Unit 1: Sequence Models

Lecture 4a: Probabilities and Estimations

Lecture 4b: Weighted Finite-State Machines

Liang Huang
Probabilities

- experiment (e.g., “toss a coin 3 times”)
- basic outcomes $\Omega$ (e.g., $\Omega=$\{HHH, HHT, HTH, ..., TTT\})
- event: some subset $A$ of $\Omega$ (e.g., $A =$ “heads twice”)
- probability distribution
  - a function $p$ from $\Omega$ to $[0, 1]$
  - $\sum_{e \in \Omega} p(e) = 1$
- probability of events (marginals)
  - $p(A) = \sum_{e \in A} p(e)$
Joint and Conditional Probs

\[ P(A, B) = P(A \cap B) \]

Conditional probability of \( A \) given \( B \)

\[ P(A | B) = \frac{P(A, B)}{P(B)} \]

\[ \begin{align*}
\text{NOTE:} \\
P(B | A) & = \frac{P(A, B)}{P(A)}
\end{align*} \]

Examples:

\( A = \) "heads twice" \( \quad P(A) = \frac{3}{8} \)
\( B = \) "first flip = heads" \( \quad P(B) = \frac{1}{2} \)

\[ P(B | A) = \text{"if you got two heads, what is chance the first flip was heads?"} \]

\[ P(A, B) \neq P(A) + P(B) \]
Multiplication Rule

\[
P(A, B) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B)
\]

Chain rule generalization

\[
P(A_1, \ldots, A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1, A_2) \ldots P(A_n | A_1, \ldots, A_{n-1})
\]
Independence

- \( P(A, B) = P(A) \cdot P(B) \quad \text{or} \quad P(A) = P(A|B) \)
- disjoint events are always dependent! \( P(A,B) = 0 \)
  - unless one of them is “impossible”: \( P(A)=0 \)
- conditional independence: \( P(A, B|C) = P(A|C) \cdot P(B|C) \)
  \[ \frac{P(A|C)}{P(A|B, C)} = P(A|B, C) \]

\( A, B \) are independent if \( P(A) = P(A|B) \)

\( P(\text{FRIDAY}) = \frac{1}{7} \)
\( P(\text{FRIDAY} | \text{RAIN}) = \frac{1}{7} \)

\( P(\text{FRIDAY}) = \frac{1}{7} \)
\( P(\text{FRIDAY} | \text{RAIN}) = 0 \)

\( P(\text{FRIDAY}) = \frac{1}{7} \)
\( P(\text{FRIDAY} | \text{RAIN}) = \frac{4}{5} \)
Marginalization

- compute marginal probs from joint/conditional probs

\[ P(B) = P(A, B) + P(\overline{A}, B) \]

\[ = P(A) \cdot P(B|A) + P(\overline{A}) \cdot P(B|\overline{A}) \]
Bayes Rules

Bayes Rule

\[ P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \]

proof: \[ P(A,B) = P(A) \cdot P(B|A) \]
\[ \overset{\text{equate & divide by } P(B)}{=} P(B) \cdot P(A|B) \]

alternative bayes rule by partition
Most Likely Event

\[
\text{most likely event}
\]

\[
\arg\max_{A_i} P(A_i)
\]

(e.g., if \(A_i \in \{\text{rain, snow, sun}\}\)

and \(P(\text{rain}) = 0.1\)

\(P(\text{snow}) = 0.1\)

\(P(\text{sun}) = 0.8\)

then

\[
\arg\max_{A_i} P(A_i) = \text{sun}
\]
\[ \arg\max_{A_i} P(A_i \mid B) = \arg\max_{A_i} P(A_i) \cdot P(B \mid A_i) \]

The right-hand side is often easier to work with.

E.g., let \( B \) be a misspelled character sequence.
Let \( A_1 \ldots A_n \) be possible corrections.

\[ P(A_i \mid B) \Rightarrow P(\text{knight} \mid \text{kmight}) \]
\[ P(A_i) \cdot P(B \mid A_i) \Rightarrow P(\text{knight}) \cdot P(\text{kmight} \mid \text{knight}) \]

The "prior" term allows us to bring a lot of background knowledge to bear.
Estimating Probabilities

- how to get probabilities for basic outcomes?
  - do experiments
  - count stuff

- e.g. how often do people start a sentence with “the”?
  - \( P(A) = \frac{\text{(# of sentences like “the ...” in the sample)}}{\text{(# of all sentences in the sample)}} \)
  - \( P(A \mid B) = \frac{\text{(count of A, B)}}{\text{(count of B)}} \)

- we will show that this is Maximum Likelihood Estimation
what is a MODEL?

