Probability for Computer Science

Spring 2021

Lecture 24

Prof. Claire Monteleoni

Today

- Parameter Estimation
- Intro. to Markov Chains

With credit to T. Jaakkola and J. Wortman Vaughan.

Parameter Estimation

We have previously introduced Maximum Likelihood and MAP for Hypothesis Testing

Now we'll study Parameter Estimation and look at Maximum Likelihood and MAP techniques for that problem.

Parameter Estimation

- Suppose we would like to estimate the unknown bias, p, of a coin, based on observations of the outcomes of n independent tosses of the coin
- Or suppose we want to estimate the approval rating of the President, by randomly polling n people and asking if they approve or disapprove.

- We can define analogs of both Maximum Likelihood and MAP for this problem of parameter estimation

Parameter Estimation

Consider n observations, X_i, of outcomes of a random variable that is parameterized by some unknown θ.

- Suppose the observations are: $X_1=k_1$, $X_2=k_2$, …, $X_n=k_n$.

The maximum likelihood (ML) estimate is the parameter value that makes the data most likely.

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

"Parameterized by θ"

θ: the parameters of the probabilistic model. We don't assume θ to be random.

Parameter Estimation

The maximum likelihood (ML) estimate is the parameter value that makes the data most likely.

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

If the X_i are independent observations, then:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

"likelihood"

Parameter Estimation

If the X_i are independent observations, then:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

$$= \arg\max_{\theta} \log \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log P(X_i = k_i; \theta)$$

"log-likelihood"

Maximum Likelihood is Consistent

Consistency: if θ is the true value of the parameter, and θ_n is the maximum likelihood estimate after n observations, then for any ε > 0,

$$\lim_{n \to \infty} P(|\theta_n - \theta| \geq \epsilon) = 0$$

In other words: as the number of observations grows large, the maximum likelihood estimate gets closer and closer to the true parameter value, as desired.

Exercise

What is the Maximum Likelihood Estimate (MLE) of the unknown bias, p, of a coin, based on observations of the outcomes of n independent tosses of the coin, X_1, …, X_n?

MAP Parameter Estimation

What if we do want to assume θ is a random variable(s)?

If there is a prior probability distribution over values of the parameter θ, one can derive a maximum a posteriori (MAP) estimate for the parameter.

Useful in cases where n is small.

Markov Chains

- We have seen how Naïve Bayes exploits independence assumptions.
- More generally, Bayesian Networks exploit independence assumptions in order to reason efficiently about complex scenarios.

- Markov chains also exploit independence assumptions, when reasoning about temporal (or sequential) scenarios.
- Markov chains underly Hidden Markov Models (HMMs).
- Markov chains can model sequential problems such as:
  - Sequence of stock prices over time
  - Sequence of pages a Web-surfer visits
  - Sequence of words in a document

Example: Topic Modeling: Approach 1

- Each document may involve many topics. We could allow each word in a document to have a different (independently chosen) topic

$$\begin{array}{cccc} z_1, & z_2, & \ldots, & z_N \\ w_1, & w_2, & \ldots, & w_N \end{array}$$

- As a result, each word is generated from a mixture

$$P(w;\theta) = \sum_{z=1}^{k} \theta_z \, \theta_{w|z}$$

and the probability of the words in a document is given by

$$P(d;\theta) = \prod_{w \in d} P(w;\theta) = \prod_{w \in d} \sum_{z=1}^{k} \theta_z \, \theta_{w|z}$$

θ_z = topic usage (per document)
$\theta_{w|z}$ = topic definitions (same across documents)

Approach 2: A simple sequence model

- We modeled a document as a sequence of topics and the corresponding words

$$\begin{array}{cccc} z_1, & z_2, & \ldots, & z_N \\ w_1, & w_2, & \ldots, & w_N \end{array}$$

