Probability and sequences: from coin tosses to DNA

Rahul Siddharthan

The Institute of Mathematical Sciences, Chennai 600113

FACETS, 6 July 2018

rsidd@imsc.res.in

Probability

What is “probability”? For example, what do the following mean:

Probability

What is “probability”? For example, what do the following mean:

- The probability of getting “heads” in a coin toss is 0.5

Probability

What is "probability"? For example, what do the following mean:

- The probability of getting "heads" in a coin toss is 0.5
- There is a 30% probability of rain today

Probability

What is "probability"? For example, what do the following mean:

- The probability of getting "heads" in a coin toss is 0.5
- There is a 30% probability of rain today
- There is a 1.7% probability of Trump winning the US presidential election (early Oct 2016)

FORECAST

PRESIDENT SENATE

By Natalie Jackson and Adam Hooper
Additional design by Alissa Scheller

PUBLISHED MONDAY, OCT. 3, 2016 12:56 P.M. EDT
UPDATED TUESDAY, NOV. 8, 2016, 12:43 A.M. EST

Probability: a measure of how likely it is that a proposition is true.

Probability: a measure of how likely it is that a proposition is true.

Frequentist definition

n/N where n = number of occurrences in a large number of "independent, identically distributed" (i.i.d.) trials N.

- "Independent" = one trial does not affect another trial (more precise definition later)
- "Identically distributed" = (one expects) each trial to behave the same way

Probability: a measure of how likely it is that a proposition is true.

Frequentist definition

n/N where n = number of occurrences in a large number of "independent, identically distributed" (i.i.d.) trials N.

- "Independent" = one trial does not affect another trial (more precise definition later)
- "Identically distributed" = (one expects) each trial to behave the same way

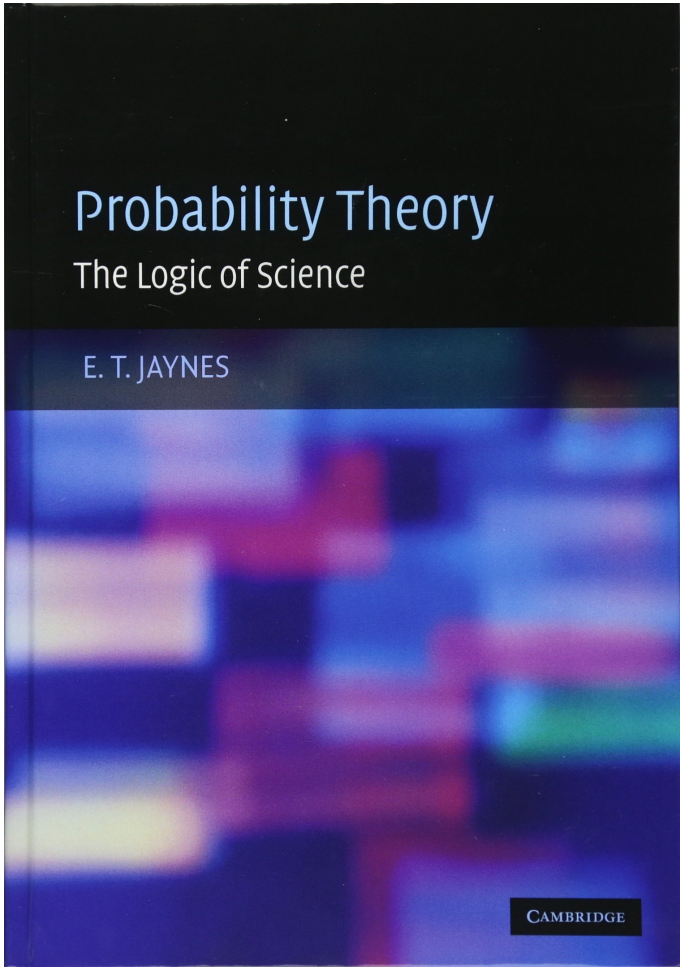
Applies to coin tosses... but not to weather, cricket matches, elections!

Bayesian definition

A real number between 0 and 1 quantifying your degree of belief in a proposition.

This made many 20th-century statisticians very uncomfortable... but this methodology is widely accepted now

A good but polemical reference



From Jaynes:

Suppose some dark night a policeman walks down a street, apparently deserted. Suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion? Let us first take a leisurely look at the general nature of such problems.

“Plausible reasoning”

Syllogism

- If A is true, B is true
  A is true
  therefore B is true

OR

- If A is true, B is true
  B is false
  therefore A is false

“Plausible reasoning”

Syllogism

- If A is true, B is true
  A is true
  therefore B is true

OR

- If A is true, B is true
  B is false
  therefore A is false

Weak syllogism

- If A is true, B is true
  A is false
  therefore B becomes less plausible

OR

- If A is true, B is true
  B is true
  therefore A becomes more plausible

- If the man were a criminal, he would probably be wearing a mask, breaking into a shop, carrying a sack, etc. He is doing that so it is plausible that he is a criminal.

- If the man were a criminal, he would probably be wearing a mask, breaking into a shop, carrying a sack, etc. He is doing that so it is plausible that he is a criminal.
- Suppose the policeman witnessed such a scene night after night and each time the man turned out to be innocent. He would gradually start considering it less plausible that the man is a criminal.
  - Cf. “The boy who cried wolf”

- If the man were a criminal, he would probably be wearing a mask, breaking into a shop, carrying a sack, etc. He is doing that so it is plausible that he is a criminal.
- Suppose the policeman witnessed such a scene night after night and each time the man turned out to be innocent. He would gradually start considering it less plausible that the man is a criminal.
  - Cf. “The boy who cried wolf”
- Bayesian reasoning = formalization of the above!

Probability theory: Some definitions

- Conditional probability: The probability of A given that B has occurred.

Probability theory: Some definitions

- Conditional probability: The probability of A given that B has occurred.
- Joint probability: The probability of both A and B occurring.

