Lecture 4: Probability

USC VSoE CSCI 544: Applied Natural Language Processing

Jonathan May -- 梅約納

September 1, 2017

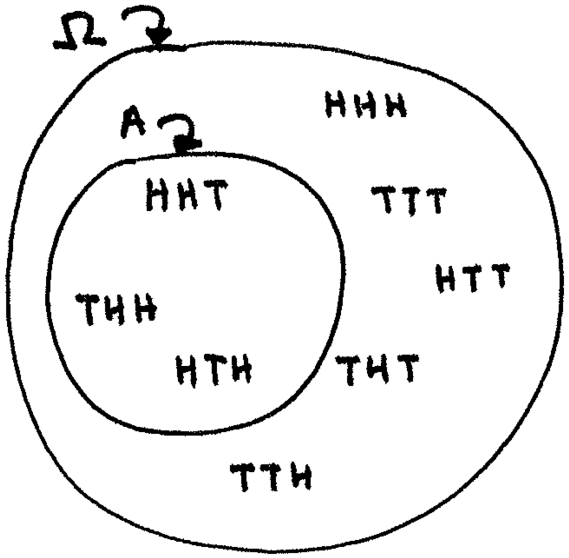
based on Kevin Knight's notes

Definitions

- Experiment: Some action that takes place in the world
- Outcomes = Sample Space = Ω = the universe; every basic outcome that could happen
- Event = $A \subseteq \Omega$ = something that happened (could be more than one basic outcome)
- Probability Distribution: $\boldsymbol{\mathcal{F}}: \Omega \rightarrow \mathbb{R}^{+}_{0}$, $\sum_{x \in \Omega} \boldsymbol{\mathcal{F}}(x)=1$
  i.e. the values sum to 1 and no value is negative

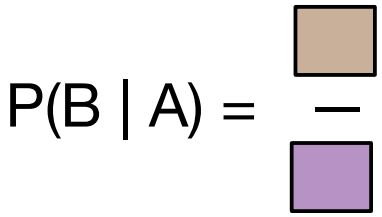
Example

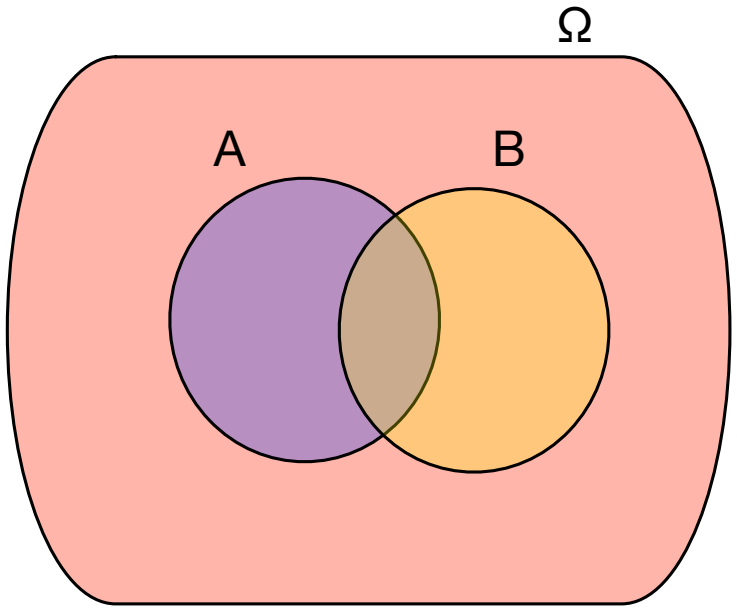
- Experiment: "toss a coin three times"
- Ω = {HHH, HHT, HTT, HTH, THH, THT, TTT, TTH}
- Event A = "exactly two heads" = {HHT, HTH, THH}
- Event B = "first one was heads" = {HHT, HTH, HTT, HHH}
- Distribution: assign a number between 0 and 1 ("probability") to each basic outcome; sum of all numbers = 1
- Uniform distribution: P(x) = 1/8 $\forall x \in \Omega$
- Probability of an event = the sum of the probabilities of its basic outcomes
- So, P(A) = ?
- P(B) = ?

