Lecture #5 - Parameter Estimation, MLE and Lagrange Multipliers

Lior Ziv
21/11/16

- We want to align two DNA sequences of a monkey and a beetle, let's go back to our two hypotheses:
  - \( H_1 \)-sequences have the same ancestor
  - \( H_0 \)-sequences are independent, derived from different ancestor

Let the alignment in index \( i \), \( A_i \), be: \( A_i = [s_i, t_i] \), what is the probability of such an alignment?

- Origin from different ancestor - independent events - \( P_0(s_i, t_i) = P_0(s_i) \cdot P_0(t_i) \)
- Sequences share a common ancestor - \( P_1(s_i, t_i) = \) ?

For now let's assume we have both of the distributions and calculate the log likelihood ratio (LLR)

- Now, for a given alignment \( A \) of \( s \) and \( t \):

\[
-\text{LLR} = \log \frac{P(A | H_1)}{P(A | H_0)} = \log \left( \frac{\prod_{i=1}^{|A|} P_1(s_i, t_i)}{\prod_{i=1}^{|A|} P_0(s_i) \cdot P_0(t_i)} \right) = \sum_{i=1}^{|A|} \log \left( \frac{P_1(s_i, t_i)}{P_0(s_i) \cdot P_0(t_i)} \right) = \sum_{i=1}^{|A|} \sigma(s_i, t_i)
\]

We define our score function this way because of Neyman-Pearson lemma which states that given certain significance level, the LLR gives us the best separator between our two hypotheses.

Parameter Estimation

- Motivation, we want to estimate the parameters for our two distributions:
  - \( P_1(a, b) \) - the probability of 'a' to be aligned with base 'b' when derived from a common ancestor
  - \( P_0(a) \) - the independent probability of base 'a'
For this we will first see an example of parameter estimation for a coin Bernoulli distribution

- Sample space \( \Sigma = \{H, T\} \)
- \( P_\theta : \Sigma \rightarrow \mathbb{R} \) s.t \( P(x) = \begin{cases} p & x = H \\ 1 - p & x = T \end{cases} \)

Now lets look at a sequence of coin tossing \( D : \{H, H, T, H\} \)

\[
L(\theta; D) = P(D|\theta) = P_\theta(D) = p^3(1-p)^1 = p^{N_H}(1-p)^{N_T}
\]

\( N_H, N_T = \) number of times the result was H or T respectively.

\( \{N_H, N_T\} \) are our sufficient statistics in this case.

**Definition:** If \( \forall D_1, \forall D_2 (s(D_1) = s(D_2)) \rightarrow \forall \theta L(\theta : D_1) = L(\theta : D_2) \), \( s \) is the sufficient statistics

In our coin tossing experiment, for every two sequences of tossing with equivalent sufficient statistics, both sequences have the same number of \( N_H, N_T \), therefore we get that their LLR is equal.

So now how can we assess \( \theta \) ?

**MLE - Maximum Likelihood Estimation**

- Where will the MLE be?
  
  Maximum likelihood refers to the maximum of the likelihood function. That is the maximum of seeing our data \( D \) given \( \theta_{MLE} \) parameters \( L(\theta_{MLE}) = P(D|\theta_{MLE}) \)
  
  Since log is monotonic, \( \arg\max_\theta \log(L) = \arg\max_\theta L \) so in our Bernoulli example, in order to find \( \theta_{MLE} \) we need to derive the function \( \log(L) \)

- Since we want to maximize this function we derive the log likelihood with respect to \( \theta \):
  
  \[
  - \log L = N_H \cdot \log \theta + N_T \cdot \log (1-\theta) \Rightarrow \frac{\partial \log L}{\partial \theta} = \frac{N_H}{\theta} - N_T \cdot \frac{1}{1-\theta} = \frac{N_H}{\theta} - \frac{N_T}{1-\theta} , \text{ in order to find the maximum we compare the result to 0}
  \]
  
  \[
  \frac{N_H}{\theta} - \frac{N_T}{1-\theta} = 0 \rightarrow N_T \theta = N_H (1-\theta) \rightarrow \theta (N_H + N_T) = N_H \rightarrow \theta = \frac{N_H}{N_H + N_T}
  \]

In the above example \( D = \{H, H, T, H\} \) \( N_H = 3, N_T = 1 \) therefore \( \theta = 0.75 \)
Lagrange Multipliers

- **Motivation**: In some cases where we will want to maximize the likelihood we will have some constraint. For example, for our \(H_0\) and \(H_1\), \(P_0(a) = \sum_b(P_1(a,b))\). That is \(P_0\) is the marginal distribution of \(P_1\), also we might want to demand some symmetry: \(P_1(a,b) = P_1(b,a)\). Therefore we will want to learn how to maximize while satisfying our constraints.

- We want to assess our parameters \(\theta\), in the case of our coin example \(\theta = p\) s.t \(0 \leq p \leq 1\) and we have seen
  \[
  L(\theta : D) = p^{N_1} \cdot (1-p)^{N_0}
  \]
  In this case, the constraint for \(0 \leq p \leq 1\) is incorporated within the function because \(L(\theta : D)\) is non-negative only when \(0 \leq p \leq 1\). In the multinomial case this is not the case.