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

$$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$$

Probability theory: Some definitions

- Conditional probability: The probability of A given that B has occurred.
- Joint probability: The probability of both A and B occurring.

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

$$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$$

- Likelihood: The probability of observed data given a particular hypothesis.

Jaynes (following Cox, Polya and others):

- If you assume these "desiderata" for a theory of probability
  1. Degrees of plausibility are represented by real numbers. Greater number = greater plausibility. Also, "continuity" property.
  2. Qualitative correspondence with common sense. For example, if

$$P(A|C') > P(A|C)$$

and

$$P(B|C') = P(B|C)$$

then

$$P(AB|C') \geq P(AB|C)$$

and

$$P(\bar{A}|C') < P(\bar{A}|C),$$

  3. The system must be "consistent" – cannot derive contradictory results
- Then you arrive at a unique theory of probability which corresponds with "frequentist" probability and with common sense!

Bayes' theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$
$$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$$

Bayes' theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$
$$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$$

-
- From first equation, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Bayes' theorem

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

$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$

-
- From first equation, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Let B = data, call it D; A = hypothesis, call it h.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

.

Bayes' theorem

$P(AB) = P(A|B)P(B) = P(B|A)P(A)$
$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$

-
- From first equation, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Let B = data, call it D; A = hypothesis, call it h.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

.

- Usually we have a set of "mutually exclusive" hypotheses h_i, observed data D, and a way to calculate $P(D|h_i)$

Bayes' theorem

$P(AB) = P(A|B)P(B) = P(B|A)P(A)$
$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$

-
- From first equation, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Let B = data, call it D; A = hypothesis, call it h.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

.

- Usually we have a set of "mutually exclusive" hypotheses h_i, observed data D, and a way to calculate $P(D|h_i)$
- If the h_i are exhaustive, $P(D) = \sum_j P(D|h_j)P(h_j)$.

Bayes' theorem

$P(AB) = P(A|B)P(B) = P(B|A)P(A)$
$P(A \text{ OR } B) = P(A) + P(B) - P(AB)$

-
- From first equation, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Let B = data, call it D; A = hypothesis, call it h.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

.

- Usually we have a set of "mutually exclusive" hypotheses h_i, observed data D, and a way to calculate $P(D|h_i)$
- If the h_i are exhaustive, $P(D) = \sum_j P(D|h_j)P(h_j)$.

Bayes' formula

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Basics: Bayes' theorem

Bayes' formula

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Terminology:

- $P(h_i)$: *prior probability* of h_i
- $P(D|h_i)$: *likelihood* of data given h_i
- $P(h_i|D)$: *posterior probability* of h_i given D

Basics: Bayes' theorem

Bayes' formula

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Terminology:

- $P(h_i)$: *prior probability* of h_i
- $P(D|h_i)$: *likelihood* of data given h_i
- $P(h_i|D)$: *posterior probability* of h_i given D

Bayesian learning...

The *posterior probabilities* calculated from the current data become the *prior probabilities* for the next set of data!

Inferring probability distributions from data

Suppose

- You have some data that could be explained by two or more possible hypotheses
- You have some prior probabilities (beliefs) for each hypothesis
- You see some new data
- What is the *likelihood* or the *posterior probability* of each hypothesis given the new data?

Is a coin fair or biased?

You have a coin that was made in a factory where one in 10,000 coins is "biased": it tosses as heads 60% of the time. You do not know whether this coin is fair or biased.

- You toss the coin 10 times, see 8 heads. Given this, what's the probability that it's biased?
- You toss the coin 100 times, see 58 heads. What's the probability that it's biased?
- You toss the coin 300 times, see 190 heads. What's the probability that it's biased?
- You toss the coin 1000 times, see 615 heads. What's the probability that it's biased?

Is a coin fair or biased?

Remember the binomial distribution. Let the data be $D = N$ tosses, n heads. Then

$$P(D|H) = \binom{N}{n} p_H^n (1 - p_H)^{N-n}$$

where p_H is the probability of tossing heads under hypothesis H.
Here, we have two hypotheses, "fair" (F) and "biased" (B), under which the probabilities of heads are p_F and p_B.
Given data D, Bayes' Theorem says

$$P(B|D) = \frac{P(D|B)P(B)}{P(D|B)P(B) + P(D|F)P(F)}$$

and the "priors" are $P(B) = 0.0001$, $P(F) = 0.9999$.

Is a coin fair or biased?

$$P(B|D) = \frac{P(D|B)P(B)}{P(D|B)P(B) + P(D|F)P(F)}$$

So we have.

- You toss the coin 10 times, see 8 heads. $P(B|D) = 0.000275\ldots$
- You toss the coin 100 times, see 58 heads. $P(B|D) = 0.000333\ldots$
- You toss the coin 300 times, see 190 heads. $P(B|D) = 0.707\ldots$
- You toss the coin 1000 times, see 615 heads. $P(B|D) = 0.999999\ldots$

Sequences

- Biology:
  `CTGACAGAGACACCCGATTACTGATTTGGGAAATTTCCCAAATTGGAAATA...`

Sequences

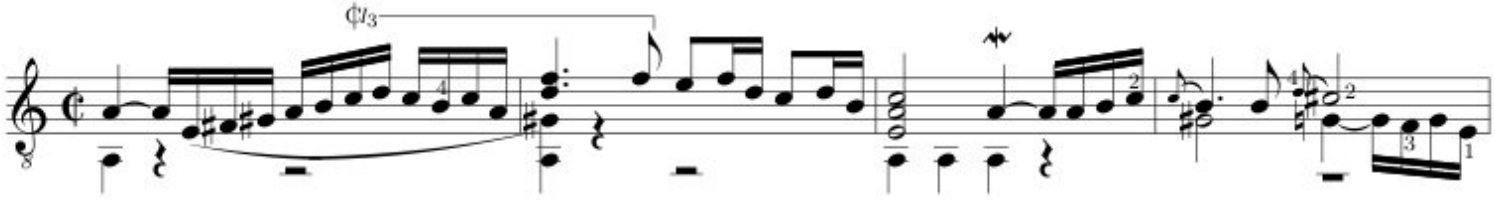
- Biology:
  `CTGACAGAGACACCCGATTACTGATTTGGGAAATTTCCCAAATTGGAAATA...`
- Language:
  `Persons attempting to find a motive in this narrative will be prosecuted; persons attempting to find a moral in it will be banished; persons attempting to find a plot in it will be shot.`