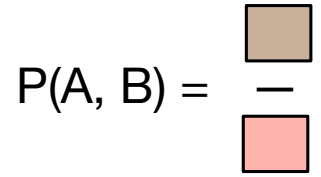


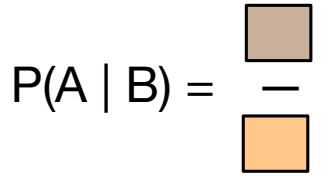
Joint and Conditional Probability

- P(A, B) = P(A∩B)
  = "joint probability of A and B"
- P(A|B) = P(A∩B)/P(B)
  = "conditional probability of A given B"
- P(B|A) = ?



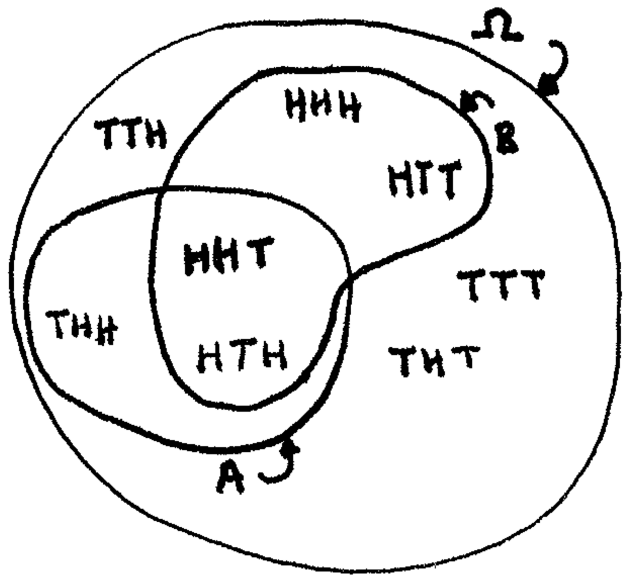




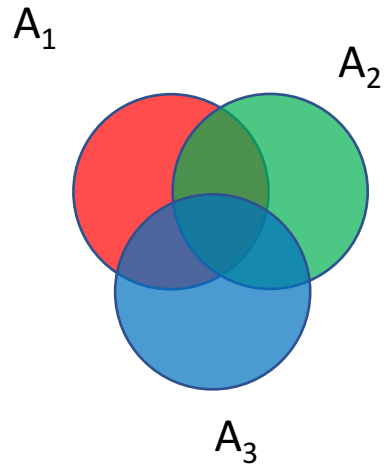


Joint and Conditional Probability

- A = "exactly two heads"
  = {HHT, HTH, THH}
- B = "first one was heads" = {HHT, HTH, HTT, HHH}
- P(A) = 3/8
- P(B) = ½
- P(A,B) = P(A) + P(B) ?
  - No. What is it?
- P(B|A) = "if you got 2 heads, what's the chance the first was heads?" = ?
- P(A|B) = "if your first was heads, what's the chance you got 2 heads" = ?



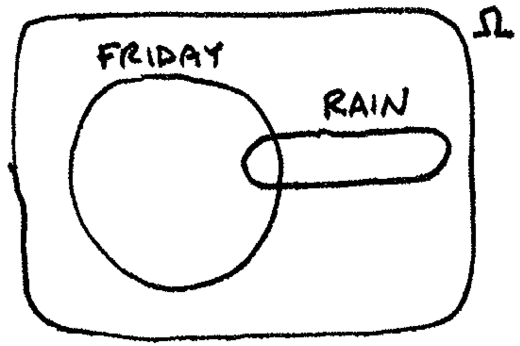
Chain Rule of Probability



- $P(A, B) = P(A \cap B)$
- $P(A|B) = P(A \cap B)/P(B) = P(A, B)/P(B)$
- Thus, $P(A, B) = P(A|B)P(B)$
- What if we had three events: A_1, A_2, A_3?
- Joint probability = $P(A_1, A_2, A_3) = P(A_1 \cap A_2 \cap A_3)$
- Define " A_2, A_3 " as "B". Then from above, $P(A_1, B) = P(A_1|B)P(B)$
- Now substitute back in: $P(A_1| A_2, A_3)P(A_2, A_3)$
- Now rewrite $P(A_2, A_3)$ in terms of conditionals:
  $P(A_1| A_2, A_3) P(A_2|A_3)P(A_3)$
- Notice: $P(A_1, A_2, A_3) = P(A_3, A_2, A_1) = P(A_3| A_2, A_1) P(A_2|A_1)P(A_1)$
- In general, $P(A_1...A_n) = P(A_1|A_2...A_n)P(A_2|A_3...A_n)...P(A_{n-1}|A_n)P(A_n)$

