditional probability,

$$P(B|A) = P(B).$$

That is the probability of B does not depend on A.

Independence

The formal definition of independence is a condition that need be checked to state that two events are independent.

pairwise independence

Two events A and B are said to be **pairwise independent** if

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

Independence, example

Randomly draw a card from a standard 52 card deck. Let A be the event that the card is a five and B be the event that the card is a spade. Are the events A and B independent?
$P(A \cap B) = 1/52$. Next $P(A) = 4/52$ and $P(B) = 13/52$ gives $P(A)P(B) = 1/52$. Therefore $P(A \cap B) = P(A)P(B)$, so the events are independent.

outline

Association Rules, motivation

From the following table, calculate and interpret $P(tea), P(coffee), P(coffee \cap tea), P(coffee|tea)$.

	coffee	coffee^c	
tea	150	50	200
tea^c	650	150	800
	800	200	1000

Table 2: Survey of coffee and tea drinkers[1].

[1] Tan et al. [2005]

Association Rules, notation

To investigate an association rule between two sets, here coffee and tea, we write

$$\{tea\} \to \{coffee\}.$$

Despite the "fancy" words used in *market basket analysis*, I encourage you to think of these concepts in terms of probability, conditional probability, and independence.

Association Rules, definitions

probability (support)

The probability (support) of an association rule $\{B\} \rightarrow \{A\}$ is the proportion of times the rule is present within the dataset, $P(A \cap B)$.

conditional probability (confidence)

The conditonal probability (confidence) of an association rule $\{B\} \rightarrow \{A\}$ is the support of A and B divided by the support of B, $P(A|B)$.

lift

The lift of an association rule $\{B\} \rightarrow \{A\}$ is the conditional probability (of the rule) divided by the probability of A, $P(A|B)/P(A)$.

Association Rules, example

From the Table 2, we have

$$P(tea) = .2$$
$$P(coffee) = .8$$
$$P(coffee \cap tea) = .15$$
$$P(coffee|tea) = .75 \ldots \text{ relatively high, no?}$$

What is the lift of $\{tea\} \rightarrow \{coffee\}$?

Association Rules, justification

Consider two events, A, B and recall independence. What condition need hold for A, B to be independent?

Association Rules, interpretting lift

Lift of the rule $\{B\} \to \{A\}$ is

$$\frac{P(A \cap B)}{P(A)P(B)}.$$

Interpret

- lift$= 1$,
- lift> 1,
- lift< 1.

outline

Take Away

- It is through probability that much of statistics is justified.
- For arbitrary events
  - now $E = \{\text{flipped head}\}$
  - soon $E = \{X > 3\}$
- Association rules are new and pretty popular in machine learning
  - check out `http://michael.hahsler.net/research/arules_RUG_2015/demo/`

outline

references I

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar.
Introduction to data mining, 2005.