Graphical Models

Lecture 5:

Parameter Estimation & Lagrange Multipliers

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.

Administration

- HW#2 due date

Learning

- Bayesian Networks can be built by hand.
  - Experts' time is expensive.
  - There may not be any experts.
  - Large models are unwieldly.
  - Knowledge doesn't always transfer across domains.
- Data is often cheap (now).
  - Remember that this was not always the case!

Notation

- P^* is the true distribution from which our samples were drawn.
- $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}$ drawn IID from P^*.

Goal of Learning?

- **Density estimation**:
  Return a model M that precisely captures P*
- **Prediction**:
  Optimize quality of answers to specific queries,
  e.g. $P(x_i | x_j, x_k)$
- **Knowledge discovery**:
  Reveal facts about the domain.

Learning Bayesian Networks

	Known structure	Unknown structure
Fully observed data	☺ (today)	hard (later)
Missing data	hard (later)	very hard

MLE Basics

- Likelihood function
- Sufficient statistic: vector representation of the data that summarizes everything you need to compute likelihood
  - If $\tau(\text{dataset}_1) = \tau(\text{dataset}_2)$ then the likelihood functions are the same.
- For distributions over one random variable, this is usually not hard.
- What about Bayesian networks?

Key Idea

- For known structure and fully observed data, MLE for a Bayesian network whose CPDs have disjoint parameters

  equates to

  MLE for each of its CPDs.

- That's it!
- Why?

Decomposability

$$\begin{aligned}
\boldsymbol{\theta}_{\text{MLE}} &= \arg\max_{\theta} \prod_t P(\boldsymbol{X} = \boldsymbol{x}^{(t)} \mid \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \prod_t \prod_i P(X_i = x_i^{(t)} \mid \text{Parents}(X_i) = \text{Parents}(x_i), \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \sum_i \log P(X_i = x_i^{(t)} \mid \text{Parents}(X_i) = \text{Parents}(x_i), \boldsymbol{\theta})
\end{aligned}$$

If the parameters **θ** are partitioned by CPT …

$$= \arg\max_{\boldsymbol{\theta}} \sum_i \sum_t \log P(X_i = x_i^{(t)} \mid \text{Parents}(X_i) = \text{Parents}(x_i), \boldsymbol{\theta}_i)$$

(swap order of sums)

Decomposability

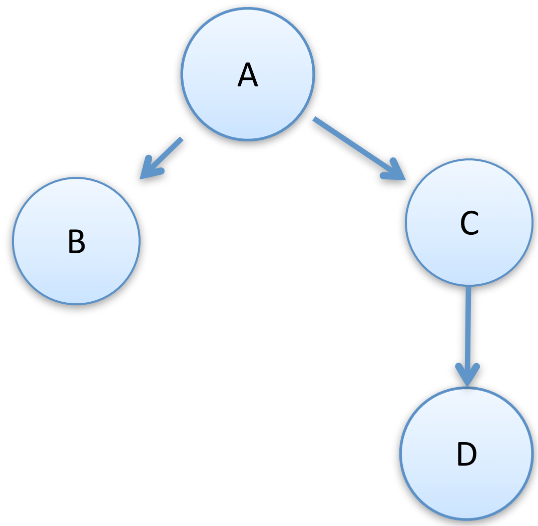
Example

$$\langle a^{(1)}, b^{(1)}, c^{(1)}, d^{(1)} \rangle$$

$$\langle a^{(2)}, b^{(2)}, c^{(2)}, d^{(2)} \rangle$$

$$\vdots$$

$$\langle a^{(M)}, b^{(M)}, c^{(M)}, d^{(M)} \rangle$$



$$\boldsymbol{\theta} = \langle \boldsymbol{\theta}_A, \boldsymbol{\theta}_{B|A}, \boldsymbol{\theta}_{C|A}, \boldsymbol{\theta}_{D|C} \rangle$$

