Foundations of Programming

Probabilities and Naïve Bayes models

Announcements

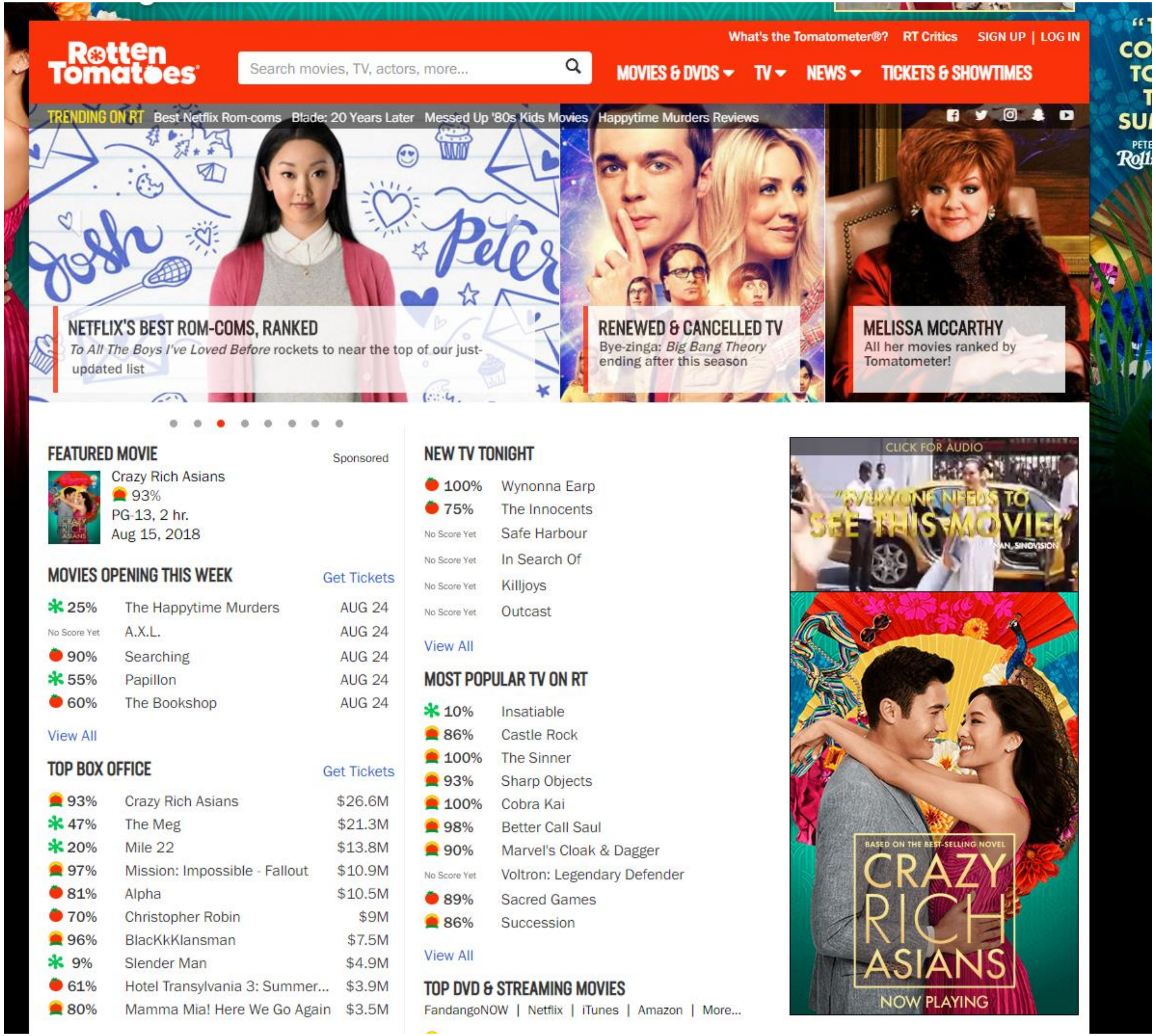
- Depth & Breadth
  - Depth: Data Mutation, nested loops, if-statements and pictures
  - Breadth: ??
- Diversity in Computing lunch form: https://goo.gl/forms/OqIkiuhmOQbWTQWs2

