



Maximum Likelihood Estimation

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

$$\rho(\vec{x} \mid \vec{\theta}) = \frac{N!}{\prod_{i} x_{i}!} \prod \theta_{i}^{x_{i}}$$
(1)

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

$$\rho(\vec{x} \mid \vec{\theta}) = \frac{N!}{\prod_{i} x_{i}!} \prod \theta_{i}^{x_{i}}$$
(1)

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

$$\ell(\vec{\theta}) \equiv \log n! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i}$$
⁽²⁾

$$p(\vec{x} \mid \vec{\theta}) = \frac{N!}{\prod_{i} x_{i}!} \prod \theta_{i}^{x_{i}}$$
(1)

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

$$\ell(\vec{\theta}) \equiv \log n! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i}$$
(2)

$$\rho(\vec{x} \mid \vec{\theta}) = \frac{N!}{\prod_{i} x_{i}!} \prod \theta_{i}^{x_{i}}$$
(1)

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

$$\ell(\vec{\theta}) \equiv \log n! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i}$$
(2)

$$p(\vec{x} \mid \vec{\theta}) = \frac{N!}{\prod_{i} x_{i}!} \prod_{j} \theta_{i}^{x_{j}}$$
(1)

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

$$\ell(\vec{\theta}) \equiv \log n! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i}$$
(2)

$$\ell(\vec{\theta}) = \log N! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i} + \lambda \left(1 - \sum_{i} \theta_{i}\right)$$
(3)
(4)

$$\ell(\vec{\theta}) = \log N! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i} + \lambda \left(1 - \sum_{i} \theta_{i}\right)$$
(3)
(4)

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

$$\ell(\vec{\theta}) = \log N! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i} + \lambda \left(1 - \sum_{i} \theta_{i}\right)$$
(3)
(4)

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

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

$$\ell(\vec{\theta}) = \log N! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i} + \lambda \left(1 - \sum_{i} \theta_{i}\right)$$
(3)
(4)

•
$$\frac{\partial \ell}{\partial \theta_i} = \frac{\mathbf{x}_i}{\theta_i} - \lambda$$

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

$$\ell(\vec{\theta}) = \log N! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i} + \lambda \left(1 - \sum_{i} \theta_{i}\right)$$
(3)
(4)

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

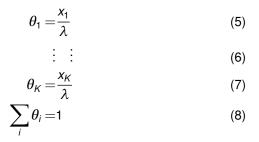
• $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

$$\ell(\vec{\theta}) = \log N! - \sum_{i} \log x_{i}! + \sum_{i} x_{i} \log \theta_{i} + \lambda \left(1 - \sum_{i} \theta_{i}\right)$$
(3)
(4)

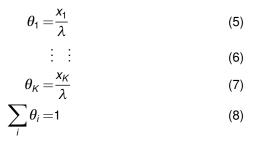
•
$$\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$$

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

We have system of equations



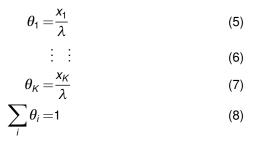
We have system of equations



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

$$\sum_{i} \frac{x_i}{\lambda} = 1 \tag{9}$$

We have system of equations

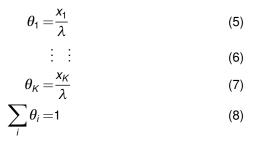


• So let's substitute the first *K* equations into the last:

$$\sum_{i} \frac{x_{i}}{\lambda} = 1 \tag{9}$$

• So
$$\lambda = \sum_i x_i = N$$
,

We have system of equations



• 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 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 α_i = 1)
- Recall that we showed Dirichlet parameter can be viewed as Pseudocounts