Sequences

- Biology:
  `CTGACAGAGACACCCGATTACTGATTTGGGAAATTTCCCAAATTGGAAATA...`
- Language:
  `Persons attempting to find a motive in this narrative will be prosecuted; persons attempting to find a moral in it will be banished; persons attempting to find a plot in it will be shot.`
  - Letter-level:
    `[P,e,r,s,o,n,s, ,a,...]`
  - Word-level:
    `[Persons, attempting, to, find...]`

Sequences

- Biology:
  `CTGACAGAGACACCCGATTACTGATTTGGGAAATTTCCCAAATTGGAAATA...`
- Language:
  `Persons attempting to find a motive in this narrative will be prosecuted; persons attempting to find a moral in it will be banished; persons attempting to find a plot in it will be shot.`
  - Letter-level:
    `[P,e,r,s,o,n,s, ,a,...]`
  - Word-level:
    `[Persons, attempting, to, find...]`
- Music:



What's common to those sequences?

- They can be written as linear sequences of a discrete, finite set of symbols ("alphabet")
- They can be very long
- They contain meaning
- At short scales, they contain correlations but are not perfectly ordered
- At longer scales, they are uncorrelated
- And more...

Why model sequences?

- Biologists want to understand the function of DNA (and protein) sequence, and design synthetic functional sequence
- Computational linguists want to use computers to parse, process, and create "natural language"
- Computer-created text and music conveys a better understanding of what goes into the "real" stuff
- Scholars want to compare and analyse works, assess authenticity, etc
- And so on...

Questions

Basic questions

Given a "model" that describes the sequence,

- What is the probability ("likelihood") of observing a particular sequence?
- Given a sequence, how do you predict ("generate") the next element?

Questions

Basic questions

Given a "model" that describes the sequence,

- What is the probability ("likelihood") of observing a particular sequence?
- Given a sequence, how do you predict ("generate") the next element?

If we can do the above:

Given multiple models for generating a sequence, how do we choose the more probable model?

Language models

Definition

A probability distribution $p(s)$ over strings s that attempts to reflect how frequently a string s occurs as a sentence.

Chen and Goodman, 1998

Markov chains

- A Markov chain is a sequence of symbols where each symbol depends only on its predecessor (or n predecessors)

Markov chains

- A Markov chain is a sequence of symbols where each symbol depends only on its predecessor (or n predecessors)
- Consider a sequence of symbols

$$S_1 S_2 S_3 S_4 S_5$$

What is the probability of observing this sequence?

Markov chains

- A Markov chain is a sequence of symbols where each symbol depends only on its predecessor (or n predecessors)
- Consider a sequence of symbols

$$S_1 S_2 S_3 S_4 S_5$$

What is the probability of observing this sequence?

- "Exact" answer:

$$P(S_1)P(S_2|S_1)P(S_3|S_1S_2)P(S_4|S_1S_2S_3)P(S_5|S_1S_2S_3S_4)$$

Markov chains

- A Markov chain is a sequence of symbols where each symbol depends only on its predecessor (or n predecessors)
- Consider a sequence of symbols

$$S_1 S_2 S_3 S_4 S_5$$

What is the probability of observing this sequence?

- "Exact" answer:

$$P(S_1)P(S_2|S_1)P(S_3|S_1S_2)P(S_4|S_1S_2S_3)P(S_5|S_1S_2S_3S_4)$$

- Markov approximation:

$$P(S_1)P(S_2|S_1)P(S_3|S_2)P(S_4|S_3)P(S_5|S_4)$$

Example: DNA sequence

If the DNA is “non-functional”, how to model it?

Example: DNA sequence

If the DNA is "non-functional", how to model it?

- Simplest model: each nucleotide occurs independently with a certain probability.
  Eg, $P(\mathrm{A}) = P(\mathrm{T}) = 0.3$, $P(\mathrm{C}) = P(\mathrm{G}) = 0.2$ (4 probabilities)
- However, dinucleotides in DNA are not distributed according to this model!

Example: DNA sequence

If the DNA is "non-functional", how to model it?

- Simplest model: each nucleotide occurs independently with a certain probability.
  Eg, $P(\mathrm{A}) = P(\mathrm{T}) = 0.3$, $P(\mathrm{C}) = P(\mathrm{G}) = 0.2$ (4 probabilities)
- However, dinucleotides in DNA are not distributed according to this model!
- Next simplest: Each nucleotide depends on its predecessor
  $P(\mathrm{A} \text{ at site } 2 | \mathrm{C} \text{ at site } 1)$, etc. (16 such "conditional probabilities"), "Markov chain"
- Still doesn't account for "trigrams"

Example: DNA sequence

If the DNA is "non-functional", how to model it?

- Simplest model: each nucleotide occurs independently with a certain probability.
  Eg, $P(\mathrm{A}) = P(\mathrm{T}) = 0.3$, $P(\mathrm{C}) = P(\mathrm{G}) = 0.2$ (4 probabilities)
- However, dinucleotides in DNA are not distributed according to this model!
- Next simplest: Each nucleotide depends on its predecessor
  $P(\mathrm{A} \text{ at site } 2 | \mathrm{C} \text{ at site } 1)$, etc. (16 such "conditional probabilities"), "Markov chain"
- Still doesn't account for "trigrams"
- Each nucleotide depends on immediate two predecessors
  $P(\mathrm{A}|\mathrm{CG})$, etc (64 conditional probabilities), 2nd order Markov chain

Example: DNA sequence

If the DNA is "non-functional", how to model it?