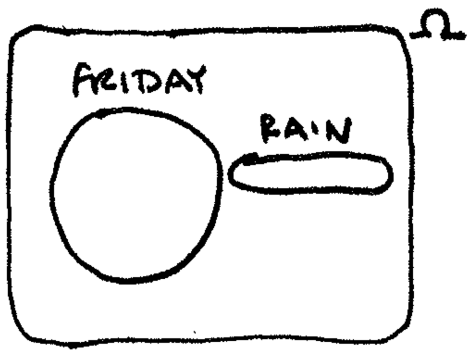
Independence

- Events A and B are <u>independent</u> if P(A) = P(A|B)



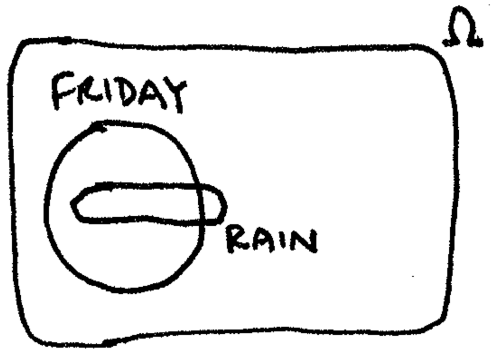
$P(\text{FRIDAY}) = 1/7$
$P(\text{FRIDAY} \mid \text{RAIN}) = 1/7$

independent



$P(\text{FRIDAY}) = 1/7$
$P(\text{FRIDAY} \mid \text{RAIN}) = 0$

dependent



$P(\text{FRIDAY}) = 1/7$
$P(\text{FRIDAY} \mid \text{RAIN}) = 4/5$

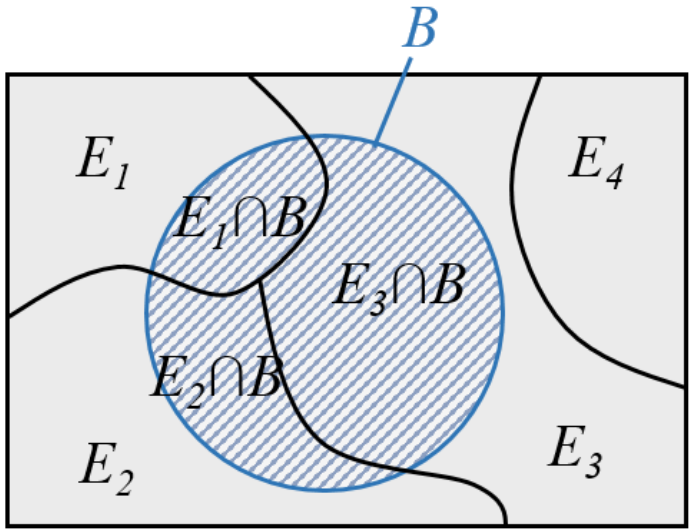
dependent

Bayes' Rule

- $P(A|B) = P(A, B)/P(B)$; $P(A, B) = P(A|B)P(B)$
- $P(B|A) = P(A, B)/P(A)$; $P(A, B) = P(B|A)P(A)$
- $P(A|B)P(B) = P(B|A)P(A)$
- $P(A|B) = P(B|A)P(A)/P(B)$

Law of Total Probability

- Let there be events $E_1 ... E_n$ that <u>partition</u> an outcome space (Ω)
  - $\sum_{i=1}^{n} P(E_i) = 1$
  - $0 \leq P(E_i) \leq 1$
- For any $B \subseteq \Omega$
  - $P(B) = \sum_{i=1}^{n} P(B, E_i)$



Using Bayes' Rule and Law of Total Probability

- Some people can read minds!
  - Not very many of them;
  - P(MR) = 1/100,000
- I have invented a mind reader detector test!
  - If you're a mind reader I can detect this with 95% accuracy : P(T|MR) = 0.95
  - If you're not, I can detect this with 99.5% accuracy: P(¬T|¬MR) = 0.995
- Jill takes the test. The test says "T" (true). How likely is it that Jill is a mind reader?

Jill Passed the Test. Is She a Mind Reader?