- a general theory of how the data is generated,
- along with a set of parameter estimates
- e.g., given this statistics
  - we can “guess” it’s generated by a 12-sided die
    - along with 11 free parameters $p(1), p(2), \ldots, p(11)$
  - alternatively, by two tosses of a single 6-sided die
    - along with 5 free parameters $p(1), p(2), \ldots, p(5)$
- which is better given the data? which better explains the data?
  - $\arg \max_m p(m|d) = \arg \max_m p(m) p(d|m)$
Maximum Likelihood Estimation

- always maximize **posterior**: what’s the best m given d?
- when do we use maximum likelihood estimation?
- with uniform **prior**, same as likelihood (explains data)
  - \[ \text{argmax}_m p(m|d) = \text{argmax}_m p(m) \cdot p(d|m) \quad \text{bayes, and } p(d)=1 \]
  - \[ = \text{argmax}_m p(d|m) \quad \text{when } p(m) \text{ uniform} \]

Suppose \[ d = H H T H \]

- \( m_1 \): coin is unbiased \[ p(d|m) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.0625 \]
- \( m_2 \): coin is biased so that \( p(H) = \frac{3}{4} \) \[ p(d|m) = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} = 0.105 \]
- \( m_3 \): coin is biased so that \( p(H) = \frac{9}{10} \) \[ p(d|m) = \frac{9}{10} \cdot \frac{9}{10} \cdot \frac{1}{10} \cdot \frac{9}{10} = 0.073 \]
How do we rigorously derive this?

- assuming any $p_m(H) = \theta$ is possible, what’s the best $\theta$?
- e.g.: data is still H, H, T, H.
- $\arg\max_\theta p(d|m; \theta) = \arg\max_\theta \theta^3 (1-\theta)$
  - take derivatives, make it zero: $\theta = 3/4$.
  - works in the general case: $\theta = n / (n+m)$ (n heads, m tails)
- this is why MLE is just counting in the discrete case
What if we have some prior?

- what if we have arbitrary prior
  - like $p(\theta) = \theta (1-\theta)$
- maximum a posteriori estimation (MAP)
- MAP approaches MLE with infinite
- MAP = MLE + smoothing
  - this prior is just “extra two tosses, unbiased”
  - you can inject other priors, like “extra 4 tosses, 3 Hs”
Probabilistic Finite-State Machines

• adding probabilities into finite-state acceptors (FSAs)

• FSA: a set of strings; WFSA: a distribution of strings

```
<table>
<thead>
<tr>
<th>FSA</th>
<th>x</th>
</tr>
</thead>
</table>
| accepts/generates some strings, 
  rejects others. |
```

```
<table>
<thead>
<tr>
<th>WFSA</th>
<th>P(x)</th>
<th>x</th>
</tr>
</thead>
</table>
| associates a prob. P(x) w/each x. 
  if P(x) = 0, x is rejected. |
```

weighted FSA represents set of pairs, e.g.:

\[
\left\{ \langle \text{you too, } 10^{-6} \rangle, \\
\langle \text{you two, } 10^{-7} \rangle, \\
\langle \text{you to, } 10^{-21} \rangle, \\
\langle \text{he the, } 0 \rangle, \\
\langle \text{i eat fish, } 10^{-8} \rangle \\
\ldots \right\}
\]
WFSA

- normalization: transitions leaving each state sum up to 1
- defines a distribution over strings?
- or a distribution over paths?
- \( \Rightarrow \) also induces a distribution over strings

\[ p(\text{he ran home}) = 0.08 \]
Max / Sum Probs

• in a WFSA, which string x has the greatest p(x)?
  • graph search (shortest path) problem
    • Dijkstra;
    • or Viterbi if the FSA is acyclic
  • does it work for NFA?
    • best path much easier than best string
    • you can determinize it (with exponential cost!)
    • popular work-around: n-best list crunching

Edsger Dijkstra (1930-2002)
“GOTO considered harmful”

Dr. Andrew Viterbi
(b. 1932)
Viterbi Alg. (1967)
CMDA, Qualcomm
A Note on Two Problems in Connexion with Graphs

By
E. W. DIJKSTRA

We consider \( n \) points (nodes), some or all pairs of which are connected by a branch; the length of each branch is given. We restrict ourselves to the case where at least one path exists between any two nodes. We now consider two problems.

