Inference and Representation

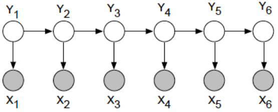
Rachel Hodos

New York University

Lab 3, September 16, 2015

Outline

Hidden Markov Models



- Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1 \mid y_1) \prod_{t=2}^{T} p(y_t \mid y_{t-1})p(x_t \mid y_t)$$

- A **homogeneous** HMM uses the same parameters (β and α below) for each transition and emission distribution (**parameter sharing**):

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)\alpha_{x_1,y_1} \prod_{t=2}^{T} \beta_{y_t,y_{t-1}}\alpha_{x_t,y_t}$$

How many parameters need to be learned?

An unexpected lesson on feature correlation

- Linear regression using pymc3
- Model: $Y \sim \mathcal{N}(\mu, \sigma^2), \mu = \alpha + \beta_1 X_1 + \beta_2 X_2$
- Parameters: $\alpha = 1, \sigma = 1, \underline{\beta_1 = 1, \beta_2 = 2.5}$
- Estimated coefficients:

```
{'alpha': array(1.0136638069892534), 'beta': array([ 1.46791629,
0.29358326]), 'sigma_log': array(0.11928770010017063)}
```

- We generated data using the following lines of code:

X1 = np.linspace(0, 1, n); X2 = np.linspace(0, 0.2, n)

- What is the correlation between X1 and X2?
- Why would this cause a problem?

Applying probabilistic modeling in the real world

- Some questions addressed yesterday:
  - Can I quantify your political stance based on who you follow?
  - What general topics are being discussed on Twitter?
    - How does this change over time?
    - Who is talking about what?
  - How much dialogue occurs on social media between people with different ideologies?
  - Are representatives of Congress affected by what their followers are discussing?
  - How can we interpret neuronal spiking patterns in the brain?
  - What makes neurons spike together?

Miscellaneous comments

- Clarification: inverse logit = logistic function =

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

- Both speakers used hidden variables, the state of the hidden variables told them something interesting
- Naive Bayes

Modeling latent political ideologies

Spatial following model

- Users' and politicians' ideology (θ_i and ϕ_j) are defined as latent variables to be estimated.
- Data: "following" decisions, a matrix of binary choices ($\mathbf{Y_{ij}}$).
- Spatial following model: for *n* users, indexed by *i*, and *m* political accounts, indexed by *j*:

$$P(y_{ij} = 1 | \alpha_j, \beta_i, \gamma, \theta_i, \phi_j) = \text{logit}^{-1}\left(\alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2\right)$$

where:

α_j measures *popularity* of politician *j*

β_i measures *political interest* of user *i*

γ is a normalizing constant

Modeling latent political ideologies

Estimation

- Goal of learning:
  - θ_i: ideological positions of users $i = 1, \ldots, n$
  - ϕ_j: ideological positions of political accounts $j = 1, \ldots, m$
- Likelihood function:

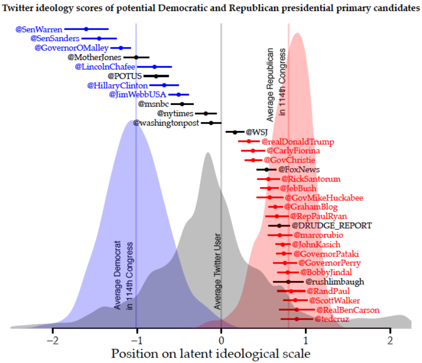
$$p(\mathbf{y}|\theta, \phi, \alpha, \beta, \gamma) = \prod_{i=1}^{n} \prod_{j=1}^{m} \text{logit}^{-1}(\pi_{ij})^{y_{ij}} (1 - \text{logit}^{-1}(\pi_{ij}))^{1-y_{ij}}$$

$$\text{where } \pi_{ij} = \alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2$$

- Exact inference is intractable $\rightarrow$ MCMC (approx. inference)
- Estimation:
  - First stage: HMC in *Stan* with random sample of **Y** to compute posterior distribution of j-indexed parameters.
  - Second stage: parallelized MH in *R* for rest of i-indexed parameters (assuming independence), on NYU's HPC.

Modeling latent political ideologies

Application: Ideology of Presidential Candidates

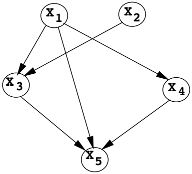


Barberá "Who is the most conservative Republican candidate for president?" *The Washington Post*, June 16 2015

Outline

Parameter Learning

- Given:
  - A Bayesian network structure.



