Mathematica 应用与经验系列_28

2015年11月

中科院软件中心 Mathematica 产品与服务部

邮箱: mathematica@sec.ac.cn
电话: 86-10-82622887
手机: 15011518369（软件/培训咨询）王树志
13552986975（软件/培训咨询）陶冕
18518414242（应用工程师）李想

星期六, 2015.11.30

Comment 如果您希望用Mathematica的最新版本解决当前问题, 请联系我们. 我们会竭诚与您一道探讨, 并提供解决方案.

FindDistribution & FindFormula - 探索数据背后的结构

FindDistribution 和 FindFormula 根据数据拟合出（但不是过度拟合）的相应的公式模型，这两个函数使用都是非常非常简单来应用的.

示例: 对于给出的文本, 出现的单词符合哪种分布. 以维基百科中伦敦条目为例进行分析

```
text = WikipediaData["London"];
words = TextCases[DeleteStopwords[text], "Word"];
```

统计出现的词频

```
wordCount = Tally[words][[All, 2]];
```

```
Reverse[Sort@wordCount][[;; 50]]
```

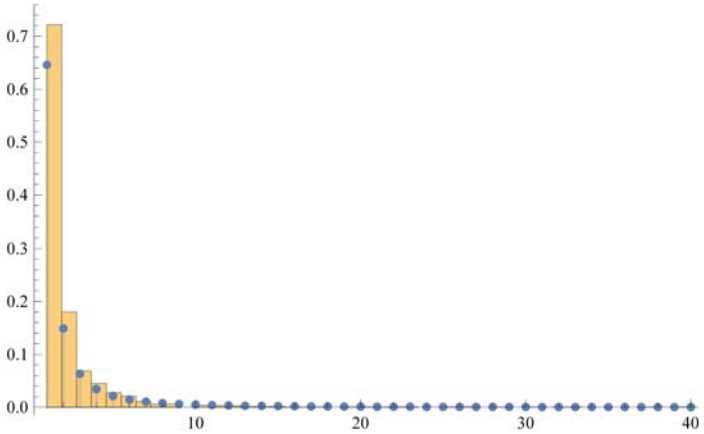
```
{461, 74, 52, 44, 43, 39, 39, 37, 33, 32, 30, 28, 27, 27, 27, 26, 26,
  23, 23, 22, 22, 21, 20, 20, 19, 19, 18, 18, 17, 16, 15, 15, 15, 15,
  14, 14, 14, 13, 13, 13, 13, 12, 12, 12, 12, 12, 12, 12, 12, 11}
```

用 FindDistribution 找出文本符合参数 ρ 为1.116 的齐夫分布

```
dist = FindDistribution[wordCount]
ZipfDistribution[1.116]
```

可以绘制出原始数据与相应的概率密度直方图

```
Show[
  Histogram[wordCount, {1, 40, 0.9}, "ProbabilityDensity"], DiscretePlot[
    PDF[dist, x], {x, 1, 40}, PlotStyle → PointSize[Medium], PlotRange → All]]
```



示例: 女性身高和体重

查看身高和体重的单位, 然后将数据导入

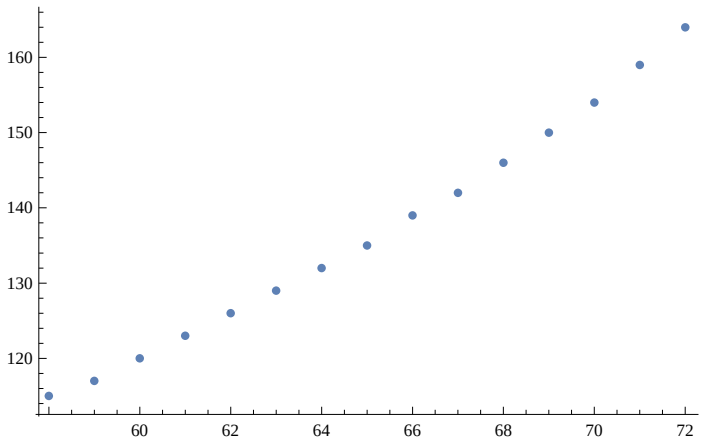
```
ExampleData[{"Statistics", "FemaleHeightsAndWeights"},
    "ColumnDescriptions"] // Column
data = ExampleData[{"Statistics", "FemaleHeightsAndWeights"}]
```

```
Height in inches.
Weight in pounds.
```

```
{{58, 115}, {59, 117}, {60, 120}, {61, 123}, {62, 126},
  {63, 129}, {64, 132}, {65, 135}, {66, 139}, {67, 142},
  {68, 146}, {69, 150}, {70, 154}, {71, 159}, {72, 164}}
```

