MLE

1. 算法简介

最大似然估计提供了一种给定观察数据来评估模型参数的方法，即：“模型已定，参数未知”。最大似然估计的一般求解过程如下：

1. 写出似然函数
2. 对似然函数取对数，并整理
3. 对不同参数求偏导
4. 解似然方程

http://blog.csdn.net/hezhourongro/article/details/17167717?locationNum=15(该博客讲的更加详细)。

2. 实例讲解

假设我们要统计全国人口的身高，首先假设这个身高服从服从正态分布，但是该分布的均值与方差未知。我们没有人力与物力去统计全国每个人的身高，但是可以通过采样，获取部分人的身高，然后通过最大似然估计来获取上述假设中的正态分布的均值与方差。

1. 身高服从正态分布 $f(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，分布参数为 $\theta=(\mu,\sigma^2)$，所以似然函数如下所示：

$$f(x_1,\ldots,x_n\,|\,\theta)=f(x_1\,|\,\theta)*\ldots,f(x_n\,|\,\theta)$$

$$L(\theta\,|\,x_1,\ldots,x_n)=f(x_1,\ldots,x_n\,|\,\theta)=\prod_{i=1}^{n}f(x_i,\theta)$$

2. 对似然函数取对数得到如下形式：

$$\ell(\theta \mid x_1,\ldots,x_n) = \log[L(\theta \mid x_1,\ldots,x_n)] = \log[\prod_{i=1}^{n} f(x_i,\theta)] = \sum_{i=1}^{n} \log f(x_i,\theta)$$

3. 对分布参数分别求偏导

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} x_i - n\mu$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_{i=1}^{n} (x_i - \mu)^2 - n\sigma^2$$

上面求偏导前面的常数省略了，因为这不影响第四步的求解。

4. 解释然方程，令第三步得到的偏导为零，可以得到最优的 μ 和 σ。

$$\mu = \frac{\sum_{i=1}^{n} x_i}{\mathrm{n}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

3. 工程文件简介

src 文件夹下是源代码：

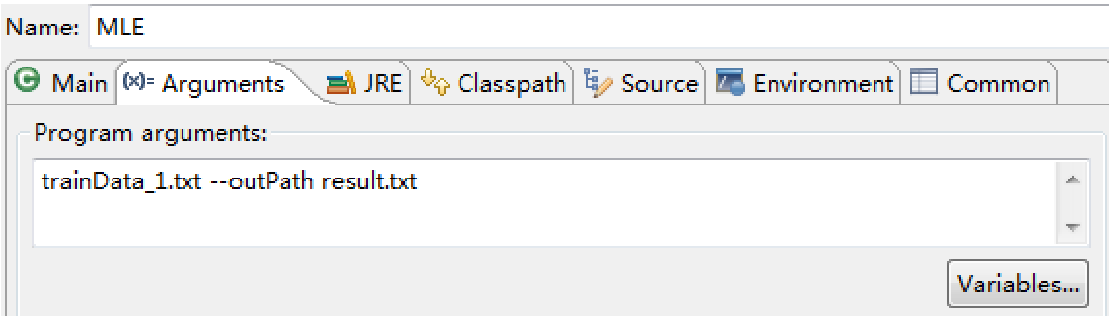
MLEDataProc.java => 读取训练数据。

MLEOption.java => 参数类，该 MLE 算法支持的参数。

MLE.java => 求解 μ 和 σ。

4. 程序调用方式

- 在 **windows** 下可以用 eclipse 打开该工程，在 Run =>Run Configurations 界面设置程序需要的参数，参数细节在后面会给出解释。



- 在 **Linux** 下可以可以通过如下命令进行调用。

  把文件夹下的 MLE.jar 拷贝到 linux 下某个目录，通过如下命令进行调用。

  java –jar MLE.jar trainDataFilePath [--outPath outPathValue]

方括号中表示可选参数。

trainDataFilePath：表示训练数据文件路径

outPath：把分布参数最优值输出到文件中。如果不传，默认为空，则输出到终端。

5. 数据集

在文件夹下有一个数据集文件 trainData_1.txt。该文件记录了一部分人的身高。

程序的时间复杂度是 $O(n)$,其中 n 表示数据集的大小。