python_backup

pymc

医药数据统计分析项目：QQ231469242

http://hao.jobbole.com/pymc/

PyMC是一个实现贝叶斯统计模型和马尔科夫链蒙塔卡洛采样工具拟合算法的Python库。PyMC的灵活性及可扩展性使得它能够适用于解决各种问题。除了包含核心采样功能，PyMC还包含了统计输出、绘图、拟合优度检验和收敛性诊断等方法。

特性

PyMC使得贝叶斯分析尽可能更加容易。以下是一些PyMC库的特性：

- 用马尔科夫链蒙特卡洛算法和其他算法来拟合贝叶斯统计分析模型。
- 包含了大范围的常用统计分布。
- 尽可能地使用了NumPy的一些功能。
- 包括一个高斯建模过程的模块。
- 采样循环可以被暂停和手动调整，或者保存和重新启动。

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

python(20)
statistics(11)
自然语言(2)
nltk(2)
癌症(2)
编码(1)
分子式(1)
分子重量(1)
机器人(1)
美国在研新药(1)
更多

随笔分类

excel操作(1)
Json and API(1)
navicat/mysql(3)
numpy(1)
porn
pyautogui
pyinstaller(2)
pypdf2=(8)

- 创建包括表格和图表的摘要说明。
- 算法跟踪记录可以保存为纯文本，pickles，SQLite或MySQL数据库文档或HDF5文档。
- 提供了一些收敛性诊断方法。
- 可扩展性：引入自定义的步骤方法和非常规的概率分布。
- MCMC循环可以嵌入在较大的程序中，结果可以使用Python进行分析。

使用

首先，在文件中定义你的模型，并命名为mymodel.py

```
# -*- coding: utf-8 -*-
"""
Created on Mon Jul 24 10:56:07 2017

@author: toby
"""

# Import relevant modules
import pymc
import numpy as np

# Some data
n = 5*np.ones(4,dtype=int)
x = np.array([-.86,-.3,-.05,.73])

```


随笔档案

```
# Priors on unknown parameters
alpha = pymc.Normal('alpha',mu=0,tau=.01)
beta = pymc.Normal('beta',mu=0,tau=.01)

# Arbitrary deterministic function of parameters
@pymc.deterministic
def theta(a=alpha, b=beta):
    """theta = logit^{-1}(a+b)"""
    return pymc.invlogit(a+b*x)

# Binomial likelihood for data
d = pymc.Binomial('d', n=n, p=theta, value=np.array([0.,1.,3.,5.]),\
                  observed=True)
```

测试脚本

```
# -*- coding: utf-8 -*-
"""
Created on Mon Jul 24 11:21:23 2017

@author: toby
"""

import pymc
import mymodel

S = pymc.MCMC(mymodel, db='pickle')
S.sample(iter=10000, burn=5000, thin=2)
pymc.Matplot.plot(S)
```

2016年3月 (3)
2016年2月 (6)

最新评论

1. Re:opencv_判断两张图片是否相同
@云敛晴空用Python3...
--python_backup
2. Re:bootstrap
顶一个
--今天我过生
3. Re:opencv_判断两张图片是否相同
你好，我直接用你的代码，出现这样的提示：Traceback (most recent call last): File "C:\Python27\test.py", line 17, in .....
--云敛晴空
4. Re:nltk.download()出错解决
合集的压缩包链接不能用啦 方便再给一下链接吗 然后下完之后直接压缩和解压的都放nltk_data底下不行吗
--Dragon5

阅读排行榜

1. PyQt 5.4参考指南 ---- PyQt5和PyQt4之间的差异(3281)
2. 自然语言11_情感分析(2327)
3. nltk.download()出错解决(1915)

这个例子会产生10000个后验样本。这个样本会存储在Python序列化数据库中。

教程示例

教程会指导用户完成常见的PyMC应用。

如何用MCMC来拟合模型

PyMC提供了一些可以拟合概率模型的方法。最主要的拟合模型方法是MCMC，即马尔科夫蒙特卡洛算法。生成一个MCMC对象来处理我们的模型，导入disaster_model.py并将其作为MCMC的参数。

调用MCMC中的sample()方法（或者交互采样函数isample())来运行采样器

```
# -*- coding: utf-8 -*-
"""
Created on Mon Jul 24 11:26:27 2017

@author: toby
"""

from pymc.examples import disaster_model
from pymc import MCMC
M = MCMC(disaster_model)
M.sample(iter=10000, burn=1000, thin=10)
```

等待几秒钟后，便可以看到采样过程执行完成，模型已经完成拟合。

4. 完美配置Python3.5+Anaconda+PyQt5，实现UI和其他模块的结合(1904)
5. 自然语言18.2_NLTK命名实体识别(1554)