- We made strong assumptions about how they are generated

$$P(z_1, \ldots, z_N, w_1, \ldots, w_N) = P(w_1, \ldots, w_N | z_1, \ldots, z_N) P(z_1, \ldots, z_N)$$

$$\stackrel{(1)}{=} \left[\prod_{i=1}^{N} P(w_i | z_i) \right] P(z_1, \ldots, z_N)$$

$$\stackrel{(2)}{=} \left[\prod_{i=1}^{N} P(w_i | z_i) \right] \left[\prod_{i=1}^{N} P(z_i) \right]$$

1) words are independent given topics, depend only on the current topic, 2) topics are independent along the sequence

Approach 3: Markov Chain

- We could extend the simple independent topic model by "coupling" the choice of topics along the sequence

$$z_1, \quad z_2, \quad \ldots, \quad z_N$$
$$w_1, \quad w_2, \quad \ldots, \quad w_N$$

- For example, we could relate successive topic choices by assuming that they are governed by a Markov chain

$$P(z_1, \ldots, z_N) = P(z_1)P(z_2|z_1)P(z_3|z_1, z_2) \cdots P(z_N|z_{N-1}, \ldots, z_1)$$
$$\overset{(2)}{=} P(z_1)P(z_2|z_1)P(z_3|z_2) \cdots P(z_N|z_{N-1})$$

where in 2) we assume that each topic selection can depend on the preceding selection but not further back.

We model the sequence of topics (or hidden states in the HMM) using a Markov Chain.

(First-order) Markov Assumption

First-order Markov Assumption:

The next state is conditionally independent of all past states, given the current state.

$$\forall \tau < t \ : \ P(s_{t+1}|s_t, s_\tau) = P(s_{t+1}|s_t)$$

Markov Chains

A Markov chain is specified by:

- The state space, the set of possible states, S = {1, …, m}

- The transitions dynamics:
  - For each pair of states, (i, j) in S x S, for which a transition from state i to state j is possible, a transition probability,

    $p_{ij} > 0$.

The Markov chain is then a sequence of r.v.s X_0, X_1, X_2,… that take values in S, such that for all times n, all states i,j in S, and all possible sequences of earlier states, i_0,…,i_{n-1} , the following holds:

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p_{ij}$$

Markov Chains

- The Markov chain is homogeneous if the "transition probabilities" $P(z_i|z_{i-1})$ do not depend on the position i along the sequence.

  In this case, we can parameterize the chain

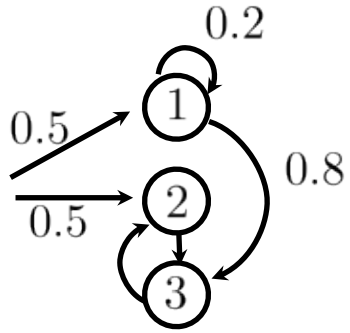
$$P(z_1, \ldots, z_N; \theta) = q(z_1) \prod_{i=2}^{N} q(z_i|z_{i-1})$$

  by specifying the initial "state" (here topic) distribution $q(z_1)$ and the state transition probability matrix $q(z_i|z_{i-1})$ that is reused along the sequence

Markov Chains: representation

- A state transition diagram: nodes correspond to "states" (here topics), arcs represent possible transitions between states

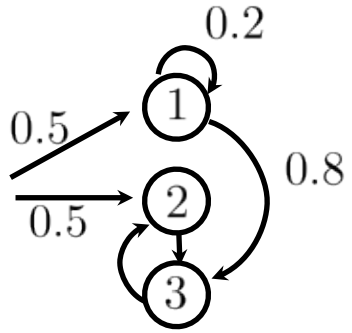
$$z_i = 1, \ldots, 3$$



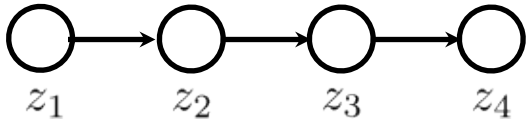
Markov Chains: representation

- A state transition diagram: nodes correspond to "states" (here topics), arcs represent possible transitions between states

$$z_i = 1, \ldots, 3$$



- A graphical model: nodes in the graph correspond to variables (state/topic selections) and arcs represent how the variables depend on each other



