Independent vs dependent events

Consider the following two scenarios:

- I draw a card from a standard deck of 52 cards. Then I return the card to the deck, shuffle it, and draw another card. I repeat this until I have drawn 3 cards total.
- I draw a card from a standard deck of 52 cards. Then I *set the card aside* and draw another card. I repeat this until I have drawn 3 cards total.

For each of these scenarios, what is the probability that all the cards I draw are spades?

Conditional probability

The probability of the event "A given B" is

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

Why?

Conditional probability

Once you know that events are dependent, you will often be interested in *conditional probability*. To determine if an event is conditional, try and find an "IF … THEN …" structure to the scenario.

Conditional probability

Once you know that events are dependent, you will often be interested in *conditional probability*. To determine if an event is conditional, try and find an "IF ... THEN ..." structure to the scenario.

From Example 2.5.3:

> *Suppose that 15% of visitors to a web site are from the USA, 35% are from Australia, and 50% from the rest of the world. The probabilities that a visitor from that region purchases something are .01, .05 and .02 respectively. We want to find the probability that a visitor purchases something.*

Conditional probability

Once you know that events are dependent, you will often be interested in *conditional probability*. To determine if an event is conditional, try and find an "IF ... THEN ..." structure to the scenario.

From Example 2.5.3:

> *Suppose that 15% of visitors to a web site are from the USA, 35% are from Australia, and 50% from the rest of the world. The probabilities that a visitor from that region purchases something are .01, .05 and .02 respectively. We want to find the probability that a visitor purchases something.*

Interpreting:

- IF a visitor is from the USA, THEN the probability they make a purchase is 0.01

Conditional probability

Once you know that events are dependent, you will often be interested in *conditional probability*. To determine if an event is conditional, try and find an "IF . . . THEN . . ." structure to the scenario.

From Example 2.5.3:

> *Suppose that 15% of visitors to a web site are from the USA, 35% are from Australia, and 50% from the rest of the world. The probabilities that a visitor from that region purchases something are .01, .05 and .02 respectively. We want to find the probability that a visitor purchases something.*

Interpreting:

- IF a visitor is from the USA, THEN the probability they make a purchase is 0.01
- IF a visitor is from Australia , THEN the probability they make a purchase is 0.05

Conditional probability

Once you know that events are dependent, you will often be interested in *conditional probability*. To determine if an event is conditional, try and find an "IF ... THEN ..." structure to the scenario.

From Example 2.5.3:

> *Suppose that 15% of visitors to a web site are from the USA, 35% are from Australia, and 50% from the rest of the world. The probabilities that a visitor from that region purchases something are .01, .05 and .02 respectively. We want to find the probability that a visitor purchases something.*

Interpreting:

- IF a visitor is from the USA, THEN the probability they make a purchase is 0.01
- IF a visitor is from Australia , THEN the probability they make a purchase is 0.05
- IF a visitor is from somewhere else, THEN the probability they make a purchase is 0.02

Binary classification

Suppose that you have implemented a machine learning model to detect and filter spam on your email server. Historical data shows that 20% of all incoming email to your server is spam. After using test data on your model, you estimate that it correctly identifies 98% of spam emails as spam, but also incorrectly identifies 3% of legitimate emails as spam.

Let S denote the event that an email is actually spam, and let T denote the event that it is identified as spam by the machine learning model.

(a) Based on the information described above, state the probabilities $P(S)$, $P(T \mid S)$ and $P(T \mid S^c)$.

(b) According to the reading materials, the probability $P(T \mid S^c)$ is called the false positive rate. What name is given to the probability $P(T \mid S)$?

(c) Calculate the probability that an email is identified as spam.

(d) Calculate $P(S \mid T)$ and $P(S \mid T^c)$.

(e) Determine $P(S^c \mid T)$ and $P(S^c \mid T^c)$.

Market basket analysis

Consider the following customer purchase transaction data set:

Transaction ID	Items
1	coffee, cake
2	coffee, newspaper, cake
3	newspaper, cake
4	chips, coffee, newspaper, cake
5	chips, coffee, cake
6	coffee, newspaper, cake
7	peanuts, chips, cake
8	peanuts, coffee
9	peanuts, coffee, newspaper
10	peanuts, coffee, newspaper, cake

- Calculate the support, confidence and lift of the association rule $\{\text{coffee}\} \Rightarrow \{\text{newspaper}, \text{cake}\}$.
- Find all frequent item-sets with minimum support 0.4. Then find all rules of the form $\{A, B\} \Rightarrow \{C\}$ that have a minimum support of 0.4 and minimum confidence of 0.7.