CPSC 440: Advanced Machine Learning

Markov Chains

Mark Schmidt

University of British Columbia

Winter 2022

Example: Vancouver Rain Data

- Consider density estimation on the "Vancouver Rain" dataset:

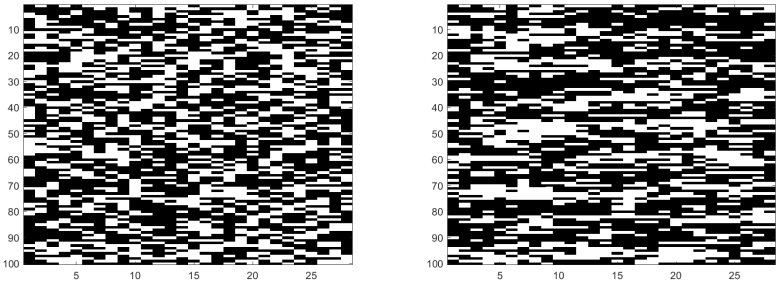
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	...
Month 1	0	0	0	1	1	0	0	1	1	
Month 2	1	0	0	0	0	0	1	0	0	
Month 3	1	1	1	1	1	1	1	1	1	
Month 4	1	1	1	1	0	0	1	1	1	
Month 5	0	0	0	0	1	1	0	0	0	
Month 6	0	1	1	0	0	0	0	1	1	

- Variable $x_j^i = 1$ if it rained on day j in month i.
  - Each row is a month, each column is a day of the month.
  - Data ranges from 1896-2004.

- The strongest signals in the data:
  - It tends to rain more in the winter than the summer.
  - If it rained yesterday, it's likely to rain today ($\approx 70\%$ chance of $(x_j^i == x_{j-1}^i)$).

Rain Data with Product of Bernoullis

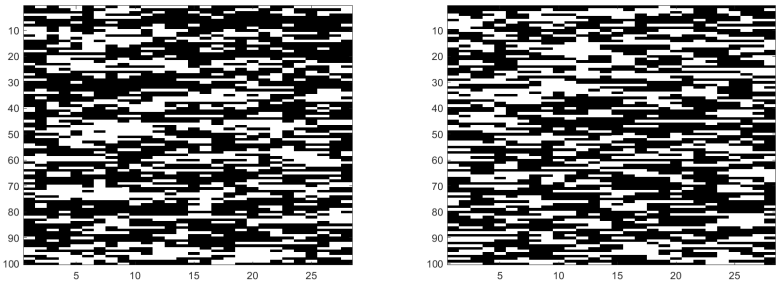
- With product of Bernoullis, we get $p(x_j^i = \text{“rain”}) \approx 0.41$ (sadly).
  - Samples from product of Bernoullis model (left) vs. real data (right):



- Making days independent misses seasons and misses correlations.

Markov Chains

- A better model for the between-day correlations is a Markov chain.
  - Models $p(x_j^i \mid x_{j-1}^i)$: probability of rain today given yesterday's value.
    - Captures dependency between adjacent days.



- It can perfectly capture the "position-independent" between-day correlation.
  - With only a few parameters and a closed-form MLE.

Markov Chain for Rain

- Markov chain ingredients and MLE for rain data:
  - State space:
    - At time j, we can be in the "rain" state or the "not rain" state.
  - Initial probabilities:

c	$p(x_1 = c)$
Rain	0.37
Not Rain	0.63

  - Transition probabilities (assumed to the same for all times j):

c'	c	$p(x_j = c \mid x_{j-1} = c')$
Rain	Rain	0.65
Rain	Not Rain	0.35
Not Rain	Rain	0.25
Not Rain	Not Rain	0.75

- Becuase of "sum to 1" constraints, there are only 3 parameters in this model.

Markov Chain Ingredients

- Markov chain ingredients and MLE for rain data:
  - State space:
    - Set of possible states (indexed by c) we can be in at time j ("rain" or "not rain").
  - Initial probabilities:
    - $p(x_1 = c)$: probability that we start in state c at time $j = 1$ (p("rain") on day 1).
  - Transition probabilities:
    - $p(x_j = c \mid x_{j-1} = c')$: probability that we move from state c' to state c at time j.
    - Probability that it rains today, given what happened yesterday.