评论排行榜

1. opencv_判断两张图片是否相同(2)
2. bootstrap(1)
3. nltk.download()出错解决(1)

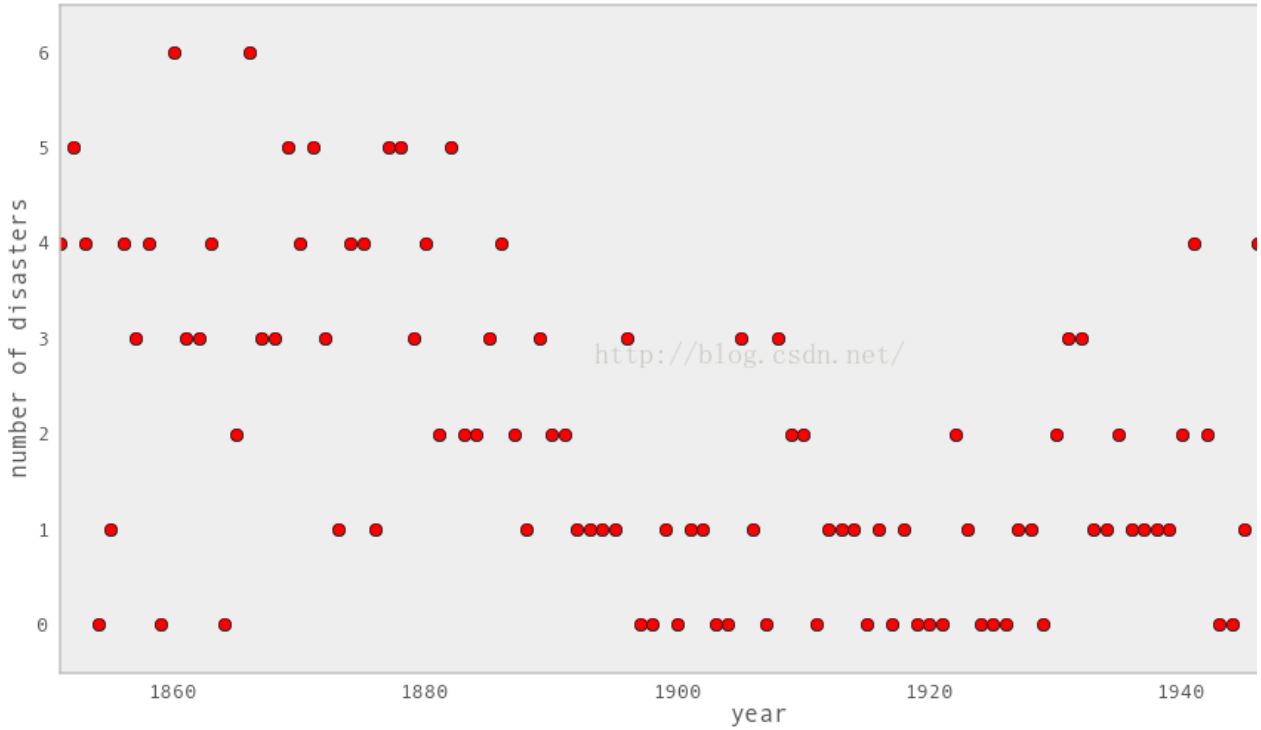
推荐排行榜

1. bootstrap(1)

http://blog.csdn.net/dmsgames/article/details/52525636

1、一个统计模型

有这样一个数据集，它按照时间顺序，收录了英国从1851年到1962年每年的矿难发生次数。如下图所示：



我们可以假设，矿难发生的概率服从一个Poisson过程，在某一年泊松过程的参数发生了变化，在时间轴的早些时候，矿难发生的概率较高，后来矿难发生的概率比较低。

我们将上述概念模型转化为统计模型：

$$(D_t|s,e,l) \sim \text{Poisson}(r_t), \quad r_t = \begin{cases} e & \text{if} \quad t < s \\ l & \text{if} \quad t \geq s \end{cases}, \quad t \in [t_l, t_h]$$

$$s \sim \text{Discrete Uniform}(t_l, t_h)$$

$$e \sim \text{Exponential}(r_e)$$

$$l \sim \text{Exponential}(r_l)$$

以上模型参数定义如下：

- D_t: 第t年矿难发生的次数；
- r_t: 第t年Posson过程的参数；
- s: 泊松过程参数发生改变的那一年；
- e: 第s年之前，泊松过程的参数；
- l:第s年之后，泊松过程的参数；
- t_l,t_h: 年份t的下限和上限；
- r_e,r_l:e和l的先验分布