  - A data set

X_1	X_2	X_3	X_4	X_5
0	0	1	1	0
1	0	0	1	0
0	1	0	0	1
0	0	1	1	1
⋮	⋮	⋮	⋮	⋮

- Estimate conditional probabilities:

$$P(X_1), P(X_2), P(X_3|X_1, X_2), P(X_4|X_1), P(X_5|X_1, X_3, X_4)$$

Outline

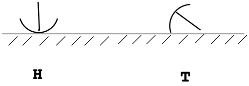
1 Problem Statement

2 Principles of Parameter Learning
- Maximum likelihood estimation
- Bayesian estimation
- Variable with Multiple Values

3 Parameter Estimation in General Bayesian Networks
- The Parameters
- Maximum likelihood estimation
- Properties of MLE
- Bayesian estimation

Single-Node Bayesian Network

X: result of tossing a thumbtack



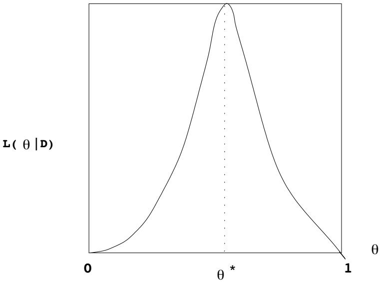
- Consider a Bayesian network with one node X, where X is the result of tossing a thumbtack and $\Omega_X = \{H, T\}$.
- Data cases:
  $D_1 = H$, $D_2 = T$, $D_3 = H$, ..., $D_m = H$
- Data set: $\mathbf{D} = \{D_1, D_2, D_3, \ldots, D_m\}$
- Estimate parameter: $\theta = P(X{=}H)$.

Likelihood

- Data: $\mathbf{D} = \{H, T, H, T, T, H, T\}$
- As possible values of θ, which of the following is the most likely? Why?
  - $\theta = 0$
  - $\theta = 0.01$
  - $\theta = 10.5$

- $\theta = 0$ contradicts data because $P(\mathbf{D}|\theta = 0) = 0$.It cannot explain the data at all.
- $\theta = 0.01$ almost contradicts with the data. It does not explain the data well. However, it is more consistent with the data than $\theta = 0$ because $P(\mathbf{D}|\theta = 0.01) > P(\mathbf{D}|\theta = 0)$.
- So $\theta = 0.5$ is more consistent with the data than $\theta = 0.01$ because $P(\mathbf{D}|\theta = 0.5) > P(\mathbf{D}|\theta = 0.01)$
  It explains the data the best among the three and is hence the most likely.

Maximum Likelihood Estimation



- In general, the larger $P(\mathbf{D}|\theta = v)$ is, the more likely $\theta = v$ is.
- Likelihood of parameter θ given data set:

$$L(\theta|\mathbf{D}) = P(\mathbf{D}|\theta)$$

- The **maximum likelihood estimation (MLE)** θ^* of θ is a possible value of θ such that

$$L(\theta^*|\mathbf{D}) = sup_\theta L(\theta|\mathbf{D}).$$

MLE best explains data or best fits data.

i.i.d and Likelihood

- Assume the data cases $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data cases are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$\begin{aligned} L(\theta|\mathbf{D}) &= P(\mathbf{D}|\theta) = P(D_1, \ldots, D_m|\theta) \\ &= \prod_{i=1}^{m} P(D_i|\theta) = \theta^{m_h}(1-\theta)^{m_t} \end{aligned} \tag{1}$$

where m_h is the number of heads and m_t is the number of tail. **Binomial likelihood**.

Example of Likelihood Function

- Example: $\mathbf{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

$$
\begin{aligned}
L(\theta|\mathbf{D}) &= P(\mathbf{D}|\theta) \\
&= P(D_1 = H|\theta)P(D_2 = T|\theta)P(D_3 = H|\theta)P(D_4 = H|\theta)P(D_5 = T|\theta) \\
&= \theta(1-\theta)\theta\theta(1-\theta) \\
&= \theta^3(1-\theta)^2.
\end{aligned}
$$

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathbf{D}) = logL(\theta|\mathbf{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- By Corollary 1.1 of Lecture 1, the following value maximizes $l(\theta|\mathbf{D})$:

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

- MLE is intuitive.
- It also has nice properties:
  - E.g. **Consistence**: θ^* approaches the true value of θ with probability 1 as m goes to infinity.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence θ=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30,000, 70,000)$. MLE: $\theta = 0.3$.
    - Reasonable.
    - Data suggest that the coin is after all biased, overshadowing our prior.
  - MLE does not differentiate between those two cases. It doe not take prior information into account.

