Course › Unit 4 Unsupervised Learning (2 weeks) › Lecture 15. Generative Models ›
6. MLE for Multinomial Distribution

6. MLE for Multinomial Distribution

Maximum Likelihood Estimate

0:00 / 0:00 1.25x

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Deriving MLE for a General Multinomial Model: Likelihood

1/1 point (graded)

In the following problems, we will derive the maximum likelihood estimates for a multinomial model with more than 2 parameters. We will employ the method of lagrange multipliers for the optimization problem.

Let the document D be a sequence of words $w_1, \ldots, w_n$ from a collection W consisting of N words. For simplicity, we assume that w_i's are independent, and that the probability of a word w is given by the parameter θ_w, and denote by $\theta = \{\theta_w\}_{w \in W}$.

Let $P(D|\theta)$ be the probability of D being generated by the simple model described above.

Find $P(D|\theta)$.

○ $P(D|\theta) = \sum_{w \in W} \theta_w^{\text{count}(w)}$

● $P(D|\theta) = \prod_{w \in W} \theta_w^{\text{count}(w)}$

○ $P(D|\theta) = \prod_{w \in W} \text{count}(w)^{\theta_w}$

○ $P(D|\theta) = \prod_{w \in W} \theta_w + \text{count}(w)$

✓

Solution:

The probability of D

$$P(D|\theta) = \prod_{i=1}^{n} \theta_{w_i} = \prod_{w \in W} \theta_w^{\text{count}(w)}$$

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Constraints on the Parameters

1/1 point (graded)

What are the constraints on the parameters θ_w in the model described in the previous problem?

● $\theta_w \geq 0, \sum_{w\in W} \theta_w = 1$

○ $\theta_w \geq 0, \sum_{w\in W} \theta_w < 1$

○ $\theta_w < 0, \sum_{w\in W} \theta_w > -1$

○ $\theta_w \geq 0, \sum_{w\in W} \theta_w \geq 1$

✓

Solution:

Since θ_w denotes the probability of the word w under the model, its value must lie between 0 and 1. Therefore, $0 \leq \theta_w \leq 1$.

Further, all the above probability values must also sum up to 1. That is, $\sum_{w\in W} \theta_w = 1$.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Stationary Points of the Lagrange Function

2/2 points (graded)

The maximum likelihood estimate of θ is the value of θ that maximizes the likelihood function:

$$P\left(D|\theta\right)=\prod_{w\in W}\left(\theta_w\right)^{\text{count}(w)}.$$

Maximizing $P\left(D|\theta\right)$ is equivalent to maximizing $\log P\left(D|\theta\right)$, so we take the natural logarithm on both sides of the equation above to bring down the exponents:

$$\log P\left(D|\theta\right)=\sum_{w\in W}\text{count}\left(w\right)\log\theta_w.$$

Recall that θ is subject to the following constraint:

$$\sum_{w\in W}\theta_w=1.$$

To maximize $\log P\left(D|\theta\right)$ subject to the contraint $\sum_{w\in W}\theta_w=1$, we use the Lagrange multiplier method.

Method of Lagrange Multipliers

Problem

Let f be a function from $\mathbb{R}^N$ to $\mathbb{R}$. Find the (local) maxima/minima of f **subject to a given constraint** $g=0$, where g is a function $\mathbb{R}^N$ to $\mathbb{R}$.

A two dimensional example is: Find the local extrema of $f\left(x,y\right)=x^2$ subject to the constraint $x^2+y^2=1$ i.e. optimize the function f on the unit circle.

Method of Lagrange Multipliers

Without the constraint, the optimization problem can be solved as usual by setting the gradient of f to zero i.e.

$$\nabla f = 0.$$

With the constraint, we can solve the following equation instead:

$$\nabla f = \lambda \nabla g$$

where λ is a constant scalar. Geometrically, for $\lambda \neq 0,$ a solution to the equation above is a point in $\mathbb{R}^N$ where the gradient of f is "parallel" to the gradient of g, or equivalently, where the gradient of f is perpendicular to the tangent of the curve defined by $g = k$ for some k. In other words, at a solution point, the directional derivative of f is zero along the direction tangent to the curve $g = k$ for some constant k, and hence f is stationary as we travel along $g = k$.

Finally, we impose the constraint $g = 0$ to find the local extrema of f on $g = 0$.

Since the equation $\nabla f = \lambda \nabla g$ is equivalent to $\nabla L = 0$ where $L = f - \lambda g$, the problem of optimizing f subject to $g = 0$ can be reformulated as optimizing the function L along with the constraint $g = 0$. We call the function L the **Lagrangian function** , and the scalar λ the **Lagranage multiplier** .

Note that we can equally define $L = f + \lambda g$, since λ is an unknown scalar we will solve for.

Example

Find the local extrema of $f(x, y) = x^2$ subject to the constraint $x^2 + y^2 = 1$. Geometrically, the function f is a parabolic cylinder, i.e. f is a parabolic in the x direction with constant values in the y direction. The constraint is a unit circle.