由于在模型中我们定义了D依赖于s,e,l，所以我们把D称作s,e,l的子变量，类似的，s,e,l称为D的父变量。

2、变量的两种类型

PyMC包中定义类两种随机变量类型，分别为stochastic和Deterministic。

模型中唯一的Deterministic变量是r，因为当我们知道r的父参数（s,l,e）后，我们可以准确地计算出r的值。

另一方面，s，D（在观察到数据之前）是stochastic变量，因为即使观察到他们的父变量，任然不能确定它们的值。

我们将模型写在一个名为 disaster_model.py 的Python脚本中：

```
"""
导入numpy和pymc
"""
from pymc import DiscreteUniform, Exponential, deterministic, Poisson, Uniform
import numpy as np
"""
导入英国矿难数据集
"""
disasters_array = \
np.array([ 4, 5, 4, 0, 1, 4, 3, 4, 0, 6, 3, 3, 4, 0, 2, 6,
3, 3, 5, 4, 5, 3, 1, 4, 4, 1, 5, 5, 3, 4, 2, 5,
2, 2, 3, 4, 2, 1, 3, 2, 2, 1, 1, 1, 1, 3, 0, 0,
1, 0, 1, 1, 0, 0, 3, 1, 0, 3, 2, 2, 0, 1, 1, 1,
0, 1, 0, 1, 0, 0, 0, 2, 1, 0, 0, 0, 1, 1, 0, 2,
3, 3, 1, 1, 2, 1, 1, 1, 1, 2, 4, 2, 0, 0, 1, 4,
0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1])

"""
定义转折点s：
取值范围0-110
均匀离散分布
"""
switchpoint = DiscreteUniform('switchpoint', lower=0, upper=110, doc='Switchpoint[year]')

"""
定义e、l
指数分布
"""
early_mean = Exponential('early_mean', beta=1.)
late_mean = Exponential('late_mean', beta=1.)

"""
```

```
定义r
"""
@deterministic(plot=False)
def rate(s=switchpoint, e=early_mean, l=late_mean):
''' Concatenate Poisson means '''
out = np.empty(len(disasters_array))
out[:s] = e
out[s:] = l
return out

"""
定义矿难发生次数
服从泊松分布
"""
disasters = Poisson('disasters', mu=rate, value=disasters_array, observed=True)
```

来自CODE的代码片

snippet_file_0.txt

3、父变量与子变量

我们已经使用PyMC创建了统计模型，PyMC中提供方法查看模型中参数之间的关系，试例代码如下：

```
from pymc.examples import disaster_model
disaster_model.switchpoint.parents #显示s的父参数
#输出{'lower': 0, 'upper': 110}
disaster_model.disasters.parents #显示disasters的父参数
#输出{'mu': <pymc.PyMCObjects.Deterministic 'rate' at 0x000000000B791BE0>}
disaster_model.rate.children #显示rate的子参数
#输出{<pymc.distributions.new_dist_class.<locals>.new_class 'disasters' at 0x000000000B791C18>}
```

来自CODE的代码片

snippet_file_0.txt

4、变量的值

所有的PyMC变量都具有value属性，查看value值示例代码如下：

```
disaster_model.disasters.value
"""输出
array([4, 5, 4, 0, 1, 4, 3, 4, 0, 6, 3, 3, 4, 0, 2, 6, 3, 3, 5, 4, 5, 3, 1,
4, 4, 1, 5, 5, 3, 4, 2, 5, 2, 2, 3, 4, 2, 1, 3, 2, 2, 1, 1, 1, 1, 3,
0, 0, 1, 0, 1, 1, 0, 0, 3, 1, 0, 3, 2, 2, 0, 1, 1, 1, 0, 1, 0, 1, 0,
0, 0, 2, 1, 0, 0, 0, 1, 1, 0, 2, 3, 3, 1, 1, 2, 1, 1, 1, 1, 2, 4, 2,
0, 0, 1, 4, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1])
"""
disaster_model.switchpoint.value
#输出 array(40)

disaster_model.early_mean.value
#输出 array(1.1444157379406001)

disaster_model.late_mean.value
#输出 array(0.027985496189503425)
```

来自CODE的代码片

snippet_file_0.txt

5、使用马尔科夫链蒙特卡洛（MCMC）拟合模型

PyMC提供MCMC方法拟合模型，使用方法如下：

```
from pymc.examples import disaster_model
from pymc import MCMC
M = MCMC(disaster_model)
M.sample(iter=10000, burn=1000, thin=10)
```

来自CODE的代码片

snippet_file_0.txt

MCMC算法输出模型中每个变量的样本，获得样本方法如下：

```
M.trace('switchpoint')[:]
#输出array([43,43,44,...44,44])
```