- Simplest model: each nucleotide occurs independently with a certain probability.
  Eg, $P(\mathrm{A}) = P(\mathrm{T}) = 0.3$, $P(\mathrm{C}) = P(\mathrm{G}) = 0.2$ (4 probabilities)
- However, dinucleotides in DNA are not distributed according to this model!
- Next simplest: Each nucleotide depends on its predecessor
  $P(\mathrm{A} \text{ at site } 2 | \mathrm{C} \text{ at site } 1)$, etc. (16 such "conditional probabilities"), "Markov chain"
- Still doesn't account for "trigrams"
- Each nucleotide depends on immediate two predecessors
  $P(\mathrm{A}|\mathrm{CG})$, etc (64 conditional probabilities), 2nd order Markov chain

If not good enough: go to higher-order Markov.

Example: DNA sequence

If the DNA is "non-functional", how to model it?

- Simplest model: each nucleotide occurs independently with a certain probability.
  Eg, $P(\mathrm{A}) = P(\mathrm{T}) = 0.3$, $P(\mathrm{C}) = P(\mathrm{G}) = 0.2$ (4 probabilities)
- However, dinucleotides in DNA are not distributed according to this model!
- Next simplest: Each nucleotide depends on its predecessor
  $P(\mathrm{A} \text{ at site } 2 | \mathrm{C} \text{ at site } 1)$, etc. (16 such "conditional probabilities"), "Markov chain"
- Still doesn't account for "trigrams"
- Each nucleotide depends on immediate two predecessors
  $P(\mathrm{A}|\mathrm{CG})$, etc (64 conditional probabilities), 2nd order Markov chain

If not good enough: go to higher-order Markov.
Drawback: for alphabet size ℓ, there are ℓ^n n-grams! Lots of data needed to estimate these.

Example: Shannon, 1948

(C. E. Shannon, *A mathematical theory of communication*, 1948)

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

Example: Shannon, 1948

(C. E. Shannon, *A mathematical theory of communication*, 1948)

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that these samples have reasonably good structure out to about twice the range that is taken into account in their construction. Thus in (3) the statistical process insures reasonable text for two-letter sequences, but four-letter sequences from the sample can usually be fitted into good sentences. In (6) sequences of four or more words can easily be placed in sentences without unusual or strained constructions. The particular sequence of ten words "attack on an English writer that the character of this" is not at all unreasonable. It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source.

The zero problem, and smoothing

If you are training from insufficient data, some possible observations may never occur in your data.

Simple example: coin-tossing:

Suppose you have a possibly unfair coin, toss it N times, and see n heads. What is the probability of seeing heads on the next toss?
"Maximum likelihood" answer = $\frac{n}{N}$. Bad answer!

The zero problem, and smoothing

If you are training from insufficient data, some possible observations may never occur in your data.

Simple example: coin-tossing:

Suppose you have a possibly unfair coin, toss it N times, and see n heads. What is the probability of seeing heads on the next toss?
"Maximum likelihood" answer = $\frac{n}{N}$. Bad answer!

Laplace's rule of succession

If you are completely ignorant about the coin's bias, the answer is

$$P(\text{heads}) = \frac{n+1}{N+2}$$

(doesn't usually apply to real coins!)

The zero problem, and smoothing

Laplace's rule: Generalization

If there are ℓ possible symbols that you can observe, and in N observations you have observed the i'th symbol n_i times; Answer, assuming complete independence of observations and complete ignorance,

$$P(i) = \frac{n_i + 1}{N + \ell}$$

Details...

Bernoulli trials (eg coins)

- Each trial has two possible outcomes, S (success) and F (failure)
- Probability of S is p, with p unknown
- You conduct N trials and S occurs n times
- What is the probability of S the next time?

Binomial distribution

- Distribution for N Bernoulli trials = *Binomial distribution*
- Probability of seeing n successes is

$$P(n; N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

where

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

- But what if p is unknown?

Binomial distribution

- Probability of seeing n successes is

$$P(n; N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

Binomial distribution

- Probability of seeing n successes is

$$P(n; N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

- If p is unknown, we integrate over all values of p.
- But are all values of p equally probable, *a priori*? We should consider a "prior" distribution on p, $P(p)$.

Binomial distribution

- Probability of seeing n successes is

$$P(n; N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

- If p is unknown, we integrate over all values of p.
- But are all values of p equally probable, *a priori*? We should consider a "prior" distribution on p, $P(p)$.
- Suppose they are equally probable ("ignorance prior", $P(p) = 1$).

$$\begin{aligned} P(n; N, p) &= \binom{N}{n} \int_0^1 p^n (1-p)^{N-n} \\ &= \binom{N}{n} B(n+1, N-n+1) = \binom{N}{n} \frac{n!(N-n)!}{(N+1)!} \end{aligned}$$

Laplace's rule of succession

$$P(n; N, p) = \binom{N}{n} \frac{n!(N-n)!}{(N+1)!}$$

Having observed n successes in N trials, what is the probability of observing a success next time, $P(\mathrm{S}|n \text{ in } N)$?

Laplace's rule of succession

$$P(n; N, p) = \binom{N}{n} \frac{n!(N-n)!}{(N+1)!}$$

Having observed n successes in N trials, what is the probability of observing a success next time, $P(\mathrm{S}|n \text{ in } N)$?

- $P(A|B) = P(AB)/P(B)$

Laplace's rule of succession

$$P(n; N, p) = \binom{N}{n} \frac{n!(N-n)!}{(N+1)!}$$

Having observed n successes in N trials, what is the probability of observing a success next time, $P(\mathrm{S}|n \text{ in } N)$?