Outline

The Parameters

- n variables: X_1, X_2, ..., X_n.
- Number of states of X_i: 1, 2, ..., $r_i{=}|\Omega_{X_i}|$.
- Number of configurations of parents of X_i : 1, 2, ..., $q_i{=}|\Omega_{pa(X_i)}|$.
- Parameters to be estimated:

$$\theta_{ijk} = P(X_i = j|pa(X_i) = k), \qquad i = 1, \ldots, n; j = 1, \ldots, r_i; k = 1, \ldots, q_i$$

- Parameter vector: $\theta = \{\theta_{ijk}|i = 1, \ldots, n; j = 1, \ldots, r_i; k = 1, \ldots, q_i\}$.
  Note that $\sum_j \theta_{ijk} = 1 \forall i, k$
- $\theta_{i..}$: Vector of parameters for $P(X_i|pa(X_i))$

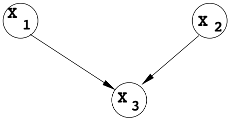
$$\theta_{i..} = \{\theta_{ijk}|j = 1, \ldots, r_i; k = 1, \ldots, q_i\}$$

- $\theta_{i.k}$: Vector of parameters for $P(X_i|pa(X_i){=}k)$

$$\theta_{i.k} = \{\theta_{ijk}|j = 1, \ldots, r_i\}$$

The Parameters

- Example: Consider the Bayesian network shown below. Assume all variables are binary, taking values 1 and 2.



$$
\begin{aligned}
\theta_{111} &= P(X_1{=}1), \theta_{121} = P(X_1{=}2) \\
\theta_{211} &= P(X_2{=}1), \theta_{221} = P(X_2{=}2) \\
pa(X_3) = 1 : \theta_{311} &= P(X_3{=}1|X_1 = 1, X_2 = 1), \theta_{321} = P(X_3{=}2|X_1 = 1, X_2 = 1) \\
pa(X_3) = 2 : \theta_{312} &= P(X_3{=}1|X_1 = 1, X_2 = 2), \theta_{322} = P(X_3{=}2|X_1 = 1, X_2 = 2) \\
pa(X_3) = 3 : \theta_{313} &= P(X_3{=}1|X_1 = 2, X_2 = 1), \theta_{323} = P(X_3{=}2|X_1 = 2, X_2 = 1) \\
pa(X_3) = 4 : \theta_{314} &= P(X_3{=}1|X_1 = 2, X_2 = 2), \theta_{324} = P(X_3{=}2|X_1 = 2, X_2 = 2)
\end{aligned}
$$

Data

- A complete case D_l: a vector of values, one for each variable.
- Example: $D_l = (X_1 = 1, X_2 = 2, X_3 = 2)$
- Given: A set of complete cases: $\mathbf{D} = \{D_1, D_2, \ldots, D_m\}$.
- Example:

X_1	X_2	X_3	X_1	X_2	X_3
1	1	1	2	1	1
1	1	2	2	1	2
1	1	2	2	2	1
1	2	2	2	2	1
1	2	2	2	2	2
1	2	2	2	2	2
2	1	1	2	2	2
2	1	1	2	2	2

- Find: The ML estimates of the parameters θ.

The Loglikelihood Function

- Loglikelihood:

$$l(\theta|D) = logL(\theta|D) = logP(D|\theta) = log \prod_l P(D_l|\theta) = \sum_l logP(D_l|\theta).$$

- The term $logP(D_l|\theta)$:
  - $D_4 = (1, 2, 2)$,

$$\begin{aligned} logP(D_4|\theta) &= logP(X_1 = 1, X_2 = 2, X_3 = 2) \\ &= logP(X_1{=}1|\theta)P(X_2{=}2|\theta)P(X_3{=}2|X_1{=}1, X_2{=}2, \theta) \\ &= log\theta_{111} + log\theta_{221} + log\theta_{322}. \end{aligned}$$

Recall:
$\theta = \{\theta_{111}, \theta_{121}; \theta_{211}, \theta_{221}; \theta_{311}, \theta_{312}, \theta_{313}, \theta_{314}, \theta_{321}, \theta_{322}, \theta_{323}, \theta_{324}\}$

The Loglikelihood Function

- Define the **characteristic function** of case D_l:

$$\chi(i, j, k : D_l) = \begin{cases} 1 & \text{if } X_i = j,\ pa(X_i) = k \text{ in } D_l \\ 0 & \text{otherwise} \end{cases}$$

- When l=4, $D_4 = (1, 2, 2)$.

$$\chi(1, 1, 1 : D_4) = \chi(2, 2, 1 : D_4) = \chi(3, 2, 2 : D_4) = 1$$

$$\chi(i, j, k : D_4) = 0 \text{ for all other i, j, k}$$

- So, $logP(D_4|\theta) = \sum_{ijk} \chi(i, j, k; D_4) log\theta_{ijk}$
- In general,

$$logP(D_l|\theta) = \sum_{ijk} \chi(i, j, k : D_l) log\theta_{ijk}$$

The Loglikelihood Function

- Define

$$m_{ijk} = \sum_{l} \chi(i, j, k : D_l).$$

It is the number of data cases where $X_i = j$ and $pa(X_i) = k$.