来自CODE的代码片

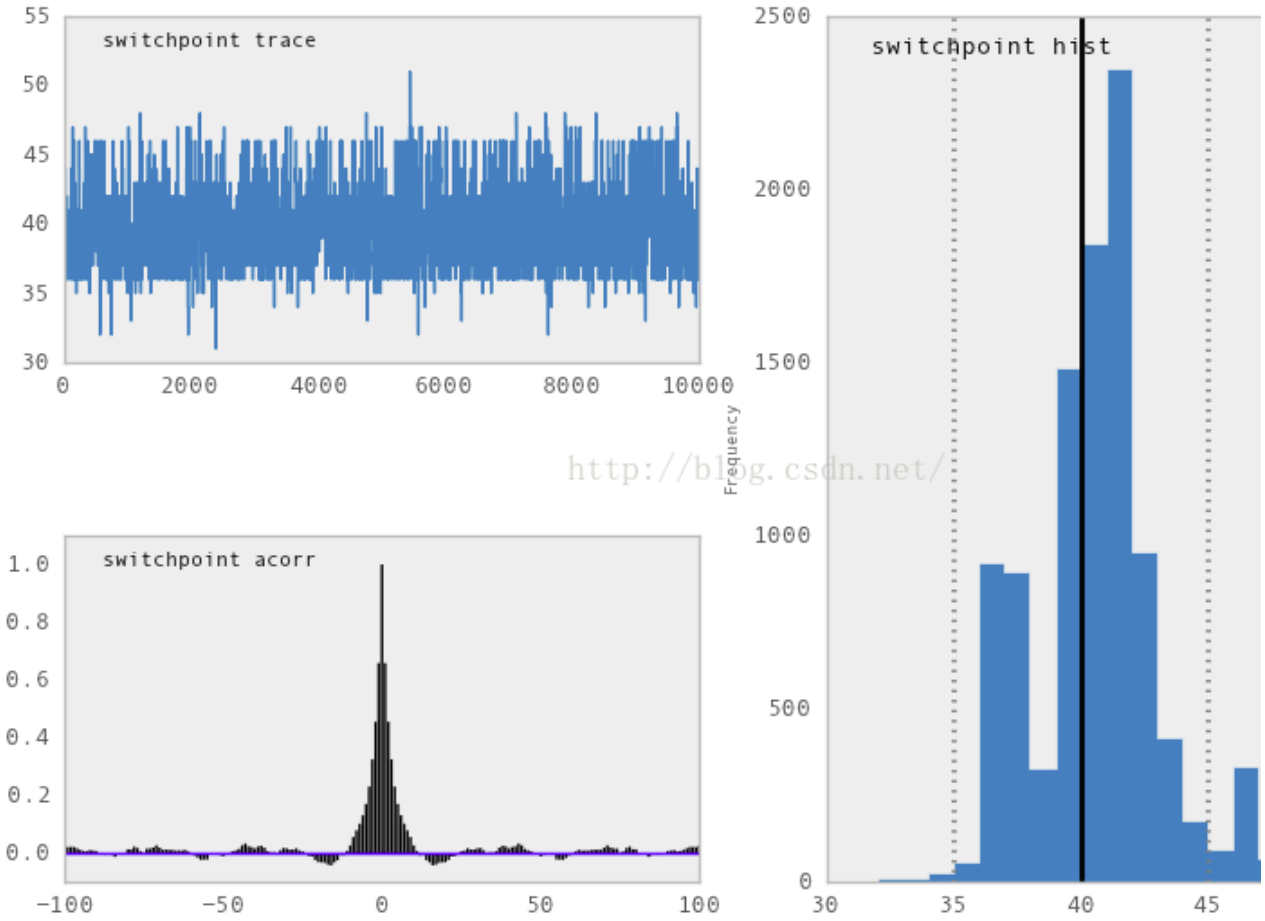
snippet_file_0.txt

画出每个变量的采样序列图、后验边缘分布直方图、自相关性图，代码如下：

```
from pymc.Matplot import plot
plot(M)
```

来自CODE的代码片

snippet_file_0.txt



采样序列图可以判断MCMC是否收敛，如果采样序列分布近似于白噪声，那么可以判断MCMC已经收敛。

分类: 统计

python_backup
关注 - 4
粉丝 - 12

0 0

+加关注

« 上一篇：Logistic Ordinal Regression
» 下一篇：PDF文本内容批量提取到Excel

posted @ 2017-07-24 11:35 python_backup 阅读(12) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

最新IT新闻:

· 再见了！Windows 10 Build 10586明起彻底停更
· 趣店10月18日在纽交所上市 预计融资7.69亿美元
· 支付宝小程序编程大赛在即：程序员鼓励师曝光
· 小米华东总部落户南京 未来将成为手机研发中心
· 京东建成全球首个全流程无人仓
» 更多新闻...

最新知识库文章:

· 实用VPC虚拟私有云设计原则
· 如何阅读计算机科学类的书
· Google 及其云智慧
· 做到这一点，你也可以成为优秀的程序员
· 写给立志做码农的大学生
» 更多知识库文章...