Learning outcomes/key ideas

- Basics of probability
- Representing Joint Probability
- Inference by enumeration
- Bayes Rule
- Naïve Bayes for sentiment classification

Text sentiment classification

- 93%
- 47%
- 20%
- 97%
- 81%
- 70%
- 96%



Basic Probability Theory

- An **experiment** has a set of potential outcomes, e.g., throw a dice
- The **sample space** of an experiment is the set of all possible outcomes, e.g., {1, 2, 3, 4, 5, 6}
  - A **random variable** can take on any value in the sample space
- An **event** is a subset of the sample space.
  - {2}
  - {3, 6}
  - even = {2, 4, 6}
  - odd = {1, 3, 5}

Probability as Relative Frequency

Total Flips: 10
Number Heads: 5
Number Tails: 5

Probability of Heads:
Number Heads / Total Flips = 0.5

Probability of Tails:
Number Heads / Total Flips = 0.5 = 1.0 – Probability of Heads

The experiments, the sample space and the events must be defined clearly for probability to be meaningful

Probabilities and classification

We are trying to determine that the probability that a given movie review is positive vs. negative. We will select the classification that has a higher probability.

One quantity we care about in this task is the ***prior probability*** of the sentiment of a review. In other words:

$P(Sentiment = pos)$ and $P(Sentiment = neg)$

True or false: $P(Sentiment = pos) + P(Sentiment = neg) = 1$

A. True B. False C. It depends

Prior probability

Of 100 students completing a course, 20 were business major. Ten students received As in the course, and three of these were business majors.

What is the probability that a randomly selected student got an A?
P(A = True) = ?

A. 0.03 B. 0.1 C. 0.3 D. 0.2 E. None of these

Joint probability: Two or more events BOTH happening

Of 100 students completing a course, 20 were business major. Ten students received As in the course, and three of these were business majors.

What is the probability that a randomly selected student is a Business student and gets an A?
P(A = True, B = True) = ?

A. 0.03 B. 0.1 C. 0.3 D. 0.2 E. None of these

"the probability of A and B"

	B(usiness student) = True	B(usiness student) = False
A = True		
A = False		

P(A=True) = P(A=True, B=True) + P (A=True, B=False)

marginalization

Will the values in a joint probability table always sum to 1?
A. Yes B. No

But how does evidence change things? *Conditional* probability

Of 100 students completing a course, 20 were business major. Ten students received As in the course, and three of these were business majors What is the probability of A after knowing B is true?

What is the probability that a randomly selected *business student* gets an A?
P(A = True | B = True) = ? A. 0.03 B. 0.1 C. 0.3 D. 0.2 E. None of these

"the probability of A ***given*** *B"*

	B(usiness student) = True	B(usiness student) = False
A = True	0.03	0.07
A = False	0.17	0.73

P(A=True | B = True) = P(A=True, B=True) / P(B=True)

Inference

- Often we will be able to measure some information, but we want to make statements about things that are not directly in our table of information.
- E.g., we want to know what is the probability of a toothache in general (the prior probability), never mind whether there's a cavity, catch, etc.

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Inference = Using known facts to derive others

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

Inference by enumeration

- Start with the joint probability distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	.108	.012	.072	.008
$\neg$ *cavity*	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.108+0.012}{0.108 + 0.012 + 0.016 + 0.064}$$

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

These terms are a pain to compute

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(\mathit{cavity} \mid \mathit{toothache}) = \frac{0.12}{0.108 + 0.012 + 0.016 + 0.064}$$

$$P(\neg \mathit{cavity} \mid \mathit{toothache}) = \frac{0.08}{0.108 + 0.012 + 0.016 + 0.064}$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

P(*cavity* | *toothache*) = α 0.12

P(¬*cavity* | *toothache*) = α 0.08

$$\alpha = \frac{1}{0.108 + 0.012 + 0.016 + 0.064}$$

P(¬*cavity* | *toothache*) + P(*cavity* | *toothache*) = α 0.08 + α 0.12 = ??

