1 Outline

- Review: HMMs
- Length distribution of states and generalized HMMs
- Conditional random fields

2 HMM review

An HMM is defined by the following parameters:

- \( Q \): A set of states
- \( O \): An alphabet of output symbols
- \( \pi \): initial state distribution
- \( A \): matrix of transition probabilities between all pairs of states in \( Q \), with \( a_{kl} = P(s_i = l | s_{i-1} = k) \)
- \( E \): Model emission probabilities. \( e_k(b) = P(e_i = b | s_i = k) \).

As a review, take a look at the following code fragment from the second problem set:

```python
base_idx = { 'A' : 0, 'G' : 1, 'C' : 2, 'T' : 3 }
state_idx = { '+' : 0, '-' : 1 }

# initial distribution over states, i.e. probability of starting in state k
init_dist = [0.5, 0.5]

# transition probabilities
tr = [
    [0.99, 0.01], # from + to -
    [0.01, 0.99]  # from - to +
]

# emission probabilities
em = [
    [0.20, 0.30, 0.30, 0.20], # +
    [0.30, 0.20, 0.20, 0.30]   # -
]
```

A couple of easy questions just to recall the different components of a hidden Markov model and some of the algorithms that we encountered in class:
• What is the set of states? (Note: “+” refers to high-GC, “-” refers to low-GC).
• What is the set of output symbols?
• What is the initial state distribution?
• What is the probability of transitioning from a high-GC state to a low-GC state?
• What is the probability of emitting a G when we are in a high-GC region?
• What is the probability of emitting a G when we are in a low-GC region?
• Suppose we want to generate an output sequence drawn from this model—how could we do that?
• Suppose we know a sequence of states and we want to compute its probability given the model—how?
• Suppose we have an output sequence and we want to determine its probability based on the model—how?
• Suppose we have a sequence and want to use the model to determine the most likely state at each position—which algorithm would we use for this?
• How about if we want to determine the most likely path of states?
• Suppose we have a sequence and we want to learn the model—how can we do that?

3 Length distributions of states and generalized hidden Markov models

Given a Markov chain with the transition from any state to the end state having probability \( \tau \), the probability of generating a sequence of length \( L \) (and then finishing with a transition to the end state) is given by:

\[
\tau (1 - \tau)^{L-1}
\]

Similarly, in the HMMs that we have been examining, the length of states will be exponentially distributed, which is not appropriate for many purposes. (For example, in a genomic sequence, an exponential distribution does not accurately capture the lengths of genes, exons, introns, etc).

How can we construct a model that does not output state sequences with an exponential distribution of lengths? Suppose we want to make sure that our sequence has length exactly 5? We might construct a sequence of five states with only a single path permitted by transition probabilities. If we include a self loop in one of the states, we will output sequences of minimum length 5, with longer sequences exponentially distributed. Suppose we have a chain of \( n \) states, with all chains starting with state \( a_1 \) and transitioning to an end state after \( a_n \). Also assume that the transition probability between state \( a_i \) and \( a_{i+1} \) is \( 1 - p \), while the self transition probability of state \( a_i \) is \( p \). The probability that a sequence generated by this Markov chain has length \( L \) is given by:

\[
\binom{L-1}{n-1}p^{L-n}(1-p)^n
\]

This is called the negative binomial distribution.

More generally, we can adapt hidden Markov models to produce output sequences of arbitrary length. In a generalized hidden Markov model (also known as a hidden semi-Markov model), the output of each state is a string of symbols, rather than an individual symbol. The length as well as content of this output string can be chosen based on a probability distribution. Many gene finding tools are based on generalized hidden Markov models.

4 Conditional random fields

We didn’t have time to cover conditional random fields in recitation. But if you are curious about them, you might be interested in the following resources:
• Lecture on Conditional Random Fields from Probabilistic Graphical Models course: [https://class.coursera.org/pgm/lecture/preview/33](https://class.coursera.org/pgm/lecture/preview/33) For background, you might also want to watch the two previous segments, on pairwise Markov networks and general Gibbs distributions.


5 References

Introduction to GHMMs: [www.cs.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec07/node28.html](www.cs.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec07/node28.html)

Durbin and Eddy, Biological sequence analysis, chapter 3