- P(MR|T) = ?
- P(MR|T) = P(T|MR)P(MR)/P(T)
- How do we use P(T|MR) and law of total probability to get P(T)?
  - P(T) = P(T,MR)+P(T,¬MR)
  - P(T,MR) = P(T|MR)P(MR) = .95*.00001=.0000095
  - P(T,¬MR)=P(T|¬MR)P(¬MR) = .005*.99999 = .00499995
  - P(T) = .00500945
- P(MR|T) = .95*.00001/.00500945 ≈ 0.002

P(MR)	
MR	.00001
¬MR	.99999

P(T|MR)	T	¬T
MR	.95	.05
¬MR	.005	.995

Argmax and Bayes' Rule

- $\text{argmax}_{Ai} P(A_i) \stackrel{\text{def}}{=}$ "The A_i with highest probability"
  - A_i = {rain, snow, sun}
  - P(rain) = 0.1; P(snow) = 0.1; P(sun) = 0.8;
  - $\text{argmax}_{Ai} P(A_i)$ = ? sun
- Consider argmax in the conditional case
- $\text{argmax}_{Ai}\ P(A_i|B)$
- = $\text{argmax}_{Ai}\ P(A_i)P(B|A_i)/P(B)$
- = $\text{argmax}_{Ai}\ P(A_i)P(B|A_i)$
  - why?

Argmax and Bayes' Rule

- $\text{argmax}_{Ai}\ P(A_i)P(B|A_i)$ may be easier to work with than $\text{argmax}_{Ai}\ P(A_i|B)$
- Let B be a misspelled character sequence ("limf")
- Let $A_1 ... A_n$ be possible corrections ("lime", "limb", "limn")
- $P(A_i|B) = ?$ keyboard distance?
- $P(A_i) \quad * \quad P(B|A_i)$

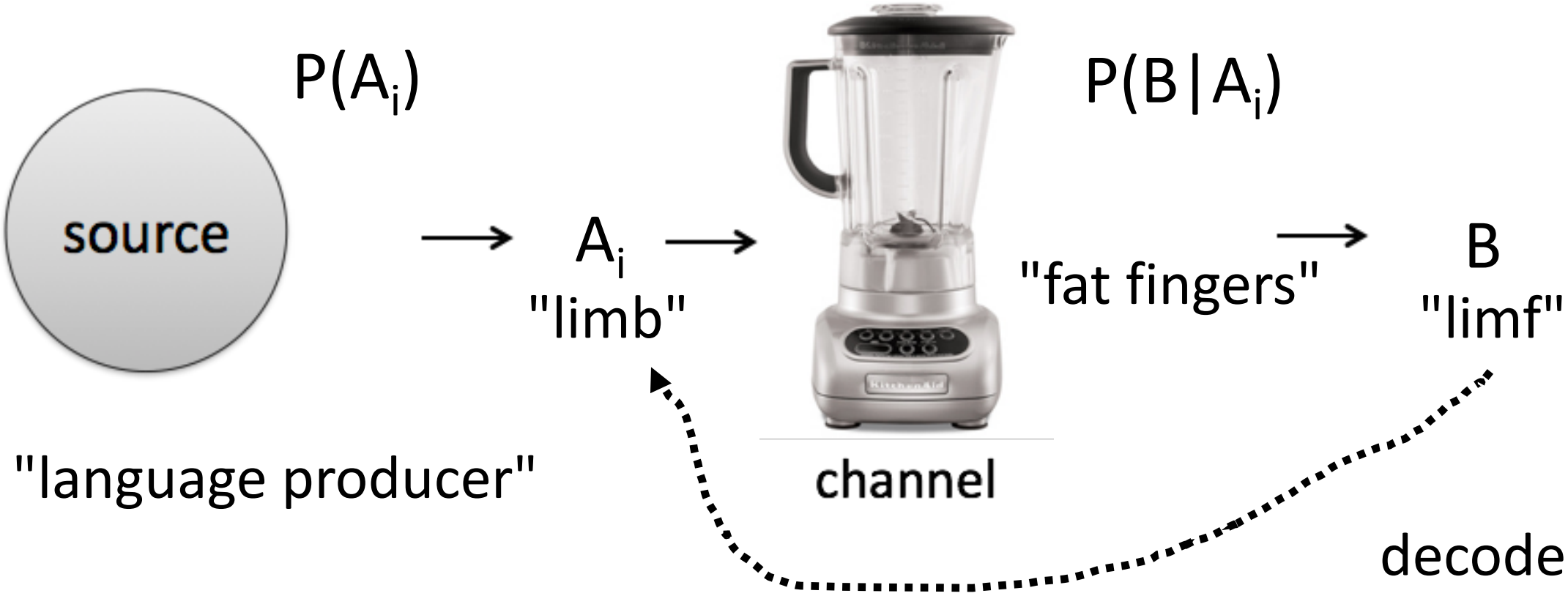
word likelihood

"We must surgically amputate the limf"

- Can bring more knowledge to help!