**Problem 1.** Construct the tree of minimum total length between the \( n \) nodes. (A tree is a graph with one and only one path between every two nodes.)

In the course of the construction that we present here, the branches are subdivided into three sets:

I. the branches definitely assigned to the tree under construction (they will form a subtree);

II. the branches from which the next branch to be added to set I, will be selected;

III. the remaining branches (rejected or not yet considered).

Numerische Mathematik 1, 269—271 (1959)

that’s min. spanning tree!
Jarnik (1930) - Prim (1957) - Dijkstra (1959)

Edsger Dijkstra (1930-2002)
“GOTO considered harmful”
the data for at most $n$ branches, viz. the branches in sets I and II and the branch under consideration in step 2.

**Problem 2.** Find the path of minimum total length between two given nodes $P$ and $Q$.

We use the fact that, if $R$ is a node on the minimal path from $P$ to $Q$, knowledge of the latter implies the knowledge of the minimal path from $P$ to $R$. In the solution presented, the minimal paths from $P$ to the other nodes are constructed in order of increasing length until $Q$ is reached.

In the course of the solution the nodes are subdivided into three sets:

A. the nodes for which the path of minimum length from $P$ is known; nodes will be added to this set in order of increasing minimum path length from node $P$;

B. the nodes from which the next node to be added to set A will be selected; this set comprises all those nodes that are connected to at least one node of set A but do not yet belong to A themselves;

C. the remaining nodes.
IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-13, NO. 2, APRIL 1967

Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm

ANDREW J. VITERBI, SENIOR MEMBER, IEEE

Abstract—The probability of error in decoding an optimal convolutional code transmitted over a memoryless channel is bounded from above and below as a function of the constraint length of the code. For all but pathological channels the bounds are asymptotically (exponentially) tight for rates above $R_0$, the computational cutoff rate of sequential decoding. As a function of constraint length the performance of optimal convolutional codes is shown to be superior to that of block codes of the same length, the relative improvement increasing with rate. The upper bound is obtained for a specific probabilistic nonsequential decoding algorithm which is shown to be asymptotically optimum for rates above $R_0$ and whose performance bears certain similarities to that of sequential decoding algorithms.

I. SUMMARY OF RESULTS

SINCE Elias\cite{11} first proposed the use of convolutional (tree) codes for the discrete memoryless channel, it has been conjectured that the performance of this class of codes is potentially superior to that of block codes of the same length. The first quantitative verification of this conjecture was due to Yudkin\cite{21} who obtained

The author is with the Department of Engineering, University of California, Los Angeles, Calif.
Sum Probs

- what is $P(x)$ for some particular $x$?
  - for DFA, just follow $x$
  - for NFA,
    - get a subgraph (by composition), then sum ??
      - acyclic $\Rightarrow$ Viterbi
      - cyclic $\Rightarrow$ compute strongly connected components
        - SCC-DAG cluster graph (cyclic locally, acyclic globally)
        - do infinite sum (matrix inversion) locally, Viterbi globally
    - refer to extra readings on course website
WFSTs

- **FST**: a relation over strings (a set of string pairs)
- **WFST**: a probabilistic relation over strings (a set of \( <s, t, p> \): strings pair \( <s, t> \) with probability \( p \))
- What is \( p \) representing?

\[ \{ <\text{they can fish}, \text{PRO AUX V}, 0.99>, <\text{they can fish}, \text{PRO V N}, 0.01>, <\text{they can fish}, \text{DT DT DT}, 0.00>, <\text{they sing}, \text{PRO V}, 1.00>, \ldots \} \]

```
  a: a/0.9
  a: *ek/0.1
  b: b/0.9
  b: *ek/0.1
```