For \(X \sim Mult(\theta_1,...\theta_k)\), \(\theta = \langle \theta_1,..,\theta_k \rangle\) we have some constraints:

1. \(\sum_{i=1}^{k} \theta_i = 1\)
2. \(\forall (1 \leq i \leq n) \cdot 0 \leq \theta_i \leq 1\)

The likelihood is defined then by

\[
\rightarrow L(\theta : D) = \theta_1^{N_1} \cdots \theta_k^{N_k}
\]

For example in a sequence of dice tossing if we get \(D = \{1, 1, 1, 2, 2\}\)

\[
N_1 = 3, N_2 = 2 N_{3-6} = 0 \rightarrow L(\theta : D) = \theta_1^{N_1} \cdots \theta_6^{N_6}
\]

Pay attention that we can choose \(\theta\) to be as large as we want because the function for itself doesn’t have any constraints on \(\theta\), which is our motivation for Lagrange Multipliers.

**Lagrange Multipliers** is an optimization method under some given constraints, more formally we want to get the \(x\) that maximizes \(f(x)\) under the constraints \(g_1(x) = 0...g_l(x) = 0\)

\[
\text{argmax } f(x) \quad \text{s.t.}
\]

\[
g_1(x) = 0 \quad g_{l-1}(x) = 0
\]

\[
g_l(x) = 0
\]
For each constraint $g_j(x)$ we will define a Lagrange Multiplier $\lambda_j$ (scalar) s.t $0 \leq j \leq l$. We can look on each $\lambda_j$ as a weight that defines how meaningful each constraint is, here however, we will demand that our answer satisfies all constraints.

**Definition** - Given (our problem written above) , $J(\vec{x}, \vec{\lambda}) = f(\vec{x}) - \Sigma g_j(\vec{x}) \cdot \lambda_j$ is defined as the Lagrangian.

- Then with the Lagrange multipliers we are guaranteed to get the $x$ that maximizes $f(x)$ under our constraints when $x$ satisfies $\vec{x}$: $\nabla J(\vec{x}, \vec{\lambda}) = 0$

  We demand $\nabla J(\vec{x}, \vec{\lambda}) = <\frac{\partial f}{\partial x}, \frac{\partial f}{\partial \lambda}> = <\frac{\lambda \nabla g - \nabla f}{\partial x}, -g>$ so we get:

  - $\lambda \nabla g - \nabla f = 0$ (*)
  - $-g = 0$ (means we meet the constraints)

(*)At the maximal point our $\nabla f$ gradient should be perpendicular to the surface created by our constraint (for each constraint) otherwise we can find a better maximum for our function,

(it means we can go up the gradient since we are not at a point where the projection of the gradient of $f$ equals zero which is a maxima)

Obviously the $\nabla g$ is also perpendicular to the surface since it is constant (the surface is defined by a constant value of $g$)
As a result we get $\nabla g, \nabla f$ are pointing at the same direction which leads us to the fact they are parallel in the maximum point that satisfies the constraints.

**Example**

We want to solve the following problem:

$$\arg\max_{x,y} f(x, y) = 3x + 4y \quad g_1 \rightarrow x^2 + y^2 = 1$$

Figure 2: 3D plot of $f(x) = 3x + 4y, x^2 + y^2 = 1$

3D plot:

- The Lagrangian is defined then by: $J(x, y, \lambda)$
  - Because the gradient of $f$ is parallel to the gradient of $g$ we have:
    
    $$J = 3x + 4y - \lambda(x^2 + y^2 - 1)$$

    $$\nabla f = <3, 4> \Rightarrow 3 = 2x\lambda \Rightarrow x = \frac{3}{2\lambda}$$

    $$\nabla g = <2x, 2y> \Rightarrow 4 = 2y\lambda \Rightarrow y = \frac{1}{2\lambda}$$

- Now let’s place $x,y$ back to our constraint equation

  $$(\frac{3}{2\lambda})^2 + (\frac{1}{2\lambda})^2 = 1 \rightarrow \frac{\lambda}{2\lambda} = \pm \sqrt{2.5}$$

  - for $\lambda = \sqrt{2.5} \rightarrow (\frac{3}{2}, \frac{1}{2})$ maximum point
  - for $\lambda = -\sqrt{2.5} \rightarrow (-\frac{3}{2}, -\frac{1}{2})$ minimum point

  We found two points which maximize our $f$ and uphold out constraints
Back to the case of Multinomial distribution -

We have the constraint \( \sum_{i=1}^{k} \theta_i = 1 \)

and the function we want to maximize is - \( L(\theta : D) = \prod_{i=1}^{k} \theta_i^{N_i} \)

We use the log in order to make it easier to derive - \( LL(\Theta : D) = \sum_{i=1}^{k} N_i log\theta_i \)

By the definition of J we get \( \Rightarrow J(\theta, \lambda) = \sum_{i=1}^{k} N_i log\theta_i - \lambda (\sum_{i=1}^{k} \theta_i - 1) \)

Now we want to derive \( J(\theta, \lambda) \) in order to find the maximum points

\[ 0 = \sum_{i=1}^{k} \frac{N_i \theta_i}{\lambda} - \lambda = \Rightarrow \theta_i = \frac{N_i}{\lambda} \]

like in the example we place back \( \theta_i \) in our constraint

\[ 1 = \sum_{i=1}^{k} \theta_i = \frac{k}{\lambda} \sum_{i=1}^{k} N_i \Rightarrow \lambda = \frac{k}{\sum_{i=1}^{k} N_i} = N \]

• So, for all \( i, \theta_i = \frac{N_i}{N} \) This means that \( \theta_1 = \frac{N_1}{N}, \ldots, \theta_k = \frac{N_k}{N} > \)

satisfies the Multinomial constraint and maximizes the likelihood of such data (\( N_k \) observations of item \( k \)).