Decomposability

$$\begin{aligned}
\boldsymbol{\theta}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A=a^{(t)}, B=b^{(t)}, C=c^{(t)}, D=d^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A=a^{(t)}) + \log P(B=b^{(t)} \mid A=a^{(t)}) \\
&\qquad + \log P(C=c^{(t)} \mid A=a^{(t)}) + \log P(D=d^{(t)} \mid C=c^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A=a^{(t)}) + \sum_t \log P(B=b^{(t)} \mid A=a^{(t)}) \\
&\qquad + \sum_t \log P(C=c^{(t)} \mid A=a^{(t)}) + \sum_t \log P(D=d^{(t)} \mid C=c^{(t)}) \\
&= \Big\langle \arg\max_{\boldsymbol{\theta}_A} \sum_t \log P(A=a^{(t)}), \arg\max_{\boldsymbol{\theta}_{B|A}} \sum_t \log P(B=b^{(t)} \mid A=a^{(t)}), \\
&\qquad \arg\max_{\boldsymbol{\theta}_{C|A}} \sum_t \log P(C=c^{(t)} \mid A=a^{(t)}), \arg\max_{\boldsymbol{\theta}_{D|C}} \sum_t \log P(D=d^{(t)} \mid C=c^{(t)}) \Big\rangle
\end{aligned}$$

Decomposability

$$\begin{aligned}
\boldsymbol{\theta}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}, B = b^{(t)}, C = c^{(t)}, D = d^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}) + \log P(B = b^{(t)} \mid A = a^{(t)}) \\
&\qquad\qquad + \log P(C = c^{(t)} \mid A = a^{(t)}) + \log P(D = d^{(t)} \mid C = c^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}) + \sum_t \log P(B = b^{(t)} \mid A = a^{(t)}) \\
&\qquad + \sum_t \log P(C = c^{(t)} \mid A = a^{(t)}) + \sum_t \log P(D = d^{(t)} \mid C = c^{(t)}) \\
&= \Big\langle \arg\max_{\boldsymbol{\theta}_A} \sum_t \log P(A = a^{(t)}), \arg\max_{\boldsymbol{\theta}_{B|A}} \sum_t \log P(B = b^{(t)} \mid A = a^{(t)}), \\
&\qquad \arg\max_{\boldsymbol{\theta}_{C|A}} \sum_t \log P(C = c^{(t)} \mid A = a^{(t)}), \arg\max_{\boldsymbol{\theta}_{D|C}} \sum_t \log P(D = d^{(t)} \mid C = c^{(t)}) \Big\rangle
\end{aligned}$$

Decomposability

$$
\begin{aligned}
\boldsymbol{\theta}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}, B = b^{(t)}, C = c^{(t)}, D = d^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}) + \log P(B = b^{(t)} \mid A = a^{(t)}) \\
&\qquad + \log P(C = c^{(t)} \mid A = a^{(t)}) + \log P(D = d^{(t)} \mid C = c^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}) + \sum_t \log P(B = b^{(t)} \mid A = a^{(t)}) \\
&\qquad + \sum_t \log P(C = c^{(t)} \mid A = a^{(t)}) + \sum_t \log P(D = d^{(t)} \mid C = c^{(t)}) \\
&= \Bigg\langle \arg\max_{\boldsymbol{\theta}_A} \sum_t \log P(A = a^{(t)}), \arg\max_{\boldsymbol{\theta}_{B|A}} \sum_t \log P(B = b^{(t)} \mid A = a^{(t)}), \\
&\qquad \arg\max_{\boldsymbol{\theta}_{C|A}} \sum_t \log P(C = c^{(t)} \mid A = a^{(t)}), \arg\max_{\boldsymbol{\theta}_{D|C}} \sum_t \log P(D = d^{(t)} \mid C = c^{(t)}) \Bigg\rangle
\end{aligned}
$$