Bayes' Rule and Noisy Channel Model

- By reformulating in this way we're taking a particular view of how we view events



Quiz Questions

- Experiment: I have two bowls, each with 100 balls numbered 1 through 100. I pick 1 ball at random from each bowl
- 1) What is the size of the sample space (Ω)?
  - A) 200
  - B) 100
  - C) 10,000
  - D) infinite
- 2) What is the set of outcomes for the event "the sum of the balls is 200"?
  - A) {(200)}
  - B) {(100, 100)}
  - C) ∅
  - D) {(4, 50), (50, 4), (10, 20), (20, 10)}
- What is the probability of the event in Q2 (assume uniform distribution)?

Quiz Questions

- 3) Which of these are possible probability distributions (multiple or no selections allowed)?
  - A) (1.3, 2)
  - B) (0.2, 0.2, 0.2, 0.2)
  - C) (0.2, 0.2, 0.2, 0.2, -0.1, 0.3)
  - D) (0.2, 0.2, 0.2, 0.2, 0.2)
  - E) (0, 1, 0)
  - F) (1)
  - G) (1/8, ¼, 5/8)
  - H) (-0.5, -0.5)

Estimating Probabilities

- How do we get these basic outcome probabilities?
  - Do experiments!
  - Find some data (a sample)
  - Count stuff in the sample
- e.g. "how often do sentences start with the word 'So'?
  - Experiment: Observe sentences that occur in the world
  - Basic outcomes: {every possible sentence that could occur}
  - Event A: {sentences that start with 'So'}
  - P(A) = count of sentences starting with 'So' in the sample/sentences in the sample
  - P(A|B) = count(A, B)/count(B)

Data and Model

- Suppose we have a biased 12-sided die and we throw it repeatedly:

result	count
1	0
2	7
3	26
4	45
5	119
6	70

result	count
7	31
8	14
9	6
10	0
11	4
12	2

- It's easy to estimate the 11 free parameters (why 11?)
- Total rolls = 324; P(1) = 0/324 = 0; P(5) = 119/324 = .367
- P(12) = 1-(P(1)+P(2)+...+P(11))

Data and Model

- What if you only saw the results and weren't told how they were generated?
- Another theory: sum of two 6-sided die rolls
  - What about sum of three 4-sided dice?
- In the 2d6 case, P(8) = P(3)P(5) + P(4)P(4) + P(5)P(3)
- Now only 5 free parameters

result	count
1	0
2	7
3	26
4	45
5	119
6	70

result	count
7	31
8	14
9	6
10	0
11	4
12	2

Data and Model

- Maybe it's not dice at all
- Could be number of minutes of phone calls
- Historical data shows this follows a gamma curve
- Shape is fixed but two numbers control center and 'flatness'

result	count
1	0
2	7
3	26
4	45
5	119
6	70

result	count
7	31
8	14
9	6
10	0
11	4
12	2

Data and Model

?

- General theory of how data got generated, together with a set of parameter estimates is called a **model**
- Questions:
  - Given a theory T, how do we know if one set of parameter estimates is better than another?
  - Which is better: theory T with parameters t or theory S with parameters s?
  - i.e. is one theory in general better than another?
  - can we automatically identify the best model (best theory+param combination)?

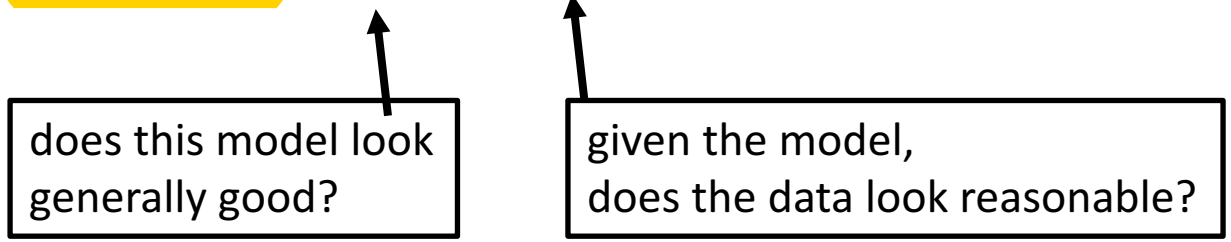
result	count
1	0
2	7
3	26
4	45
5	119
6	70

result	count
7	31
8	14
9	6
10	0
11	4
12	2

BIG MACHINE LEARNING QUESTIONS!