- $P(A|B) = P(AB)/P(B)$
- $$P(\mathrm{S}|n \text{ in } N) = \frac{P(\mathrm{S} \text{ in most recent trial AND } n \text{ in } N)}{P(\mathrm{S}|n \text{ in } N)}$$
$$= \frac{\binom{N}{n} \frac{(n+1)!(N-n)!}{(N+2)!}}{\binom{N}{n} \frac{n!(N-n)!}{(N+1)!}}$$
$$= \frac{n+1}{N+2}$$

Beta prior

Having started with a uniform "prior" over p, what is the "posterior" probability of a given p, given n successes?

Beta prior

Having started with a uniform "prior" over p, what is the "posterior" probability of a given p, given n successes?
Bayes' theorem:

$$P(p|\text{data}) = \frac{P(\text{data}|p)P(p)}{\int_0^1 P(\text{data}|p)P(p)dp}$$

Since our prior was uniform, we can easily evaluate:

$$P(p|n,N) = \frac{\binom{N}{n} p^n (1-p)^{N-n}}{\binom{N}{n} \frac{n!(N-n)!}{(N+1)!}} = \frac{(N+1)!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Beta prior

Having started with a uniform "prior" over p, what is the "posterior" probability of a given p, given n successes?
Bayes' theorem:

$$P(p|\text{data}) = \frac{P(\text{data}|p)P(p)}{\int_0^1 P(\text{data}|p)P(p)dp}$$

Since our prior was uniform, we can easily evaluate:

$$P(p|n,N) = \frac{\binom{N}{n} p^n (1-p)^{N-n}}{\binom{N}{n} \frac{n!(N-n)!}{(N+1)!}} = \frac{(N+1)!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Beta prior

$$P_{\text{Beta}}(p|c_1, c_2) \propto p^{c_1-1}(1-p)^{c_2-1}$$

Beta prior

$$P_{\text{Beta}}(p|c_1, c_2) \propto p^{c_1-1}(1-p)^{c_2-1}$$

Beta prior

$$P_{\text{Beta}}(p|c_1, c_2) \propto p^{c_1-1}(1-p)^{c_2-1}$$

Beta prior is “conjugate prior” of binomial distribution:

Beta prior

$$P_{\text{Beta}}(p|c_1, c_2) \propto p^{c_1-1}(1-p)^{c_2-1}$$

Beta prior is "conjugate prior" of binomial distribution:

- Ignorance prior ($c_1 = 1, c_2 = 1$) is special case

Beta prior

$$P_{\text{Beta}}(p|c_1, c_2) \propto p^{c_1-1}(1-p)^{c_2-1}$$

Beta prior is "conjugate prior" of binomial distribution:

- Ignorance prior ($c_1 = 1, c_2 = 1$) is special case
- If prior is of Beta form with $P_{\text{Beta}}(p; c_1, c_2)$, then posterior is also of Beta form $P_{\text{Beta}}(p; c_1 + n, c_2 + N - n) \propto p^{c_1+n-1}(1-p)^{c_2+N-n-1}$

Beta prior

$$P_{\text{Beta}}(p|c_1, c_2) \propto p^{c_1-1}(1-p)^{c_2-1}$$

Beta prior is "conjugate prior" of binomial distribution:

- Ignorance prior ($c_1 = 1, c_2 = 1$) is special case
- If prior is of Beta form with $P_{\text{Beta}}(p; c_1, c_2)$, then posterior is also of Beta form $P_{\text{Beta}}(p; c_1 + n, c_2 + N - n) \propto p^{c_1+n-1}(1-p)^{c_2+N-n-1}$
- If you observe n successes, with a Beta prior $P_{\text{Beta}}(p; c_1, c_2)$, then the probability of another success is

$$P(\text{success}|n \text{ successes}) = \frac{n + c_1}{N + c_1 + c_2}$$

(therefore, c_1 and c_2 called "pseudocounts")

Nucleotides

- There are four nucleotides, A, C, G, T

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...
  - Describe probability of each nucleotide α by w_α

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...
  - Describe probability of each nucleotide α by w_α
  - $\sum_\alpha w_\alpha = 1$

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...
  - Describe probability of each nucleotide α by w_α
  - $\sum_\alpha w_\alpha = 1$
  - Suppose you have N observations, and each nucleotide α occurs n_α times

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...
  - Describe probability of each nucleotide α by w_α
  - $\sum_\alpha w_\alpha = 1$
  - Suppose you have N observations, and each nucleotide α occurs n_α times
  - Appropriate distribution = "multinomial distribution"

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...
  - Describe probability of each nucleotide α by w_α
  - $\sum_\alpha w_\alpha = 1$
  - Suppose you have N observations, and each nucleotide α occurs n_α times
  - Appropriate distribution = "multinomial distribution"
  - Appropriate prior = "Dirichlet prior": $P(w; c_\alpha) \propto \prod_\alpha w_\alpha^{c_\alpha - 1}$

Nucleotides

- There are four nucleotides, A, C, G, T
- Distribution = "multinomial distribution"
- Straightforward generalisation of preceding discussion...
  - Describe probability of each nucleotide α by w_α
  - $\sum_\alpha w_\alpha = 1$
  - Suppose you have N observations, and each nucleotide α occurs n_α times
  - Appropriate distribution = "multinomial distribution"
  - Appropriate prior = "Dirichlet prior": $P(w; c_\alpha) \propto \prod_\alpha w_\alpha^{c_\alpha - 1}$
  - With Dirichlet prior, after N observations as above, probability of nucleotide α in next observation is

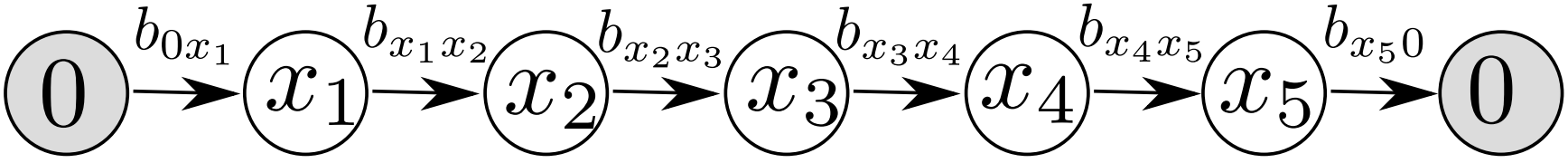
$$\frac{n_\alpha + c_\alpha}{N + C}$$

    where $C = \sum_\alpha c_\alpha$.