Decomposability

$$
\begin{aligned}
\boldsymbol{\theta}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}, B = b^{(t)}, C = c^{(t)}, D = d^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}) + \log P(B = b^{(t)} \mid A = a^{(t)}) \\
&\qquad + \log P(C = c^{(t)} \mid A = a^{(t)}) + \log P(D = d^{(t)} \mid C = c^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \log P(A = a^{(t)}) + \sum_t \log P(B = b^{(t)} \mid A = a^{(t)}) \\
&\qquad + \sum_t \log P(C = c^{(t)} \mid A = a^{(t)}) + \sum_t \log P(D = d^{(t)} \mid C = c^{(t)}) \\
&= \Bigg\langle \arg\max_{\boldsymbol{\theta}_A} \sum_t \log P(A = a^{(t)}), \arg\max_{\boldsymbol{\theta}_{B|A}} \sum_t \log P(B = b^{(t)} \mid A = a^{(t)}), \\
&\qquad \arg\max_{\boldsymbol{\theta}_{C|A}} \sum_t \log P(C = c^{(t)} \mid A = a^{(t)}), \arg\max_{\boldsymbol{\theta}_{D|C}} \sum_t \log P(D = d^{(t)} \mid C = c^{(t)}) \Bigg\rangle
\end{aligned}
$$

Deriving the MLE

- Many distributions have a closed form for the MLE.
- Solve (analytically, and with constraints), ∀j:

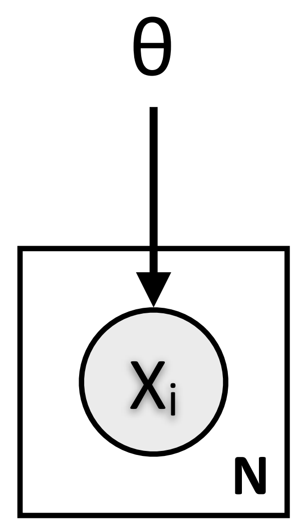
$$\frac{\partial}{\partial\theta_j}\sum_t \log P(X_i = x_i^{(t)} \mid \text{Parents}(X_i) = \text{Parents}(x_i^{(t)})) \quad = \quad 0$$

- Typically *convex*.
- Eg: Gaussian, binomial, multinomial.
- Today: Binomial and multinomial,
  with Lagrange Multipliers.

Binomial Distribution

- $P(Y = \text{heads}) = \theta$, $P(Y = \text{tails}) = 1 - \theta$
- "IID" assumption
  - Each flip is independent of the others.
  - All flips are distributed identically.

$$P(\boldsymbol{Y} \mid \theta, N) = \theta^{\#\text{heads}(\boldsymbol{Y})} \times (1-\theta)^{\#\text{tails}(\boldsymbol{Y})}$$



Maximum Likelihood Estimation

- Data: sequence **Y** of flip outcomes
- Assumption: binomial distribution; flips are IID
- Goal: select θ
- Maximum likelihood estimation: treat this as an optimization problem over θ

$$\begin{aligned}\theta_{\mathrm{MLE}} &= \arg\max_{\theta} P(\boldsymbol{Y} \mid \theta) \\ &= \arg\max_{\theta} \log P(\boldsymbol{Y} \mid \theta)\end{aligned}$$

MLE for the Binomial

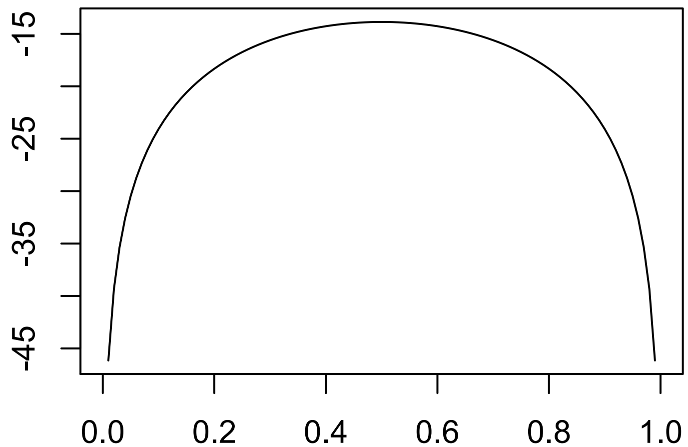
$$\begin{aligned}\theta_{\mathrm{MLE}} &= \arg\max_{\theta} P(\boldsymbol{Y} \mid \theta) \\ &= \arg\max_{\theta} \log P(\boldsymbol{Y} \mid \theta)\end{aligned}$$