Data and Model

- Ask the data!
- For data D and models m in M, find $\text{argmax}_m P(m|D)$
- By Bayes' rule this is equal to $\text{argmax}_m P(m)\, P(D|m)$

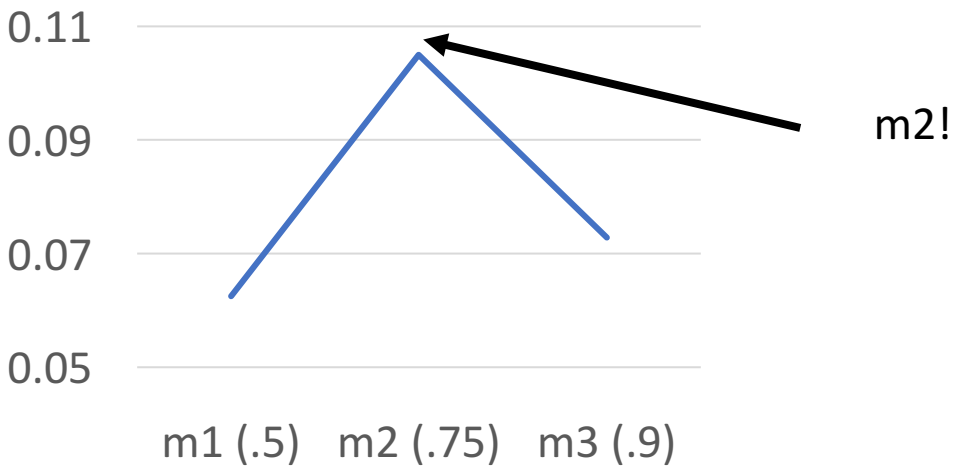


- If we don't have any prior preferences, i.e. distribution over M is uniform, then it's just $\text{argmax}_m P(D|m)$

Determining P(D|m)

- Quantitative!
  - Get some D (a sample)
  - apply each m
- $P(D|m) = \prod_{d \in D} P(d|m)$
- m1: ½ * ½ * ½ * ½ = .0625
- m2: ¾ * ¾ * ¼ * ¾ = .105
- m3: 9/10 * 9/10 * 1/10 * 9/10 = .073
- $\text{argmax}_m P(D|m)$ = m2
- This is the maximum likelihood estimation (MLE)

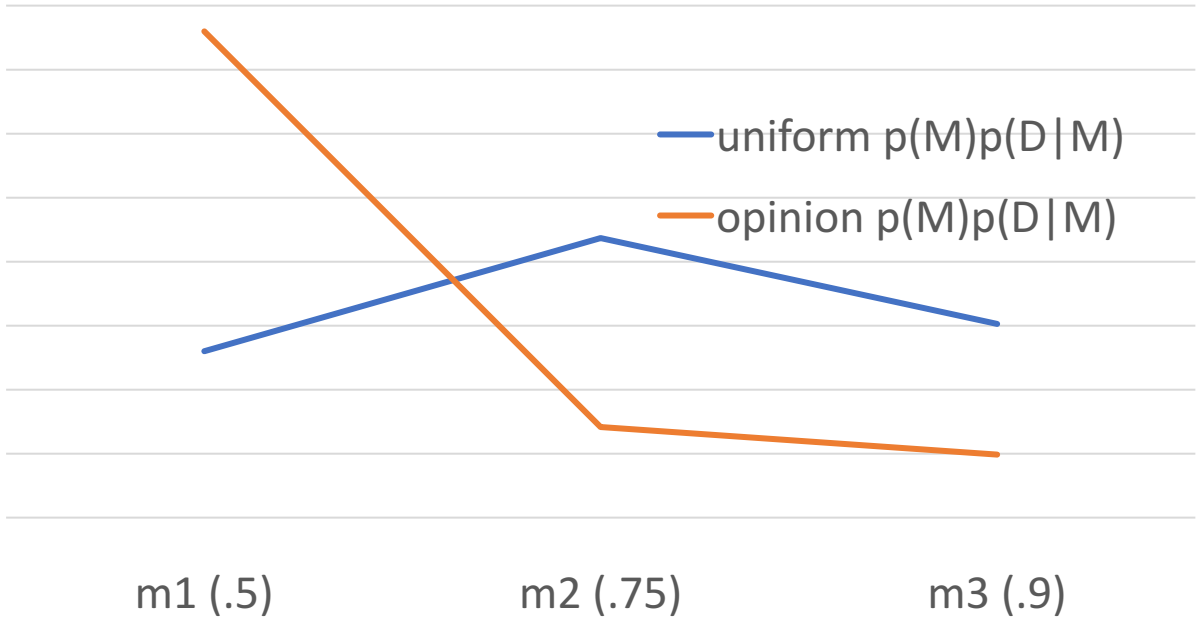
Model	P(H)	P(T)
m1	.5	.5
m2	.75	.25
m3	.9	.1