What value is ??

A. 1
B. α
C. 0.108 + 0.012 + 0.016 + 0.064
D. Something else
E. I have no idea!

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

P(*cavity* | *toothache*) + P(¬ *cavity* | *toothache*) = α 0.12 + α 0.08 = 1

$\boxed{\alpha = 1 / (0.12 + 0.08)}$

α = 1 / 0.2 = 5

<P(*cavity* | *toothache*), P(¬ *cavity* | *toothache*)> = < α 0.12, α 0.08 > = <0.6, 0.4>

Compare to computing: $\alpha = \dfrac{1}{0.108 + 0.012 + 0.016 + 0.064}$

Bayes Theorem

If $P(E2)>0$, then

$P(E1|E2) = P(E2|E1)P(E1) \; / \; P(E2)$

This can be derived from the definition of conditional probability.

Bayes Rule example

A patient takes a lab test and the result comes back positive. The test has a false negative rate of 2% and false positive rate of 3%. Furthermore, 0.8% of the entire population have this cancer.

What is the probability of cancer if we know the test result is positive?

"False negative" = Considering only situations where there is cancer, test is negative
"False positive" = Considering only situations where there is not cancer, test is positive

Bayes Rule example

A patient takes a lab test and the result comes back positive. The test has a false negative rate of 2% and false positive rate of 3%. Furthermore, 0.8% of the entire population have this cancer.

What is the probability of cancer if we know the test result is positive?

<u>We know:</u>

<u>We want:</u>

"False negative" = Considering only situations where there is cancer, test is negative
"False positive" = Considering only situations where there is not cancer, test is positive

Bayes Rule example

A patient takes a lab test and the result comes back positive. The test has a false negative rate of 2% and false positive rate of 3%. Furthermore, 0.8% of the entire population have this cancer.

What is the probability of cancer if we know the test result is positive?

We know:

P(Test=pos|¬cancer) = 0.03 FP

P(Test=neg|cancer) = 0.02 FN

P(cancer) = 0.008 Prior

We want:

P(cancer|Test=pos)

Bayes Rule example

A patient takes a lab test and the result comes back positive. The test has a false negative rate of 2% and false positive rate of 3%. Furthermore, 0.8% of the entire population have this cancer.

What is the probability of cancer if we know the test result is positive?

We know:

$P(\text{Test=pos}|\neg\text{cancer}) = 0.03$ FP

$P(\text{Test=neg}|\neg\text{cancer}) = 0.97$

$P(\text{Test=pos}|\text{cancer}) = 0.98$

$P(\text{Test=neg}|\text{cancer}) = 0.02$ FN

$P(\text{cancer}) = 0.008$ Prior

$P(\neg\text{cancer}) = 0.992$

We want:

$P(\text{cancer}|\text{Test=pos}) =$

$$\frac{\overset{0.98}{P(\text{Test=pos}|\text{cancer})}\,\overset{0.008}{P(\text{cancer})}}{\underset{0.0376}{P(\text{Test=pos})}} \quad \boxed{= 0.21}$$

$$P(\text{Test=pos}) = \overset{0.00784}{P(\text{Test=pos, cancer})} + \underset{0.02976}{P(\text{Test=pos}, \neg\text{cancer})}$$

$$P(\text{Test=pos, cancer}) = \overset{0.98}{P(\text{Test=pos}|\text{cancer})}\,\overset{0.008}{P(\text{cancer})}$$

$$P(\text{Test=pos}, \neg\text{cancer}) = \overset{0.03}{P(\text{Test=pos}|\neg\text{cancer})}\,\overset{0.992}{P(\neg\text{cancer})}$$