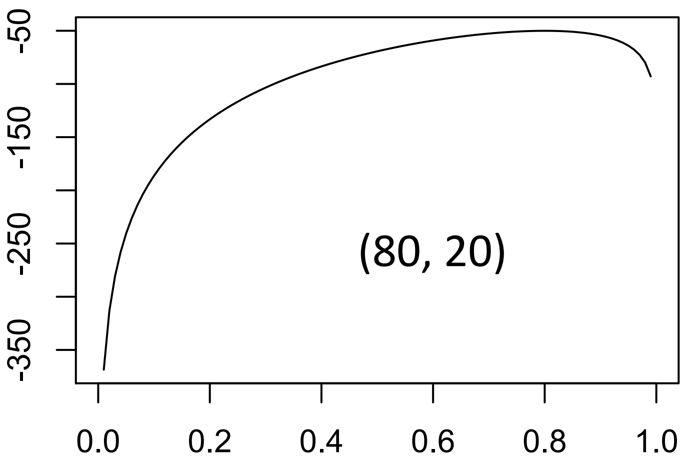
$$P(\boldsymbol{Y} \mid \theta, N) = \theta^{\#\mathrm{heads}(\boldsymbol{Y})} \times (1-\theta)^{\#\mathrm{tails}(\boldsymbol{Y})}$$

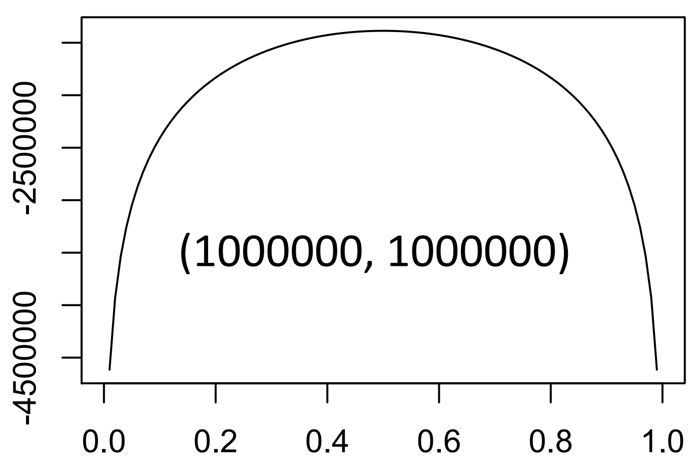
$$\arg\max_{\theta} \#\mathrm{heads}(\boldsymbol{Y}) \log\theta + \#\mathrm{tails}(\boldsymbol{Y}) \log(1-\theta)$$

MLE for the Binomial

(10, 10)







Deriving the Binomial MLE

- Board work
- Use a little calculus...

Deriving the Multinomial MLE

- Board work
- Introduce Lagrange Multipliers
- Use them to solve for MLE of a multinomial.

Deriving Functional Form for Maximum Entropy Classifiers

- Board work
- Lagrange again...

Generalized Linear Model

- Score is defined as a *linear* function of **X**:

$$f(\boldsymbol{X}) = \underbrace{w_0 + \sum_i w_i X_i}_{Z}$$

Z = f(X) is a random variable

- Probability distribution over binary value Y is defined by:

$$P(Y = 1) = \text{sigmoid}(f(\boldsymbol{X}))$$

- Sample Y.

From lecture 3!

$$\text{sigmoid}(z) = \frac{e^z}{1 + e^z}$$

Markov Networks as a Generalized Linear Model

- Sigmoid equates to *binary* output log-linear model.
- More generally, *multinomial* logit:
  take a linear score (Z in lecture 3),
  exponentiate, and normalize (Z in Gibbs dist.)
  - Don't confuse the Zs.

- The generalized linear model we used for CPDs is a log-linear distribution.