What if we have an opinion about M?

- e.g. the coin doesn't look biased

M	P(M)
m1	.9
m2	.1
m3	.1



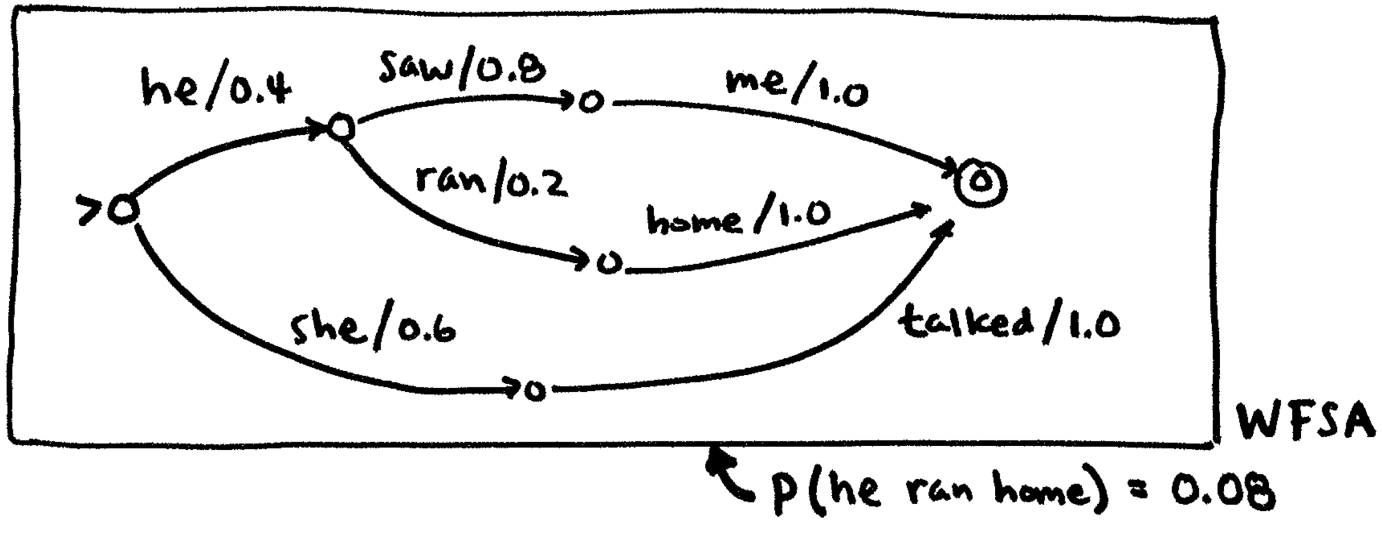
We could also compare the d12 vs 2d6 theories using the same technique!