Markov Chains

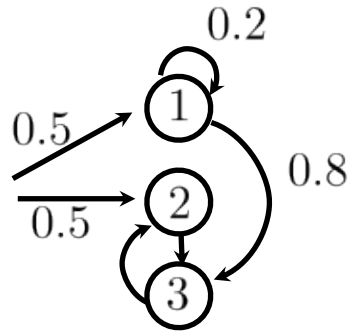
- Given any initial state distribution $q(z_1)$ and the state transition probability matrix $q(z_i|z_{i-1})$, we can generate sample sequences from the corresponding homogeneous Markov chain according to

$$z_1 \sim q(z_1),$$

$$z_i \sim q(z_i|z_{i-1}), \quad \text{where } z_{i-1} \text{ is fixed}, \quad i = 2, \ldots, N$$

(Note: we do not model the length of the sequence)

$$z_i = 1, \ldots, 3$$



Example

Define the state space as S = {1, …, m}, corresponding to positions on the line.

A fly starts at one of these positions and then moves left with probability 0.3, right with probability 0.3, and stays in place with probability 0.4.

There is a spider at position 1 and at position m. If a fly lands at either of these two positions it dies.

- What is the transition matrix of this Markov chain?

- What is the transition state transition diagram?

n-Step Transitions

- We can compute the probability of any sequence of states using multiplication rule and Markov property.

- We can efficiently compute the the n-step transition probability, using the Total Probability Theorem, yielding a recursive formula:

$$P(X_n = j | X_0 = i) = \sum_{k=1}^{m} p_{kj} P(X_{n-1} = k | X_0 = i)$$

n-Step Transitions

We can efficiently compute the the n-step transition probability, using the Total Probability Theorem, yielding a recursive formula:

$$P(X_n = j|X_0 = i) = \sum_{k=1}^{m} p_{kj} P(X_{n-1} = k|X_0 = i)$$

$$r_{ij}(n) = P(X_n = j|X_0 = i) = \sum_{k=1}^{m} p_{kj} r_{ik}(n-1)$$

I.e. the n-step transition probability matrix can be obtained by multiplying the (n-1)-step transition probability matrix by the (one-step) transition probability matrix.

→ Therefore, the n-step transition probability matrix is n-th power of the (one-step) transition matrix!

n-Step Transitions

Example. Consider an unreliable router which can either be online or offline. If it is online one day, it will be online the next day with probability 0.8. If it is offline one day, it will remain offline the next day with probability 0.4.

- If the router is online on Day 0, what is the probability it will be online on Day 1? Day 2?
- Similarly, re-compute the n-step transition probabilities for Day 1 and Day 2, starting from the router being offline at day 0.
- (Optional) If time, compute the the n-step transition probabilities, for each of the two cases above, for Day 3, and Day 4.

n-Step Transitions

Example. Consider an unreliable router which can either be online or offline. If it is online one day, it will be online the next day with probability 0.8. If it is offline one day, it will remain offline the next day with probability 0.4.

If the router is online on Day 0, what is the probability it will be online on Day 1? Day 2? Day 3? Day 4?

Similarly you can compute the n-step transition probabilities starting from $X_0 = 2$.

Bonus slides

n-Step Transitions

Example. Consider an unreliable router which can either be online or offline. If it is online one day, it will be online the next day with probability 0.8. If it is offline one day, it will remain offline the next day with probability 0.4.

If the router is online on Day 0, what is the probability it will be online on Day 1? Day 2? Day 3? Day 4?

$P(X_0 = 1 \mid X_0 = 1) = 1$	$P(X_0 = 2 \mid X_0 = 1) = 0$
$P(X_1 = 1 \mid X_0 = 1) = 0.8$	$P(X_1 = 2 \mid X_0 = 1) = 0.2$
$P(X_2 = 1 \mid X_0 = 1) = 0.76$	$P(X_2 = 2 \mid X_0 = 1) = 0.24$
$P(X_3 = 1 \mid X_0 = 1) = 0.752$	$P(X_3 = 2 \mid X_0 = 1) = 0.248$
$P(X_4 = 1 \mid X_0 = 1) = 0.7504$	$P(X_4 = 2 \mid X_0 = 1) = 0.2496$

Similarly you can compute the n-step transition probabilities starting from $X_0 = 2$.

