Lecture 10

188 200
Discrete Mathematics and Linear Algebra

Pattarawit Polpinit

Department of Computer Engineering
Khon Kaen Uiversity

July 14, 2009

Overview

Topics for today.

- Independence
- Probability Examples

Reference : Section 6.1-6.9

Independence Events

Definition: If A and B are events in a sample space S, then A and B are **independent** iff

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

What we are saying is the outcome of A doesn't depend in any way on the outcome of B, and conversely.

For example, suppose a coin is tossed twice. The first toss could turn H or T, and would not depend on the outcome of the second toss. The second toss could also turn H or T and would not depend on the outcome of the first toss.

Independence Events: Example

Example: Suppose a fair coin is tossed twice. Does knowing that the coin comes up tail on the first toss help you predict whether the coin will be tail on the second toss?

Solution:

- $S = \{HH, HT, TH, TT\}$
- Let $A =$ "Coin was tail on the first toss" $= \{TH, TT\}$
- Let $B =$ "Coin was tail on the second toss" $= \{HT, TT\}$
- $P(B) = 1/2$

$$P(B|A) = P(B \cap A)/P(A) = \left(\frac{1}{4}\right) / \left(\frac{1}{2}\right) = \frac{1}{2}$$

Conclusion: Knowing the outcome of the first toss does not help you guess the outcome of the second toss.

Are Disjoint Events Independent?

Recall: A and B are independent iff $P(A \cap B) = P(A)P(B)$.

So, are disjoint events independent?

It would be natural to think that mutually disjoint events events would be independent, in fact almost the opposite is true: Disjoint events with non-zero probabilities are dependent.

Example: Let A and B be events on S, and suppose $A \cap B = \emptyset$, $P(A) \neq 0$ and $P(B) \neq 0$. Show that $P(A \cap B) \neq P(A) \cdot P(B)$.

Solution:

- Because $A \cap B = \emptyset$, $P(A \cap B) = 0$.
- But $P(A) \cdot P(B) \neq 0$ because $P(A) \neq 0$ and $P(B) \neq 0$.
- Thus $P(A \cap B) \neq P(A) \cdot P(B)$.

Independence Events: Example

Example: Suppose that A is the event that a randomly generated bit string of length four begins with a 1, and B is the event that this bit string contains an even number of 1s. Are A and B independent if all 4-bit strings are equally likely to occur?

Solution:

- By the product rule, $|S| = 2^4 = 16$
- $A = \{1111, 1110, 1101, 1011, 1100, 1010, 1001, 1000\}$
- $B = \{0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111\}$
- So $P(A) = P(B) = 8/16 = 1/2$.
- $P(A) \cdot P(B) = 1/4$
- $A \cap B = \{1111, 1100, 1010, 1001\}$
- $P(A \cap B) = 4/16 = 1/4$
- Since $P(A \cap B) = P(A) \cdot P(B)$, A and B are independent events.

Independence Events: Example 2

Example: Assume that each of the four ways that a family can have two children are equally likely. Are the events E that a family with two children has two boys, and F that a family with two children has at least one boy independent?

Solution:

- $E = \{BB\}$
- $F = \{BB, BG, GB\}$
- $P(E) = 1/4$
- $P(F) = 3/4$
- $P(E) \cdot P(F) = 3/16$
- $E \cap F = \{BB\}$
- $P(E \cap F) = 1/4$
- Since $1/4 \neq 3/16$, E and F, are not independent.

Product Rule to Determine Probability of Combinations of Events.

If probabilities are independent, we can use the product rule to determine the probabilities of combinations of events.

Example: What is the probability of flipping heads 4 times in a row using a fair coin?

Solution:

- $P(H) = 1/2$
- so $P(HHHH) = P(H) \cdot P(H) \cdot P(H) \cdot P(H) = (1/2)^4 = 1/16$ because probabilities of flipping head each time are independent.

Product Rule to Determine Probability of Combinations of Events.

Example: What is the probability of rolling the same number 3 times in a row using an unbiased 6-sided die?

Solution:

- First roll agrees with itself with probability $1/6$.
- Second roll agrees with first with probability $1/6$.
- Third roll agrees with first two with probability $1/6$.
- So probability of rolling the same number 6 times is $(1/6) \cdot (1/6) \cdot (1/6) = 1/36$.