下一步, 根据数据绘制图形

```
ListPlot[data]
```



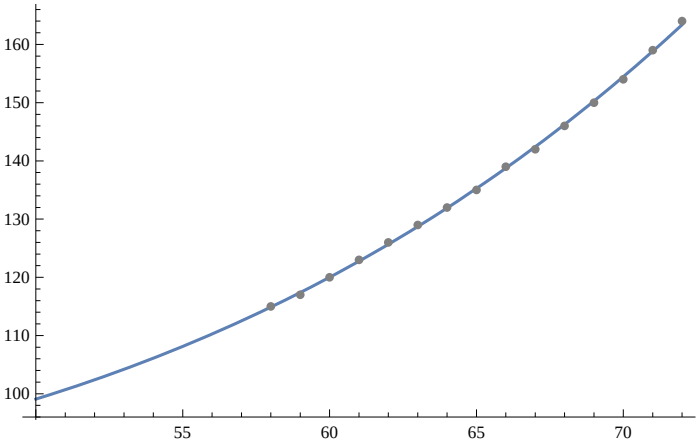
直接将数据送入软件即可

```
fit = FindFormula[data]
```

$$79.6096 + 3.11735 \times 10^{-6}\, \#1^4 \,\&$$

将原始数据跟拟合函数绘制出

```
Show[Plot[fit[x], {x, 50, 72}], ListPlot[data, PlotStyle → Gray]]
```



计算出当女性身高为 80 英寸时候其相应体重为多少?

```
fit[80]
```

```
207.296
```

女高音歌手体重的符合哪类分布? - 身高数据我们给出

```
sopranos = {64, 62, 66, 65, 60, 61, 65, 66, 65, 63, 67, 65, 62, 65, 68,
      65, 63, 65, 62, 65, 66, 62, 65, 63, 65, 66, 65, 62, 65, 66, 65, 61,
      65, 66, 65, 62, 63, 67, 60, 67, 66, 62, 65, 62, 61, 62, 66, 60, 65,
      65, 61, 64, 68, 64, 63, 62, 64, 62, 64, 65, 60, 65, 70, 63, 67, 66};
```

```
weightss = fit[sopranos];
```

数据符合均值为 132.805, 标准差为 7.6619 的正态分布

```
FindDistribution[weightss]
```

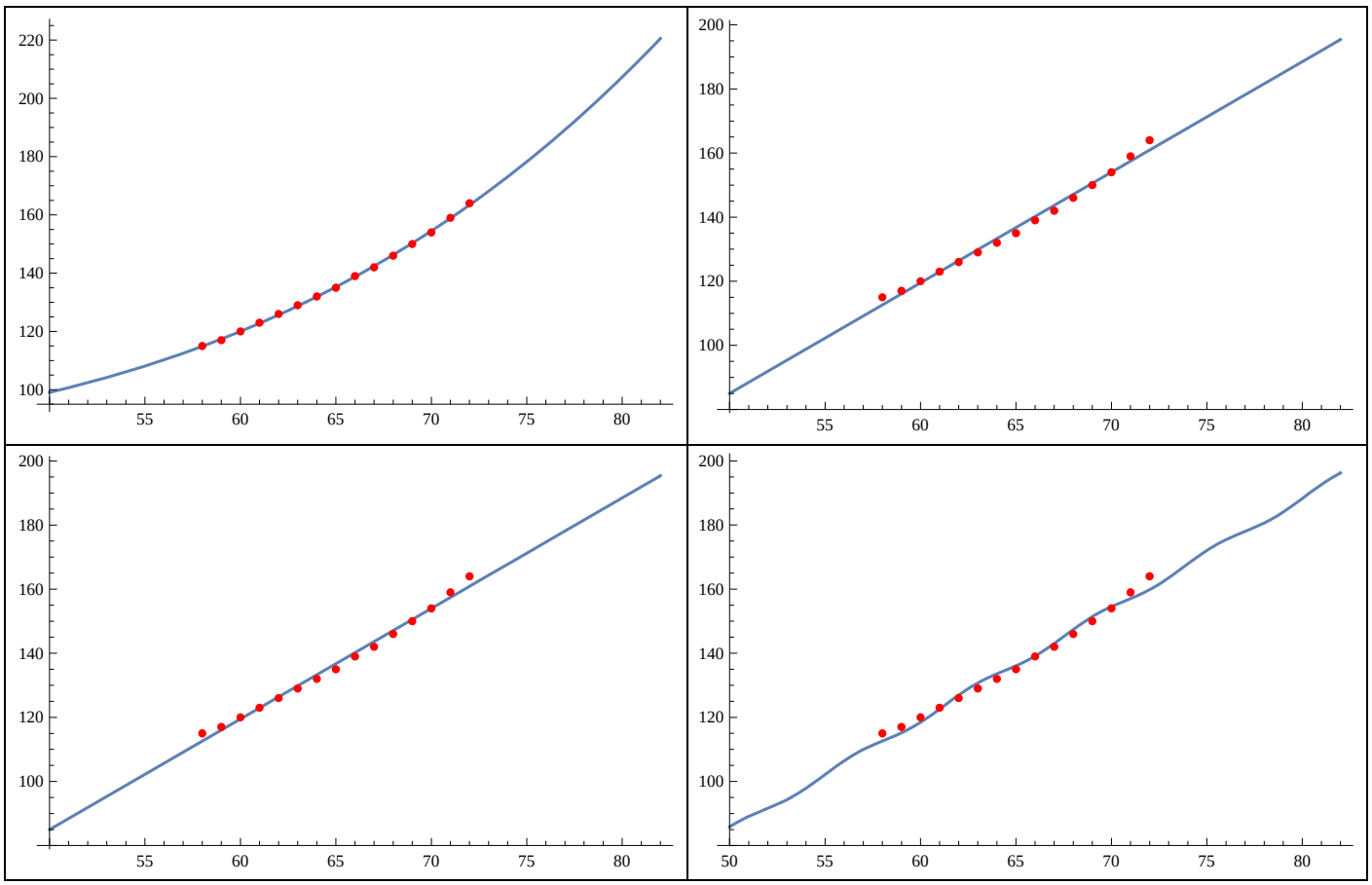
```
NormalDistribution[132.805, 7.6619]
```