Quiz Questions

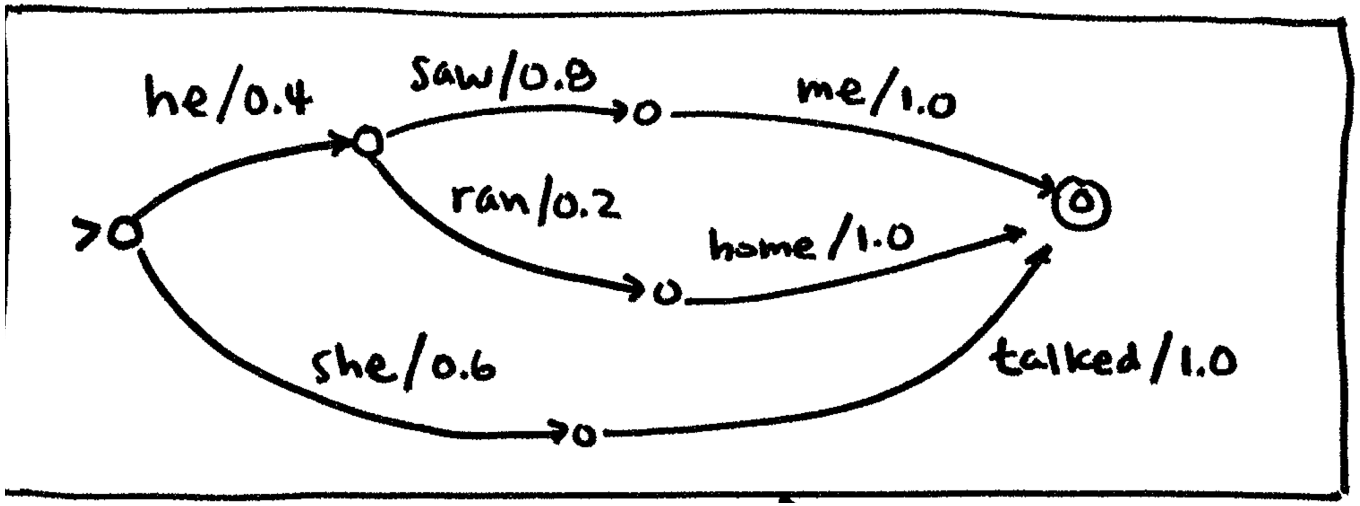
- $P(A) = ½$; $P(A \cap B) = ¼$; $P(B) = 3/10$
- 4) What is $P(A|B)$?
  - A) 5/6
  - B) 1
  - C) ¼
  - D) ½
- 5) What is $P(B|A)$? (same choices)
- 6) What is $P(A, B)$? (same choices)

Adding Probabilities to FSA

- Aside from a label, FSAs can have a **weight**; then they're **weighted finite-state automata** (wFSA)
- If the weights of arcs leaving a path sum to 1 (final state excluded) it's **probabilistic** (pFSA)
- Weight of a path = product of the weights of the arcs



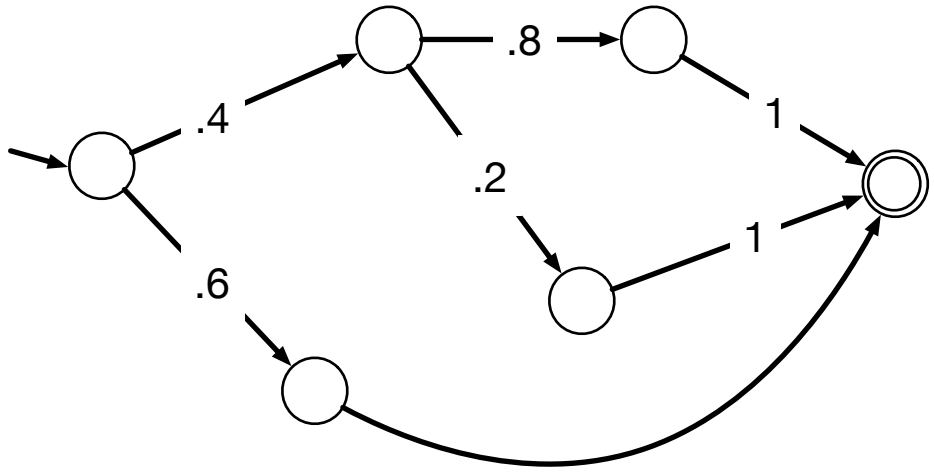
pFSA is a way to encode P(A)



"she talked" = .6
"he saw me" = .32
"he ran home" = .08

wFSA is just a weighted graph

- We can use best-path algorithms (Dijkstra) to efficiently find the most probable path (after converting to log space)
- We can use FSA intersection to efficiently look up the probability of a sentence



Adding Probabilities to FST

- FSTs can also have weighted arcs
- If the weights of all outgoing arcs from a state **with the same input symbol** sum to 1, it's a pFST
- FST transduces string x into string y
- pFST transduces x into y with probability P(y|x)
  - Note: probabilities no longer reversable!