Hidden Markov models

In Markov models

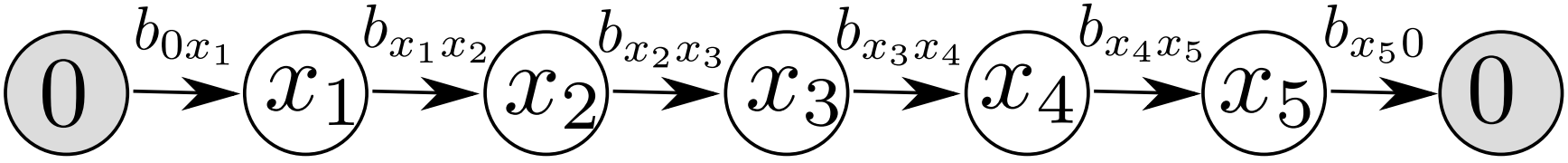
symbols are emitted probabilistically based on the previous symbol.



Hidden Markov models

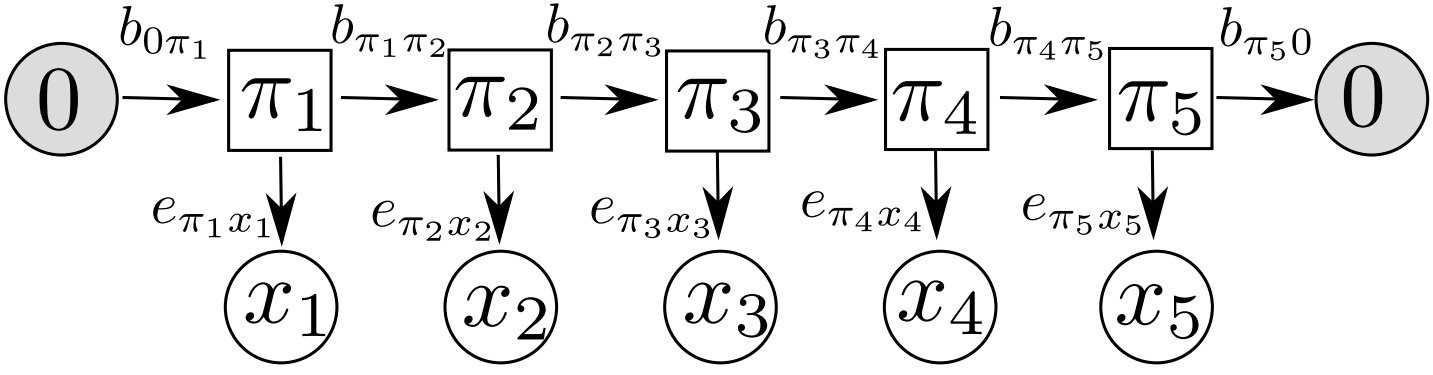
In Markov models

symbols are emitted probabilistically based on the previous symbol.

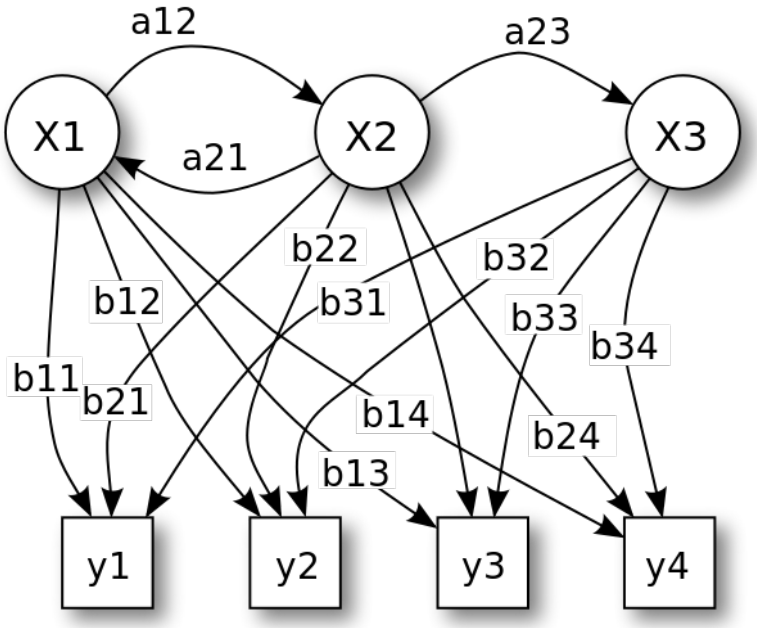


In hidden Markov models

an invisible "state" follows a Markov process. This state emits symbols that are visible. Each hidden state emits a different pattern of symbols.