- Then

$$\begin{aligned}
l(\theta|\mathbf{D}) &= \sum_{l} log P(D_l|\theta) \\
&= \sum_{l} \sum_{i,j,k} \chi(i, j, k : D_l) log \theta_{ijk} \\
&= \sum_{i,j,k} \sum_{l} \chi(i, j, k : D_l) log \theta_{ijk} \\
&= \sum_{ijk} m_{ijk} log \theta_{ijk} \\
&= \sum_{i,k} \sum_{j} m_{ijk} log \theta_{ijk}. \qquad (4)
\end{aligned}$$

MLE

- Want:

$$\arg\max_{\theta} l(\theta|\mathbf{D}) = \arg\max_{\theta_{ijk}} \sum_{i,k} \sum_{j} m_{ijk} log\theta_{ijk}$$

- Note that $\theta_{ijk} = P(X_i{=}j|pa(X_i){=}k)$ and $\theta_{i'j'k'} = P(X_{i'}{=}j'|pa(X_{i'}){=}k')$ are not related if either $i{\neq}i'$ or $k{\neq}k'$.
- Consequently, we can separately maximize each term in the summation $\sum_{i,k}[\ldots]$

$$\arg\max_{\theta_{ijk}} \sum_{j} m_{ijk} log\theta_{ijk}$$

MLE

- By Corollary 1.1 , we get

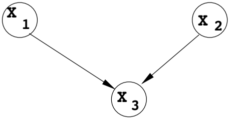
$$\theta^*_{ijk} = \frac{m_{ijk}}{\sum_j m_{ijk}}$$

- In words, the MLE estimate for $\theta_{ijk} = P(X_i{=}j|pa(X_i){=}k)$ is:

$$\theta^*_{ijk} = \frac{\text{number of cases where } X_i{=}j \text{ and } pa(X_i){=}k}{\text{number of cases where } pa(X_i){=}k}$$

Example

Example:



X_1	X_2	X_3	X_1	X_2	X_3
1	1	1	2	1	1
1	1	2	2	1	2
1	1	2	2	2	1
1	2	2	2	2	1
1	2	2	2	2	2
1	2	2	2	2	2
2	1	1	2	2	2
2	1	1	2	2	2

- MLE for $P(X_1{=}1)$ is: $6/16$
- MLE for $P(X_2{=}1)$ is: $7/16$
- MLE for $P(X_3{=}1|X_1{=}2, X_2{=}2)$ is: $2/6$
- ...

Kullback-Leibler divergence

- **Relative entropy** or **Kullback-Leibler divergence**
  - Measures how much a distribution $Q(X)$ differs from a "true" probability distribution $P(X)$.
  - **K-L divergence** of Q from P is defined as follows:

$$KL(P, Q) = \sum_X P(X) log \frac{P(X)}{Q(X)} = E_P[logP(X)] - E_P[logQ(X)]$$

$0 log \frac{0}{0} = 0$ and $plog \frac{p}{0} = \infty$ if $p \neq 0$

- Not symmetric. So, not a distance measure mathematically.

Kullback-Leibler divergence

Theorem (1.2)

*(**Gibbs' inequality**)*

$$KL(P, Q) \geq 0$$

with equality holds iff P is identical to Q

Proof:

$$\begin{aligned}
\sum_X P(X) log \frac{P(X)}{Q(X)} &= -\sum_X P(X) log \frac{Q(X)}{P(X)} \\
&\geq -log \sum_X P(X) \frac{Q(X)}{P(X)} \qquad \text{Jensen's inequality} \\
&= -log \sum_X Q(X) = 0.
\end{aligned}$$

KL distance from P to Q is larger than 0 unless P and Q are identical.

A corollary

Corollary (1.1)

Let $f(X)$ be a nonnegative function of variable X such that $\sum_X f(X) > 0$. Let $P^*(X)$ be the probability distribution given by

$$P^*(X) = \frac{f(X)}{\sum_X f(X)}.$$

Then for any other probability distribution $P(X)$

$$\sum_X f(X) log P^*(X) \geq \sum_X f(X) log P(X)$$

with equality holds iff P^* and P are identical. In other words,

$$P^* = \arg\sup_P \sum_X f(X) log P(X)$$

A corollary

Proof:

$$KL(P^*, P) = \sum_X P^*(X) log \frac{P^*(X)}{P(X)} \geq 0$$

Hence

$$\sum_X P^*(X) log P^*(X) \geq \sum_X P^*(X) log P(X)$$

$$\sum_X \frac{f(X)}{\sum_X f(X)} log P^*(X) \geq \sum_X \frac{f(X)}{\sum_X f(X)} log P(X)$$

$$\sum_X f(X) log P^*(X) \geq \sum_X f(X) log P(X)$$

Q.E.D