Returning to review sentiment classification…

Our evidence is the text in the review. We want to estimate

$$P(Sentiment = pos \mid text)$$

But we could never directly estimate this because we're unlikely to have ever seen this specific text before! How can Bayes rule help us?

$$P(Sentiment = pos \mid text) = \frac{P(text|Sentiment = pos)\, P(Sentiment = pos)}{P(text)}$$

Returning to review sentiment classification...

$$P(Sentiment = pos \mid text) \propto P(text \mid Sentiment = pos)\ P(Sentiment = pos)$$

Simplifying assumption 1: Represent text with a "bag of words" representation

$$P(w_1 = true, w_2 = true, w_3 = false, \ldots \mid Sentiment = pos)$$

But can we learn this??

Independence: Intuition

- Events are independent if one has nothing whatever to do with others. Therefore, for two independent events, knowing one happening does not change the probability of the other event happening.
  - one toss of coin is independent of another coin (assuming it is a regular coin).
  - price of tea in England is independent of the result of general election in Canada.

Independence: Definition

- Events A and B are independent iff:

$$P(A, B) = P(A) \times P(B)$$

which is equivalent to

$$P(A|B) = P(A) \text{ and}$$

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

when $P(A, B) > 0$.

T1: the first toss is a head.

T2: the second toss is a tail.

$$P(T2|T1) = P(T2)$$

Conditional Independence

- Dependent events can become independent given certain other events.
- Example,
  - Size of shoe
  - Size of vocabulary
  - ??
- Two events A, B are conditionally independent given a third event C iff

  $P(A|B, C) = P(A|C)$

Conditional Independence: Utility via Naïve Bayes

- Let E1 and E2 be two events, they are conditionally independent given E iff

  P(E1|E, E2)=P(E1|E),

  that is the probability of E1 is not changed after knowing E2, given E is true.
- Equivalent formulations:

  P(E1, E2|E)=P(E1|E) P(E2|E)

  P(E2|E, E1)=P(E2|E)

$$P(w_1 = true, w_2 = true, w_3 = false, \ldots | Sent = pos)$$
$$= P(w_1 = true | Sent = pos)\, P(w_2 = true | Sent = pos) P(w_3 = false | Sent = pos) \ldots$$

A ha! These we can learn from data!

Returning to review sentiment classification…

$$P(Sentiment = pos \mid text) \propto P(text|Sentiment = pos)\ P(Sentiment = pos)$$

Simplifying assumption 1: Represent text with a "bag of words" representation

$$P(w_1 = true, w_2 = true, w_3 = false, \ldots |Sentiment = pos)$$

Simplifying assumption 2: Words are conditionally independent given sentiment

$$P(w_1 = true, w_2 = true, w_3 = false, \ldots |Sent = pos)$$
$$= P(w_1 = true|Sent = pos)\ P(w_2 = true|Sent = pos)P(w_3 = false|Sent = pos) \ldots$$

Training a Naïve Bayes classifier

$$P(Sentiment = pos \mid text) \propto P(text|Sentiment = pos)\ P(Sentiment = pos)$$

Simplifying assumption 1: Represent text with a "bag of words" representation

$$P(w_1 = true, w_2 = true, w_3 = false, \ldots |Sentiment = pos)$$

Simplifying assumption 2: Words are conditionally independent given sentiment

$$P(w_1 = true, w_2 = true, w_3 = false, \ldots |Sent = pos)$$
$$= P(w_1 = true|Sent = pos)\ P(w_2 = true|Sent = pos) P(w_3 = false|Sent = pos) \ldots$$

Practice with Probability

- Which of the following statements are generally true? (If they are true only in certain conditions, state what the conditions are)

$P(A,B) = P(A)*P(B)$

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

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

$P(A|B) + P(A|\neg B) = 1$

$P(\neg A) + P(A) = 1$

$P(\neg A, B) + P(A, B) = P(B)$

$P(\neg A|B) + P(A|B) = P(B)$