Hidden Markov models



Source: Wikipedia

Tasks

- Infer hidden states (Viterbi algorithm)
- Infer likelihood of sequence (forward/backward algorithms)

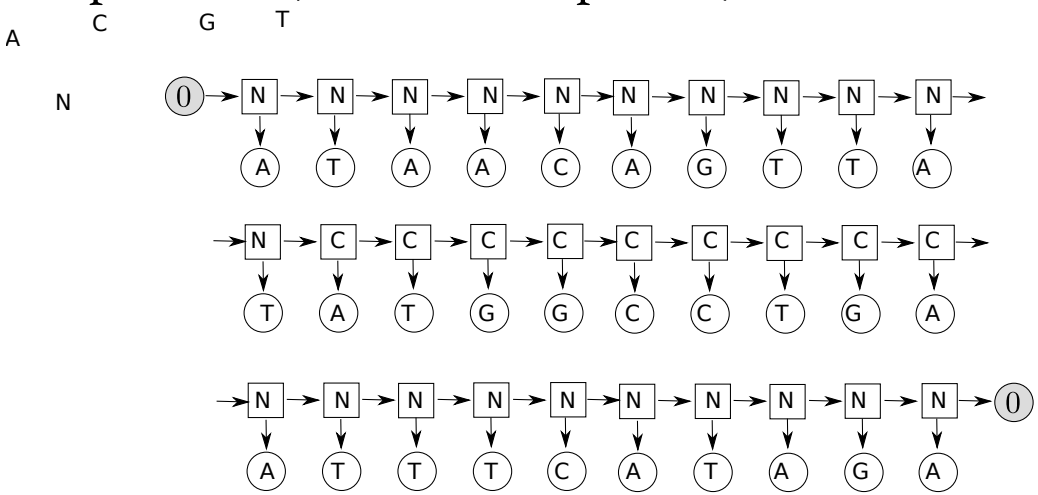
Hidden Markov models

Silly example

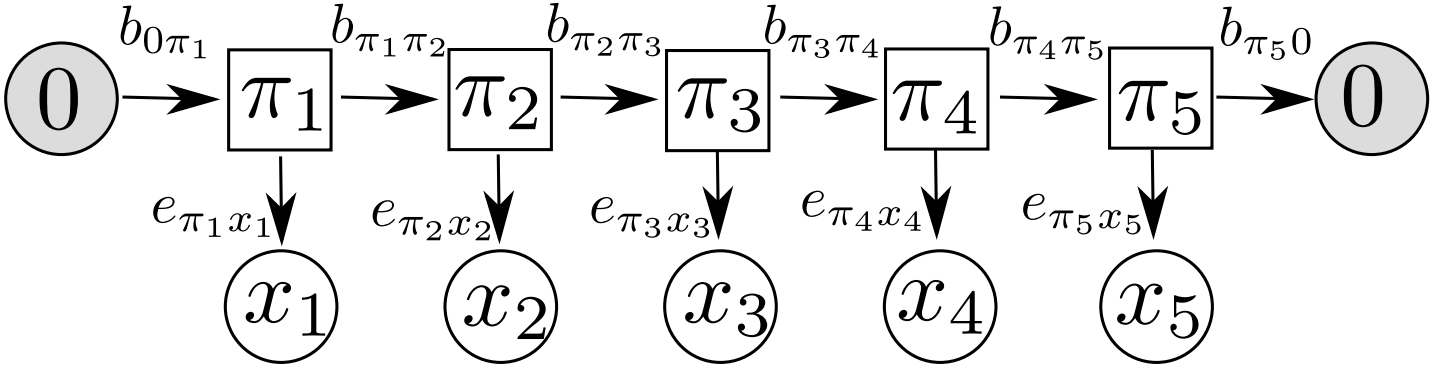
This sentence seems un peu unusual parce que it is made of deux langues différents.

Somewhat more realistic example

Imagine a very simple model of DNA where there are two kinds of regions – coding (C) and noncoding (N) – characterised mainly by differences in nucleotide densities. Non-coding regions are AT rich, coding regions are more GC rich. This is a possible (but oversimplified) HMM.



Most probable path: Viterbi algorithm



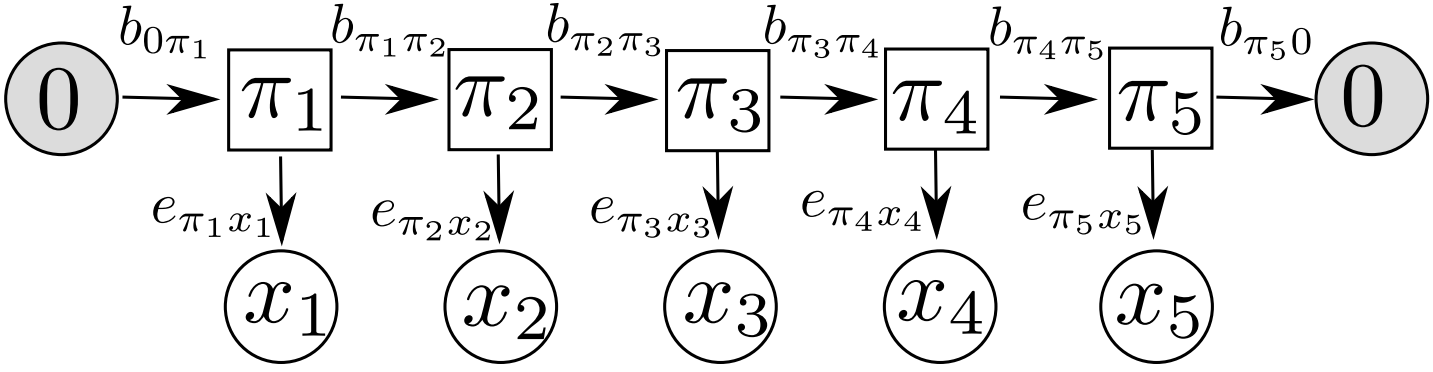
Given a sequence $x = x_1 x_2 \ldots x_n$ and a set of hidden states $\pi = \pi_1 \pi_2 \ldots \pi_n$, we can calculate the joint likelihood

$$P(x, \pi) = b_{0\pi_1} \left(\prod_{i=1}^{n} e_{\pi_i x_i} b_{\pi_i \pi_{i+1}} \right)$$

where $\pi_{n+1} = 0$.
How do we find the *most probable path*

$$\pi^* = \text{argmax}_\pi P(x, \pi)?$$

Viterbi algorithm: example



Suppose someone gives you a sequence of coin tosses, but the person has two coins and is randomly switching from one coin to another. One coin is fair ("F"). The other coin is biased ("B") and tosses heads 80% of the time. After each toss, the player can keep the same coin (probability 0.8), switch to the other coin (0.15), or end the game (0.05). Also, the probability of starting with the fair coin is 0.8. The sequence of tosses is HTTHHHHHTH. What is the most probable hidden path?

