Data, Estimation and Inference

Pedro Piniés
ppinies@robots.ox.ac.uk
Michaelmas 2016
Topic 4: Maximum Likelihood Estimation
Maximum likelihood estimation (MLE) is a method used to estimate the parameters $\theta$ or the inputs $x$ to a model $f_\theta(x, \epsilon)$ from noisy data $y$.

$$y = f_\theta(x, \epsilon)$$

The model describes how the data is generated given the parameters and the inputs.
Maximum likelihood estimation (MLE) is a method used to estimate the parameters $\theta$ or the inputs $x$ to a model $f_\theta(x, \epsilon)$ from noisy data $y$

$$y = f_\theta(x, \epsilon) = x + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
The model describes how the data is generated given the parameters and the inputs

\[ y = f_\theta(x, \epsilon) = x + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \]

A conditional distribution arises naturally from the previous statement

\[ p(y|x, \theta) \]
The model describes how the data is generated given the parameters and the inputs.

\[ y = f_\theta(x, \epsilon) = x + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \]
However what we actually know is $y$ and $\theta$ and the unknown in this case is $x$
Maximum Likelihood Intuition

\[ p(y|x, \theta) \]

\[ y = f_\theta(x, \epsilon) = x + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \]
Maximum Likelihood Intuition

\[ y = f_\theta(x, \epsilon) = x + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \]
Maximum Likelihood Estimation is about finding the value of the input $x$ or the parameter $\theta$ that maximises $p(y\mid x, \theta)$

Note that when we fix $y$ and vary either $x$ or $\theta$ then $p(y\mid x, \theta)$ is not a distribution anymore and this is why we use instead the term likelihood function and the notation $\mathcal{L}(x\mid y, \theta)$ or $\mathcal{L}(\theta\mid y, x)$

Therefore MLE searches for the parameter $\theta^*$ that maximises the likelihood function $\mathcal{L}(\theta\mid y, x)$

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta\mid y, x)$$
Log-Likelihood: Usually it is easier to work with the log-likelihood instead

\[ \theta^* = \arg\min_{\theta} - \log L(\theta | y, x) \]

The Likelihood and Log-Likelihood satisfy:

\[ \frac{\partial L(\theta | y, x)}{\partial \theta} \bigg|_{\theta^*} = 0 \]

\[ \frac{\partial^2 L(\theta | y, x)}{\partial \theta^2} \bigg|_{\theta^*} < 0 \]

\[ \frac{\partial (- \log (L(\theta | y, x)))}{\partial \theta} \bigg|_{\theta^*} = 0 \]

\[ \frac{\partial^2 (- \log (L(\theta | y, x)))}{\partial \theta^2} \bigg|_{\theta^*} > 0 \]
Exercise 1
We toss a coin $n$ times. Given the corresponding outcome, estimate the probability of Heads ($X=1$)
Exercise 2

We want to estimate the position $x$ of a robot. We mount on it a sensor that measures the distance from the current pose to the origin of the trajectory with some uncertainty (sigma = 0.5 m).

From the current pose and using that sensor we take a measurement $z_1 = 3$ meters. Given that information, What is the best estimate of the pose of the robot? What is the uncertainty of the estimate?
Exercise 2 Solution

1- Model: 
\[ z_1 = x + n \quad n \sim \mathcal{N}(0, \sigma^2) \]

2- Conditional: 
\[ p(z_1|x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(z_1-x)^2}{2\sigma^2}} \]

3- LogLikelihood: 
\[ -\log \mathcal{L}(x|z_1) = c + \frac{(z_1-x)^2}{2\sigma^2} \]

4- Condition: 
\[ \frac{\partial (-\log \mathcal{L}(x|z_1))}{\partial x} = \frac{(z_1-x)}{\sigma^2} = 0 \]

5- Solution: 
\[ x^* = z_1 \]
Exercise 3
We take a second measurement $z_2 = 3.2$ meters. Can you think of a way of combining these two measurements? What is the uncertainty of the new estimate?
Exercise 3 Solution

1- Model: $z_1 = x + n_1, \, z_2 = x + n_2 \quad n_1, \, n_2 \sim \mathcal{N}(0, \sigma^2)$