WFST

\[
P(ade | acdet) = 0.9 \cdot 0.1 \cdot 0.9 \cdot 0.9 \cdot 0.1 = 0.00729
\]
Edit Distance as WFST

- this is simplified edit distance
- real edit distance as an example of WFST, but not PFST

WFST: real edit distance

Costs:
- replacement: 1
- insertion: 2
- deletion: 2
Normalization

- if transitions leaving each state and each input symbol sum up to 1, then...
  - WFST defines conditional prob $p(y|x)$ for $x \Rightarrow y$
- what if we want to define a joint prob $p(x, y)$ for $x=\Rightarrow y$?
- what if we want $p(x \mid y)$?

$$P(\text{ade} \mid \text{acdet}) = 0.9 \cdot 0.1 \cdot 0.9 \cdot 0.9 \cdot 0.1 = .00729$$
Questions of WFSTs

• given $x$, $y$, what is $p(y|x)$?

• for a given $x$, what’s the $y$ that maximizes $p(y|x)$?

• for a given $y$, what’s the $x$ that maximizes $p(y|x)$?

• for a given $x$, supply all output $y$ w/ respective $p(y|x)$

• for a given $y$, supply all input $x$ w/ respective $p(y|x)$

\[ P(ade | acdelt) = 0.9 \times 0.1 \times 0.9 \times 0.9 \times 0.1 = 0.00729 \]
Answer: Composition

\[ p(z \mid x) = p(y \mid x) \ p(z \mid y) \]

\[ = \sum_y p(y \mid x) \ p(z \mid y) \]  have to sum up y

given y, z & x are independent in this cascade - Why?

how to build a composed WFST C out of WFSTs A, B?

again, like intersection

sum up the products

\((+, x)\) semiring
Example

\[ P(bt | dr) = 0.06 \cdot 0.5 = 0.03 \]
Example

from M. Mohri and J. Eisner

they use (min, +), we use (+, *)
Example

they use (min, +), we use (+, *)
Example

a:b .o. b:a = a:a

they use (min, +), we use (+, *)
Given $x$, supply all output $y$

no longer normalized!
Given x, y, what’s \( p(y|x) \)
Given $x$, what's max $p(y|x)$

$x = "ab"

\[ a:a/0.2 \quad b:b/0.2 \]

\[ a:A/0.8 \quad b:B/0.8 \]

\[ a:a/0.2 \quad b:b/0.2 \]

\[ a:A/0.8 \quad b:B/0.8 \]

remove input symbols

find best path

so, $\text{argmax } p(y|x|ab) = AB$
Part-of-Speech Tagging Again

Before Probabilities

\[
\begin{align*}
\text{they} & \xrightarrow{\text{Can}} \text{fish} \\
\text{Can: V} & \xrightarrow{\text{Can: N}} \text{Can: AUX} \\
\text{fish: V} & \xrightarrow{\text{no preferences}}
\end{align*}
\]
Part-of-Speech Tagging Again

Before Probabilities

After Probabilities

needs tweaking:
Adding a Tag Bigram Model (again)

\[ p(t...t|w...w) \]

wait, is that right (mathematically)?
Noisy-Channel Model
Noisy-Channel Model

\[ p(t...t) \]
**Applications of Noisy-Channel**

| Application                          | Input                  | Output                | $p(i)$                               | $p(o|i)$                                 |
|--------------------------------------|------------------------|-----------------------|--------------------------------------|------------------------------------------|
| Machine Translation                  | $L_1$ word sequences   | $L_2$ word sequences  | $p(L_1)$ in a language model         | translation model                        |
| Optical Character Recognition (OCR)  | actual text            | text with mistakes    | prob of language text                | model of OCR errors                      |
| Part Of Speech (POS) tagging         | POS tag sequences      | English words         | prob of POS sequences                | $p(w|t)$                                 |
| Speech recognition                   | word sequences         | speech signal         | prob of word sequences               | acoustic model                           |

**Diagram:**

- WFSA $\rightarrow$ $t \ldots t$ $\rightarrow$ WFST $\rightarrow$ $w \ldots w$
Example: Edit Distance

O(k) deletion arcs

O(k) insertion arcs

O(k) identity arcs

from J. Eisner
Example: Edit Distance

Best path (by Dijkstra’s algorithm)