用来看查看更多的结果, 这里找出 4 个拟合公式, 指定算法为并行退火

```
SeedRandom[989]
more = FindFormula[data, x, 4, Method → "ParallelTempering"]
```

$$\left\{79.6096 + 3.11735 \times 10^{-6}\, x^4,\ -87.5167 + 3.45\, x,\right.$$
$$\left.-87.4824 + 3.45\, x,\ -87.6094 + 3.45\, x + \text{Cos}[x]\right\}$$

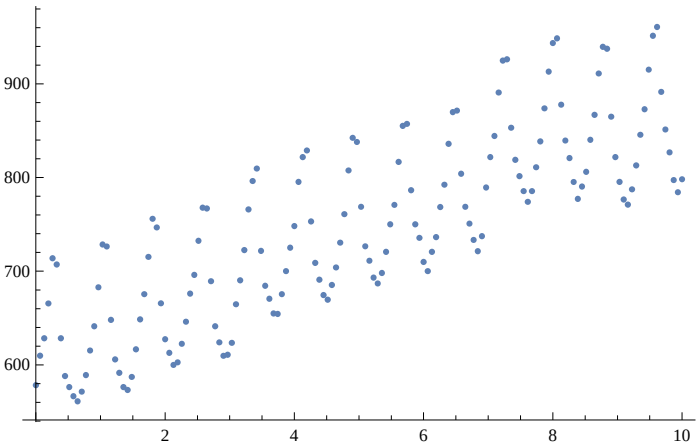
```
Multicolumn[Table[Show[Plot[i, {x, 50, 82}], ListPlot[data, PlotStyle → Red],
      ImageSize → 350 ], {i, more}], 2, Frame → All]
```



示例: 数据中应用超越函数

我们已经有厂家提供额13年中每个月奶牛产奶的数据, 现在将数据绘制出来

```
ListPlot[data]
```

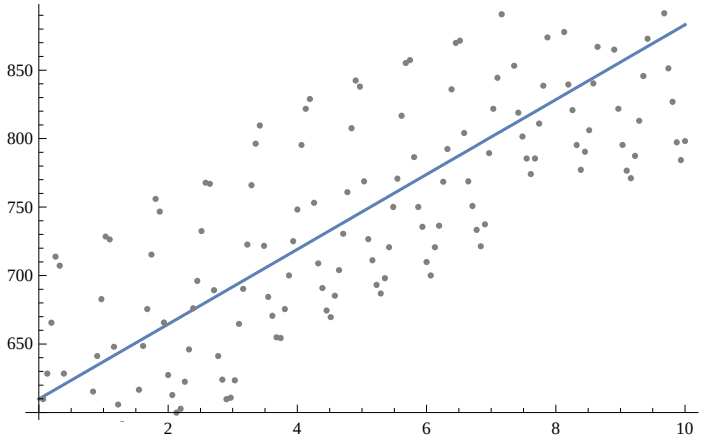


如果只是简单地应用数据, 只是能找到一个线性的模型, 显然是不足够好的

```
fit1 = FindFormula[data, x]
```

609.821 + 27.3341 x

```
Show[Plot[fit1, {x, 0, 10}], ListPlot[data, PlotStyle → Gray]]
```



我们可以给出目标模型中还有的一些算符, 如周期性函数等

```
fit2 = FindFormula[data, x, TargetFunctions → {Plus, Times, Sin, Cos}]
```

$$605.964 + 27.9547\,x + 73.237\,\text{Sin}[8.0\,x]$$

```
Show[Plot[{fit1, fit2}, {x, 0, 10}], ListPlot[data, PlotStyle → Gray]]
```