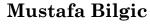
CS 581 – ADVANCED ARTIFICIAL INTELLIGENCE

TOPIC: HIDDEN MARKOV MODELS





♦ http://www.cs.iit.edu/~mbilgic



https://twitter.com/bilgicm

MOTIVATION

- Reason over time/sequence
 - Time series
 - Financial data, sensor readings, temperature, video, location, mic, ...
 - Text
 - DNA
- Some applications
 - Track current state, speech recognition, part-ofspeech tagging, machine translation, handwritten character recognition, ...
 - https://en.wikipedia.org/wiki/Hidden_Markov_model#
 Applications

HIDDEN STATES AND OBSERVATIONS

- Two types of variables
- \circ State variables: X_t
 - The (unobserved) state(s) at time t
- \circ Observation variables: \boldsymbol{o}_t
 - The observed variable(s) at time t
- Examples
 - The text is observed; unobserved states are part-ofspeech for each observed word
 - The GPS sensor readings are observed; unobserved states are the actual locations of the device

Typical Queries

- Filtering
 - $P(X_t | o_{1:t})$
- Prediction
 - $P(X_{t+k} \mid o_{1:t})$ for some k > 0
- Smoothing
 - $P(X_k \mid o_{1:t})$ for some k such that $0 \le k < t$
- Most likely explanation
 - $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t} \mid o_{1:t})$

FACTORIZATION OF THE JOINT $P(X_{0:t}, O_{1:t})$

- $P(X_{0:t}, O_{1:t}) = P(X_{0:t})P(O_{1:t} \mid X_{0:t})$
 - Conditional rule
- $P(X_{0:t}) = P(X_0) P(X_1 \mid X_0) P(X_2 \mid X_{0:1}) \dots P(X_t \mid X_{0:t-1})$
 - Chain rule
- $P(O_{1:t} \mid X_{0:t}) = P(O_1 \mid X_{0:t})P(O_2 \mid O_1, X_{0:t}) \dots P(O_t \mid O_{1:t-1}, X_{0:t})$
 - Conditional chain rule

Markov Assumption – States

- Markov assumption
 - The current state depends on only a finite fixed number of previous states
- First-order Markov assumption
 - The current state depends on only the previous state
- $P(X_{0:t}) = P(X_0) P(X_1 \mid X_0) P(X_2 \mid X_{0:1}) \dots P(X_t \mid X_{0:t-1})$
 - No assumption; just chain rule
- $P(X_{0:t}) = P(X_0) P(X_1 \mid X_0) P(X_2 \mid X_1) \dots P(X_t \mid X_{t-1})$ $= P(X_0) \prod_{i=1}^t P(X_i \mid X_{i-1})$
 - First-order Markov assumption

OBSERVATION MODEL

- The observation at time $t(O_t)$ depends only on the state at time $t(X_t)$
- $P(O_{1:t} \mid X_{0:t}) = P(O_1 \mid X_{0:t}) P(O_2 \mid O_1, X_{0:t}) \dots P(O_t \mid O_{1:t-1}, X_{0:t})$
 - No assumption; just the conditional chain rule
- $P(O_{1:t} \mid X_{0:t}) = P(O_1 \mid X_{0:t}) P(O_2 \mid O_1, X_{0:t}) \dots P(O_t \mid O_{1:t-1}, X_{0:t})$ $= P(O_1 \mid X_1) P(O_2 \mid X_2) \dots P(O_t \mid X_t)$

$$= \prod_{i=1}^t P(O_i \mid X_i)$$

REVISIT THE JOINT

 $P(X_{0:t}, O_{1:t}) = P(X_{0:t})P(O_{1:t} \mid X_{0:t})$

$$= P(X_0) \prod_{i=1}^{t} P(X_i \mid X_{i-1}) P(O_i \mid X_i)$$

• Exercise: draw this model as a Bayesian network

INFERENCE

- Filtering, prediction, and smoothing
 - Probability query
 - Variable elimination and message passing
- Most-likely explanation
 - MAP query
 - Variable elimination and message passing

FILTERING

- Given observations from the beginning to now, what is the probability distribution over the current state?
 - $P(X_t | o_{1:t})$
- Let's use variable elimination. Notice
 - The future variables, both X and O, are mathematically irrelevant. That is, $X_{t+1:\infty}$ and $O_{t+1:\infty}$ are mathematically irrelevant
 - There is a pattern (see OneNote)

PREDICTION

- Given observations from the beginning to now, what is the probability distribution over a state in the future?
 - $P(X_{t+k} \mid o_{1:t})$ for some k > 0
- Let's use variable elimination. Notice
 - All variables after time t+k are mathematically irrelevant. That is, $X_{t+k+1:\infty}$ and $O_{t+k+1:\infty}$ are mathematically irrelevant
 - $O_{t+1:k}$ are also mathematically irrelevant
 - There is a pattern; it's related to the pattern for filtering (see OneNote)

SMOOTHING

- Given observations up to time *t*, what is the probability distribution of a state variable in the past?
 - $P(X_k \mid o_{1:t})$ for some k such that $0 \le k < t$
- Let's use variable elimination. Notice
 - All variables between time θ and t are mathematically relevant. Variables after time t are mathematically irrelevant
 - There is a pattern (see OneNote)

MOST-LIKELY EXPLANATION

- Given observations up to time t, what is the most-likely sequence of states $x_{1:t}$ that could have generated the observation sequence $o_{1:t}$?
 - $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t} \mid o_{1:t})$
- Let's use variable elimination. Instead of sumproduct, we will use max-product
- The pattern is like the patterns in previous slides
- Viterbi algorithm