Viterbi algorithm: example

i \ k	0	F	B	x_i
1		0.4	0.12	H
2				T
3				T
4				H
5				H
6				H
7				H
8				H
9				T
10				H

Define a matrix v_{ki} = probability of most probable path up until state i, if $\pi_i = k$. Then

$v_{l,i+1} = e_{lx_{i+1}} \max_k \left(v_{ki} b_{kl}\right)$.

- Initialize $v_{00} = 1$, $v_{k0} = 0 \forall k > 0$.
- Fill in the matrix until the bottom right, *each time pointing back to the previous row entry that gave the best answer*
- Trace back arrows from largest entry on bottom row

Viterbi algorithm: example

i \ k	0	F	B	x_i
1		0.4	0.12	H
2		0.16		T
3				T
4				H
5				H
6				H
7				H
8				H
9				T
10				H

Define a matrix v_{ki} = probability of most probable path up until state i, if $\pi_i = k$. Then

$v_{l,i+1} = e_{lx_{i+1}} \max_k (v_{ki} b_{kl})$.

- Initialize $v_{00} = 1$, $v_{k0} = 0 \forall k > 0$.
- Fill in the matrix until the bottom right, *each time pointing back to the previous row entry that gave the best answer*
- Trace back arrows from largest entry on bottom row

Viterbi algorithm: example

i \ k	o	F	B	x_i
1		0.4	0.12	H
2		0.16	0.0192	T
3				T
4				H
5				H
6				H
7				H
8				H
9				T
10				H

Define a matrix v_{ki} = probability of most probable path up until state i, if $\pi_i = k$. Then

$v_{l,i+1} = e_{lx_{i+1}} \max_k \left(v_{ki} b_{kl}\right)$.

- Initialize $v_{oo} = 1$, $v_{ko} = 0 \forall k > 0$.
- Fill in the matrix until the bottom right, *each time pointing back to the previous row entry that gave the best answer*
- Trace back arrows from largest entry on bottom row

Viterbi algorithm: example

i \ k	F	B	x_i
1	0.4	0.12	H
2	0.16	0.0192	T
3	0.064		T
4			H
5			H
6			H
7			H
8			H
9			T
10			H

Define a matrix v_{ki} = probability of most probable path up until state i, if $\pi_i = k$. Then

$v_{l,i+1} = e_{lx_{i+1}} \max_k (v_{ki} b_{kl})$.

- Initialize $v_{00} = 1$, $v_{k0} = 0 \forall k > 0$.
- Fill in the matrix until the bottom right, *each time pointing back to the previous row entry that gave the best answer*
- Trace back arrows from largest entry on bottom row

Viterbi algorithm: example

i \ k	0	F	B	x_i
1		0.4	0.12	H
2		0.16	0.0192	T
3		0.064	0.0048	T
4				H
5				H
6				H
7				H
8				H
9				T
10				H

Define a matrix v_{ki} = probability of most probable path up until state i, if $\pi_i = k$. Then

$v_{l,i+1} = e_{lx_{i+1}} \max_k \left(v_{ki} b_{kl}\right)$.

- Initialize $v_{00} = 1$, $v_{k0} = 0 \forall k > 0$.
- Fill in the matrix until the bottom right, *each time pointing back to the previous row entry that gave the best answer*
- Trace back arrows from largest entry on bottom row

Viterbi algorithm: example

i \ k	0	F	B	x_i
1		0.4	0.12	H
2		0.16	0.0192	T
3		0.064	0.0048	T
4		0.0256	0.00768	H
5		0.01024	0.00491	H
6		0.00409	0.003142	H
7		0.00163	0.002011	H
8		0.00065	0.001287	H
9		0.00026	0.000205	T
10		0.00010	0.000131	H

Define a matrix v_{ki} = probability of most probable path up until state i, if $\pi_i = k$. Then

$v_{l,i+1} = e_{lx_{i+1}} \max_k \left(v_{ki} b_{kl}\right)$.

- Initialize $v_{00} = 1$, $v_{k0} = 0 \forall k > 0$.
- Fill in the matrix until the bottom right, *each time pointing back to the previous row entry that gave the best answer*
- Trace back arrows from largest entry on bottom row

Likelihood of sequence: forward algorithm

The Viterbi algorithm gives you the most probable value of $P(x, \pi)$. But what if you only care about

$$P(x) = \sum_{\pi} P(x, \pi)$$

and not about the hidden path?
"Forward algorithm" lets you do that: define

$$f_{ki} = P(x_1 \ldots x_i | \pi_i = k).$$

Then

$$f_{l,i+1} = e_{l,x_{i+1}} \sum_{k} f_{ki} b_{kl}.$$

Likelihood of sequence: forward algorithm

The Viterbi algorithm gives you the most probable value of $P(x, \pi)$. But what if you only care about

$$P(x) = \sum_{\pi} P(x, \pi)$$

and not about the hidden path?
"Forward algorithm" lets you do that: define

$$f_{ki} = P(x_1 \ldots x_i | \pi_i = k).$$

Then

$$f_{l,i+1} = e_{l,x_{i+1}} \sum_{k} f_{ki} b_{kl}.$$

There is also "backward algorithm", "posterior decoding" and much more...

Why is this useful?

- Probability theory applies to every area of science
- Much of modern "machine learning" is built on Bayesian probability theory
- Sequence models are applicable to biology, language, signal processing, music, and much more...

Thank you