API Usage

jidong

2022 年 8 月 17 日

1 Model API

模型名称：**DonutX**

1.1 创建模型

DonutX 的创建的参数及其含义如下：

参数名	默认值	作用
max_epoch	150(int)	训练轮数
batch_size	128(int)	一次投入网络的数据量
network_size	[100,100](list[int])	编码器和解码器网络维数
latent_dims	8(int)	隐变量的维数
window_size	120(int)	窗长
cuda	True(bool)	是否使用 gpu
condition_dropout_left_rate	0.9(double)	丢去多少比例的条件变量 (防止过拟合)
print_fn	print	(一般不需要更改)

表 1: 初始化参数

如果需要保存模型参数，可以直接访问 DonutX 实例中的 _model 属性并用 torch 工具保存:torch.save(model._model.state_dict(),'model.pth')。载入模型则先实例化 DonutX，然后读入模型参数:torch.save(model._model.state_dict(),'model.pth')。

1.2 模型训练

API:fit

输出: 无输出，仅训练模型。

参数名	默认值	作用
kpi	无默认值 (KPISeries)	训练数据
valid_kpi	None(KPISeries)	验证数据 (实际生产环境可以不给)

表 2: fit 函数的参数

1.3 模型预测

API:predict

参数名	默认值	作用
kpi	无默认值 (KPISeries)	预测数据
return_statistics	False(bool)	是否返回 kpi 的平均值和标准差
indicator_name	indicator(string)	无需更改

表 3: predict 函数的参数

输出: 预测得到的重建误差，重建误差越大，越可能是异常点。由于有窗长的存在，所以一个序列的前窗长个点没有预测值，因此需采集到窗长个点后才能开始进行预测，涉及到冷启动的问题。可以考虑缩短窗长，或者在前面补 0。

2 Dataset API

2.1 初始化数据集

参数名	默认值	作用
value	无 (序列)	kpi 序列值
timestamp	无 (序列)	kpi 序列对应的时间戳
label	None(序列)	kpi 序列是否异常 (1 为异常,0 为正常)，可以不给
missing	None	缺失点，可以不给
name	默认用 uuid4 生成	数据集名称，可以不给
normalized	False(bool)	是否正则化数据

表 4: 数据集初始化参数

2.2 数据集 API

API:normalize

作用: 正则化数据

参数名	默认值	作用
mean	None	按照所给均值正则化，若不给则用数据的均值
std	None	按照所给标准差正则化，若不给则用数据的标准差
return_statistic	False	是否返回统计量

表 5: normalize 参数

API:split

作用: 分割数据集，一般不需要

API:label_sampling

作用: 丢去 $(1 - sampling\_rate)$ 比例的 label，若在初始化数据集时不给

label，则不需要用这个函数。

3 工作流程

训练流程

1. 从数据库读取数据
2. 预处理数据:
   - 整理数据格式
   - 生成 KPISeries，并正则化
3. 创建模型
4. 用 fit 函数训练模型
5. 保存模型

训练流程可以定期取前一段时间的数据进行训练，具体时间可以自行安排。

预测流程

1. 取大于窗长长度的数据序列
2. 对数据进行预处理
3. 创建模型
4. 读入之前保存的模型参数
5. 预测数据，得到重建误差
6. 用自己制定的或者简单的算法得到的阈值进行检测
7. 认为大于阈值的点是异常点，发出预警

4 可能遇到的问题

Q: 如果遇到丢失点怎么办?

A:KPISeries 在创建过程中会自动检测两个时间戳之间的最小间隔，并按此间隔对缺失时间点补 0。