Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Maximum Likelihood Estimation

Introduction to Data Science Algorithms
Jordan Boyd-Graber and Michael Paul
SEPTEMBER 29, 2016

Discrete Distribution: Multinomial

- Recall the density function (N is total number of observations, x_i is the number for each cell, θ_i probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

Discrete Distribution: Multinomial

- Recall the density function (N is total number of observations, x_i is the number for each cell, θ_i probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log n! - \sum_i \log x_i! + \sum_i x_i \log \theta_i \tag{2}$$

Discrete Distribution: Multinomial

- Recall the density function (N is total number of observations, x_i is the number for each cell, θ_i probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log n! - \sum_i \log x_i! + \sum_i x_i \log \theta_i \tag{2}$$

Discrete Distribution: Multinomial

- Recall the density function (N is total number of observations, x_i is the number for each cell, θ_i probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log n! - \sum_i \log x_i! + \sum_i x_i \log \theta_i \tag{2}$$

- Recall the density function (N is total number of observations, x_i is the number for each cell, θ_i probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log n! - \sum_i \log x_i! + \sum_i x_i \log \theta_i \tag{2}$$

MLE of Multinomial θ

$$\ell(\vec{\theta}) = \log N! - \sum_i \log x_i! + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{3}$$

$$\tag{4}$$

MLE of Multinomial θ

$$\ell(\vec{\theta}) = \log N! - \sum_i \log x_i! + \sum_i x_i \log \theta_i + \lambda\left(1 - \sum_i \theta_i\right) \tag{3}$$

$$\tag{4}$$

Where did this come from? Constraint that $\vec{\theta}$ must be a distribution.

MLE of Multinomial θ

$$\ell(\vec{\theta}) = \log N! - \sum_i \log x_i! + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{3}$$

$$\tag{4}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

MLE of Multinomial θ

$$\ell(\vec{\theta}) = \log N! - \sum_i \log x_i! + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{3}$$

$$\tag{4}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

MLE of Multinomial θ

$$\ell(\vec{\theta}) = \log N! - \sum_i \log x_i! + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{3}$$

$$\tag{4}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

MLE of Multinomial θ

$$\ell(\vec{\theta}) = \log N! - \sum_i \log x_i! + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{3}$$

$$\tag{4}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

MLE of Multinomial θ

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{5}$$

$$\vdots \quad \vdots \tag{6}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{7}$$

$$\sum_i \theta_i = 1 \tag{8}$$

MLE of Multinomial θ

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{5}$$

$$\vdots \quad \vdots \tag{6}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{7}$$

$$\sum_i \theta_i = 1 \tag{8}$$

- So let's substitute the first K equations into the last:

$$\sum_i \frac{x_i}{\lambda} = 1 \tag{9}$$

MLE of Multinomial θ

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{5}$$

$$\vdots \quad \vdots \tag{6}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{7}$$

$$\sum_i \theta_i = 1 \tag{8}$$

- So let's substitute the first K equations into the last:

$$\sum_i \frac{x_i}{\lambda} = 1 \tag{9}$$

- So $\lambda = \sum_i x_i = N$,

MLE of Multinomial θ

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{5}$$

$$\vdots \quad \vdots \tag{6}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{7}$$

$$\sum_i \theta_i = 1 \tag{8}$$

- So let's substitute the first K equations into the last:

$$\sum_i \frac{x_i}{\lambda} = 1 \tag{9}$$

- So $\lambda = \sum_i x_i = N$, and $\theta_i = \frac{x_i}{N}$

But why are we adding one?

- But you told us to add one while estimating multinomials!
- Difference between MLE and MAP
- mle assumes only the data distribution
- map assumes a distribution over parameters too (technically for Laplace, Dirichlet with $\alpha_i = 1$)
- Recall that we showed Dirichlet parameter can be viewed as Pseudocounts