Spam Filter, Revisited

Recall: we were able to construct a simple filter, a single keyword filter, using Bayes' theorem.

Problem:

- What keyword should we use?
- One keyword is not enough.

How can we fix this?

- Choose keywords such that $P(\text{spam}|\text{keyword})$ is very high or very low.
- Filter based on multiple keywords.

Multiple Keywords Filter

Specifically, we want to develop a Bayesian filter for two keywords that tells us $P[A \mid (B_1 \cap B_2)]$

where $A =$ "an email is spam".

$B_1 =$ "an email contains the first questionable keyword".

$B_2 =$ "an email contains the second questionable keyword".

But first, some assumptions:

1. Events B_1 and B_2 are independent.
2. The events $B_1|A$ and $B_2|A$ are independent.
3. $P(A) = P(\overline{A}) = 0.5$

Formula for $P[A \mid (B_1 \cap B_2)]$

Now, lets derive formula for this $P[A \mid (B_1 \cap B_2)]$.

$$P[A \mid (B_1 \cap B_2)]$$

$$= \frac{P[(B_1 \cap B_2) \mid A]P(A)}{P[(B_1 \cap B_2) \mid A]P(A) + P[(B_1 \cap B_2) \mid \overline{A}]P(\overline{A})}$$

$$= \frac{P[(B_1 \cap B_2) \mid A]}{P[(B_1 \cap B_2) \mid A] + P[(B_1 \cap B_2) \mid \overline{A}]}$$

Now use the assumptions that B_1 and B_2, and $B_1|A$ and $B_2|A$ are independent.

$$= \frac{P(B_1|A)P(B_2|A)}{P(B_1|A)P(B_2|A) + P(B_1|\overline{A})P(B_2|\overline{A})}$$

Spam Filter for Two Keywords

Example: Suppose that we train a Bayesian spam filter on a set of 2000 spam emails and 1000 emails that are not spam. The word "viagra" appears in 400 spam emails and 60 good emails, and the word "discount" appears in 200 spam emails and 25 good emails. Estimate the probability that a message containing the words viagra" and "discount" is spam. Will we reject this message if our spam threshold is set at 0.9?

Step 1: Set up events

- A = email is spam, $\overline{A}$ = email is good
- B_1 = email contains the word "viagra"
- B_2 = message contains the word "discount"

Step 2: Identify probabilities

- $P(B_1|A) = 400/2000 = 0.2$
- $P(B_1|\overline{A}) = 60/1000 = 0.06$
- $p(B_2|A) = 200/2000 = 0.1$
- $p(B_2|\overline{A}) = 25/1000 = 0.025$

Spam Filter for Two Keywords, cont.

Recall:
$P(B_1|A) = P(\overline{A}) = 0.2$
$P(B_1|\overline{A}) = 0.06$
$P(B_2|A) = 0.1$
$P(B_2|\overline{A}) = 0.025$

$$P[A|(B_1 \cap B_2)] = \frac{P(B_1|A)P(B_2|A)}{P(B_1|A)P(B_2|A) + P(B_1|\overline{A})P(B_2|\overline{A})}$$

Compute $P[A \mid (B_1 \cap B_2)$

$$P[A \mid (B_1 \cap B_2)] = \frac{0.2(0.1)}{0.2(0.1) + 0.06(0.025)} = 0.9302$$

Conclusion: Since the probability that our email is spam given that it contains the string "viagra" and "discount" is approximately $0.9302 > 0.9$, we will flag this email as spam.

Bayesian Spam Filter, conclusion

What about formula for more than two keywords?

You probably can guess. For n keywords:

$P[A|(B_1 \cap B_2 \cap \ldots \cap B_n)]$

$$= \frac{P(B_1|A)P(B_2|A)\ldots P(B_n|A)}{P(B_1|A)P(B_2|A)\ldots P(B_n|A) + P(B_1|\overline{A})P(B_2|\overline{A})\ldots P(B_n|\overline{A})}$$

Advantage: it can be trained on a per-user basis.

• A scientist who is researching on Viagra won't have emails containing the word "Viagra" flagged as spam, because "Viagra" will show up often in his good emails.

Disadvantage:

• Assume that keywords are independent.

• Can't filter image.