Solution:
First, solve the equation

$$\nabla f \quad = \lambda \nabla g \qquad \text{where } g(x, y) = x^2 + y^2 - 1$$

$$\begin{aligned} \iff \quad \begin{bmatrix} 2x \\ 0 \end{bmatrix} &= \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix} \\ \iff \quad \begin{bmatrix} (1-\lambda)\,2x \\ \lambda\,(2y) \end{bmatrix} &= 0 \end{aligned}$$

The set of all possible solutions to the equation above are $(\lambda = 1, y = 0)$, or $(\lambda = 0,\ x = 0)$, or $(x = y = 0)$.

Finally, impose the constraint $x^2 + y^2 - 1 = 0$ to further pin down the local extrema. Subject to $x^2 + y^2 = 1$, $f(x, y) = x^2$ is at local maximum or mininum at $(x = 0, y = \pm 1)$ and $(y = 0, x = \pm 1)$. At $(x = 0, y = \pm 1)$, we have $\lambda = 0$ and $\nabla f = 0$. Since f has only local minima, these two points remain local minima of f on the unit circle. At $(y = 0, x = \pm 1)$, we have $\lambda = 1$ and hence $\nabla f = \nabla g$. Equivalently, the directional derivative ∇f is zero along the tangent direction of the circle at this point. Visualizing or computing second derivatives will allow us to see that these two points are local maxima of f along the unit circle.

Hide

Define the Lagrange function:

$$L = \log P(D|\theta) + \lambda \left(\sum_{w \in W} \theta_w - 1 \right)$$

where λ is a constant scalar.

Then, find the stationary points of L by solving the equation $\nabla_\theta L = 0$. The components of this equation are

$$\frac{\partial}{\partial \theta_w} \left(\log P(D|\theta) + \lambda \left(\sum_{w \in W} \theta_w - 1 \right) \right) = 0 \qquad \text{for all } w \in W.$$

Solve for θ_w from the above equation. Choose the right answer for θ_w from options below.

○ $\theta_w = \frac{-\lambda}{\text{count}(w)}$

○ $\theta_w = \lambda\text{count}(w)$

○ $\theta_w = -\lambda\text{count}(w)$

● $\theta_w = \frac{-\text{count}(w)}{\lambda}$

✓

Now, apply the constraint that $\sum_{w \in W} \theta_w = 1$ to the answer above to obtain λ.

$\lambda =$

● $\lambda = -\sum_{w \in W} \text{count}(w)$

○ $\lambda = \sum_{w \in W} \text{count}(w)$

○ $\lambda = -\sum_{w \in W} (\theta_w \text{count}(w))$

○ $\lambda = \sum_{w \in W} (\theta_w \text{count}(w))$

✓

Find θ_w that maximizes $\log P\left(D|\theta\right)$ subject to $\sum_{w \in W} \theta_w = 1$.

(There is no answer box for this final question.)

Solution:

$$\frac{\partial}{\partial \theta_w}(\log P\left(D|\theta\right) + \lambda\,(\sum_{w \in W} \theta_w - 1)) = 0$$

$$\frac{\partial \log P\left(D|\theta_w\right)}{\partial \theta_w} + \lambda = 0$$

$$\frac{\partial \log \prod_{w \in W} \theta_w^{\text{count}(w)}}{\partial \theta_w} + \lambda = 0$$

$$\frac{\partial \sum_{w \in W} \log \theta_w \times \text{count}\left(w\right)}{\partial \theta_w} + \lambda = 0$$

$$\frac{\text{count}\left(w\right)}{\theta_w} + \lambda = 0$$

$$\theta_w = -\frac{\text{count}\left(w\right)}{\lambda}$$

If we apply the constraint that $\sum_{w \in W} \theta_w = 1$ we get

$$\sum_{w \in W} \theta_w = 1$$

$$\sum_{w \in W} -\frac{\text{count}\,(w)}{\lambda} = 1$$

$$\sum_{w \in W} \text{count}\,(w) = -\lambda$$

$$\lambda = -\sum_{w \in W} \text{count}\,(w)$$

Substituting this expression for λ back into our previous expression for θ_w we get

$$\theta_w = -\frac{\text{count}\,(w)}{\lambda}$$

$$\theta_w = \frac{\text{count}\,(w)}{\sum_{w \in W} \text{count}\,(w)}$$

Note that $\theta_w > 0$ and $\sum_{w \in W} \theta_w = 1$ satisfying all the constraints. These set of θ_w parameters are the maximum likelihood estimates for this multinomial generative distribution.

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Lecture 15. Generative Models / 6. MLE for Multinomial Distribution

Add a Post

Show all posts	by recent activity
Lagrange multiplier missing details?	6
For those who don't understand lagrange method from the text Hope this video would help https://www.youtube.com/watch?v=yuqB-d5MjZA	1