Classification of States

- Definition: a state *j* is accessible from state *i* if there exists some n ≥ 0 such that the n-step probability $r_{ij}(n)$ is non-zero.

$$j \in A(i)$$

$$\Longleftrightarrow$$

$$\exists n \geq 0 : \; r_{ij}(n) > 0$$

i.e. there's a positive probability of reaching state *j* from state *i*, after some amount of time.

Classification of States

- Definition: Let A(i) be the set of states that are accessible from state *i*. State *i* is called recurrent if for every *j* in A(i), *i* is also accessible from *j*.
  - i.e. for each state *j,* that is accessible from *i, i* is also accessible from *j.*

$$j \in A(i) \implies i \in A(j)$$

- Definition: Any state that is not recurrent is transient.

Classification of States

- Recurrent: a state *i* is recurrent if:

$$j \in A(i) \implies i \in A(j)$$

- Transient: a state *i* is transient if there exists some state *j* such that:

$$j \in A(i), \quad i \notin A(j)$$

Recurrent Class

Definition: If *i* is a recurrent state, the set of states, A(i), that are accessible from *i* form a recurrent class, meaning that states in A(i) are all accessible from eachother, and no state outside of A(i) is accessible from them.

i.e. if *i* is a recurrent state, then:

$$\forall j \in A(i), \quad A(i) = A(j)$$

Proof sketch:

- By definition of accessible, if j is accessible from i, then i is accessible from j. So from j, all of A(i) is accesssible (e.g. passing through i first). Therefore A(i) is a subset of A(j).
- Meanwhile, for any k that is not accessible from i, k cannot be accessible from j otherwise it would be accessible from i, via j, which is a contradiction. So A(j) is a subset of A(i).

Long-term behavior of MCs

- By definition of a recurrent class, if a Markov chain starts in a recurrent state, it will stay in that recurrent class forever, and never visit a transient state.
  - All states in that recurrent class are accessible from eachother so they will all be visited an infinite number of times.

- If a Markov chain starts in a transient state, it will pass through 0 or more transient states before ending up in a recurrent class.
  - Once this happens the state will stay in that recurrent class forever.

Markov chain decomposition

- A Markov chain can be decomposed into one or more recurrent classes, plus possibly some transient states.
- A recurrent state is accessible from all states in its class but is not accessible from recurrent states in other classes.
- A transient state is not accessible from any recurrent state.
- One or more recurrent state(s) are accessible from any transient state.

Periodic recurrent class

- Definition: A recurrent class is periodic if its states can be grouped into d > 1 disjoint subsets S_1, …, S_d, such that all transitions from one subset lead to the next subset.
  - i.e. all transitions from S_k lead to S_{k+1} (or to S_1 if k=d).
  - The period of a recurrent class, is the number, d, of these subsets.
  - Formally: if $i \in S_k$ and $p_{ij} > 0$, then $\begin{cases} j \in S_{k+1} & k < d \\ j \in S_1 & k = d \end{cases}$

- A recurrent class is aperiodic if it is not periodic.
  - i.e., there exists some time n, such that:

$$r_{ij}(n) > 0 \quad \text{for all states } i, j \quad \text{in the class}$$

Steady-state probabilities

Consider a Markov Chain with the following properties:

- It has a single recurrent class
- This recurrent class is aperiodic

Then every state *j* has a steady-state probability.

For large *n*: $\pi_j \approx P(X_n = j)$

That is, the probability, $r_{ij}(n)$, of being in state *j* at time *n* is independent of the initial state *i*.

Steady-State Convergence Theorem

Theorem: For any Markov chain with a single, aperiodic recurrent class, there are unique values $\pi_1, \ldots, \pi_m$ that solve the balance equations:

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{kj} \quad \text{for } j = 1, \ldots, m$$

$$\sum_{k=1}^{m} \pi_k = 1$$

where for each j, π_j is its steady-state probability.