2- Conditional: $p(z_1, z_2|x) = p(z_1|x)p(z_2|x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{(z_1-x)^2}{2\sigma^2} + \frac{(z_2-x)^2}{2\sigma^2}\right)}$

3- LogLikelihood: $-\log \mathcal{L}(x|z_1, z_2) = c + \frac{(z_1-x)^2}{2\sigma^2} + \frac{(z_2-x)^2}{2\sigma^2}$

4- Condition: $\frac{\partial (-\log \mathcal{L}(x|z_1, z_2))}{\partial x} = \frac{(z_1-x)}{\sigma^2} + \frac{(z_2-x)}{\sigma^2} = 0$

5- Solution: $x^* = \frac{z_1 + z_2}{2}$
Linear regression

Given $y$ and the inputs $x$ we want to estimate the parameters of the curve $\theta$ (model) that best fits the data.

**Important:** The model may be non-linear in the inputs $x$ but it is linear in the parameters $\theta$.

$$y = f_\theta(x, \epsilon)$$
Simplest case: the model is linear in $x$

$$y_i = f_\theta(x_i, \epsilon_i) = \theta_1 x_i + \theta_0 + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
Linear regression: linear in $x$

We stack all measurements in a vector so that we can express the model in vector form

$$
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n \\
\end{bmatrix} =
\begin{bmatrix}
  x_1 & 1 \\
  x_2 & 1 \\
  \vdots & \vdots \\
  x_n & 1 \\
\end{bmatrix}
\begin{bmatrix}
  \theta_1 \\
  \theta_0 \\
  \vdots \\
  \epsilon_n \\
\end{bmatrix}
+ \begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_n \\
\end{bmatrix}
$$

$$
y = X\theta + \epsilon \quad p(\epsilon) \sim \mathcal{N}(0, \Sigma)
$$

From the previous expression, what is $p(y|X, \theta)$?
Linear regression: linear in $x$

Once we know $p(y|X, \theta)$ we can estimate $\theta$ using ML
Linear regression

General case: the model is non-linear in $x$

$$y_i = f_\theta(x_i, \epsilon_i) = \theta_m \phi_m(x_i) + \cdots + \theta_1 \phi_1(x_i) + \theta_0 + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
Linear regression: non-linear in x

Key Idea: We extend the class of models by considering linear combinations of fixed non-linear functions $\phi_j(x)$ of the input variables.

$\phi_j(x)$ are known as basis functions (or features in some contexts).

Some examples of basis functions:

$$\phi_j(x) = \{x^2, x^3, x^n, \cdots, \sin(x), \cos(x), \cdots, \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)\}$$
Linear regression: non-linear in \( x \)

Again, we stack all measurements in a vector so that we can express the model in vector form

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} = \begin{bmatrix}
\phi_m(x_1) & \cdots & \phi_1(x_1) & 1 \\
\phi_m(x_2) & \cdots & \phi_1(x_2) & 1 \\
\vdots & & \vdots & \\
\phi_m(x_n) & \cdots & \phi_1(x_n) & 1
\end{bmatrix} \begin{bmatrix}
\theta_m \\
\vdots \\
\theta_1 \\
\theta_0
\end{bmatrix} + \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]

\[
y = \Phi(x)\theta + \epsilon \quad \mathcal{P}(\epsilon) \sim \mathcal{N}(0, \Sigma)
\]
Linear regression: non-linear in $x$

As in the linear case the ML solution is obtained by solving the normal equations:

$$(\Phi^T \Sigma^{-1} \Phi) \theta_{ML} = \Phi^T \Sigma^{-1} y$$

Given a new input $x_{new}$ the ML prediction is given by

$$\phi(x_{new})^T \theta_{ML} = \sum_{i=1}^{m} \phi_i(x_{new}) \theta_{iML}$$
In summary,

1. Maximum likelihood estimation (MLE) is a method used to estimate the parameters or the inputs to a model from noisy data.

2. We use the Likelihood function notation $L(\theta|y, x)$ or $L(x|y, \theta)$ to clearly specify the variable we want to estimate.

3. It is usually easier to work with the Log-Likelihood.

4. In linear regression the model is linear in the parameters which are estimated by solving the normal equations.