- We're assuming that the order of features is meaningful.
  - We're modeling dependency of each feature on the previous feature.

Chain Rule of Probability

- By using the product rule, $p(a, b) = p(a)p(b \mid a)$, we can write any density as

$$\begin{aligned} p(x_1, x_2, \ldots, x_d) &= p(x_1)p(x_2, x_3, \ldots, x_d \mid x_1) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3, x_4, \ldots, x_d \mid x_1, x_2) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1)p(x_4, x_5, \ldots, x_d \mid x_1, x_2, x_3), \end{aligned}$$

  and so on until we get

$$p(x_1, x_2, \ldots, x_d) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_d \mid x_1, x_2, \ldots x_{d-1}).$$

- This factorization of a density is called the chain rule of probability.
  - It turns multivariate density estimation into estimating conditionals.

- But it leads to complicated conditionals:
  - For binary x_j, we need 2^d parameters for $p(x_d \mid x_1, x_2, \ldots, x_{d-1})$ alone.
  - Or we could logistic regression, neural networks, and so on to estimate conditionals.

Markov Chains

- Markov chains we simplify the distribution by assuming the Markov property:

$$p(x_j \mid x_{j-1}, x_{j-2}, \ldots, x_1) = p(x_j \mid x_{j-1}),$$

  that x_j is independent of the past given x_{j-1}.
  - "Don't care what happened 2 days ago if you know what happened yesterday".

- The probability for a sequence $x_1, x_2, \cdots, x_d$ in a Markov chain simplifies to

$$\begin{aligned} p(x_1, x_2, \ldots, x_d) &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1)\cdots p(x_d \mid x_{d-1}, x_{d-2}, \ldots, x_1) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)\cdots p(x_d \mid x_{d-1}) \end{aligned}$$

- Another way to write this joint probability is

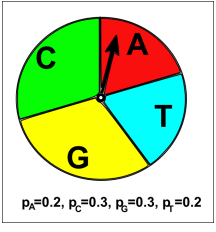
$$p(x_1, x_2, \ldots, x_d) = \underbrace{p(x_1)}_{\text{initial prob.}} \prod_{j=2}^{d} \underbrace{p(x_j \mid x_{j-1})}_{\text{transition prob.}}.$$

Example: Modeling DNA Sequences

- A nice demo of independent vs. Markov for DNA sequences:
  - http://a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/src/chapter10.html

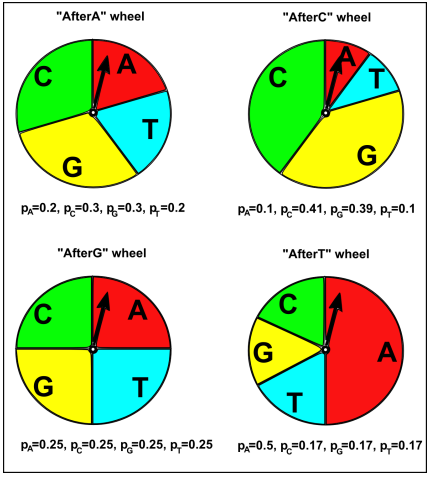
https://www.tes.com/lessons/WE5E9RncBhieAQ/dna

- Independent model for elements of sequence:



Example: Modeling DNA Sequences

- Transition probabilities in a Markov chain model for elements of sequence:



(visualizing transition probabilities based on previous symbol):

Markov Chains

- Markov chains are ubiquitous in sequence/time-series models:

Summary

- Markov chains model dependencies between adjacent features.
  - Set of possible states.
  - Initial probabilities.
  - Transition probabilities.
- Chain rule of probability.
  - Writes joint probability in terms of conditionals over "earlier" variables.
- Markov assumption.
  - Conditional independence from "past" times given previous time.

- Next time: Monte Carlo for Markov chains (MC for MC, not MCMC).