**关于血液透析的相关理解**

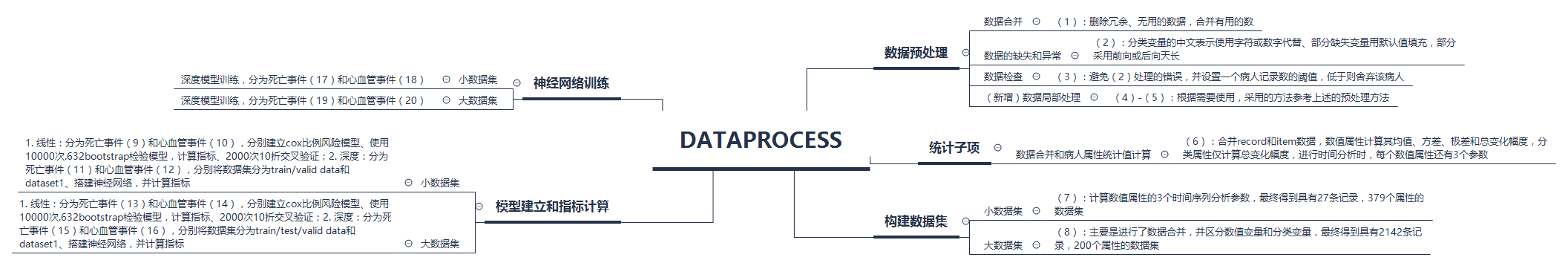
# 1 背景简要介绍

目前，我国患终末期肾脏病（ESRD，end-stage renal disease）的大部分病人的首选医治方案是维持性血液透析（MHD）。鉴于这个现实因素，医学上提出了一个重要的课题“提高血液透析患者总体存活率，进一步改善患者生存质量”。而近年来探讨影响患者生存相关因素主要集中在两个方面【1,2,3,4】，一是透析参数，二是患者自身因素。面对上述的医学数据，目前流行的是采用以Cox比例风险回归模型（简称Cox模型）【5】为主的生存分析技术。基于Cox模型，衍生出多种改进的模型，但依然保持着一致的基本假设，即患者的死亡风险是患者属性线性组合的某种关系，这个假设也称为线性比例风险假设【6】。

由于这一基本假设过于理想，不符合现实需求，因此需要一个更好的方法来满足实际需求。一个可实现的方法是：运用深度学习的技术结合生存分析来改善Cox模型，挖掘影响血液透析患者生存率的相关因素。事实上，神经网络结构是线性组合和非线性组合的结合，当输入患者属性与患者死亡风险组合为某种线性关系，再经过非线性转换得到输出，这是否在一定程度上更能表达患者属性与患者死亡风险之间的关系？此外，国外关于深度学习和生存分析相结合的理论正在不断发展，利用该方法对ESRD患者进行血液透析研究是具有一定的实际意义，也是值得探究的。

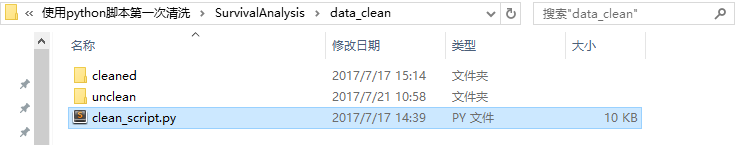
# 2 代码详解

在周泽栩师兄的代码（第一版）文件中，主要分为三个部分，分别是数据预处理、数据处理、数据可视化。由于新给的岭南院区已经做过前期的预处理，故某些处理与第一版代码有稍许差异。代码的组织结构如下所示。

[](dataprocess.xmind)

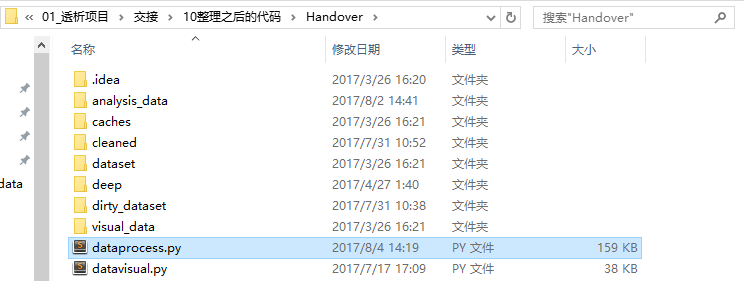
## 2.1 python清洗

clean\_script.py主要是将gb2312编码重新编码为utf-8，并且删除无用的空白符、换行符以及回车符。文件所处如下的结构。

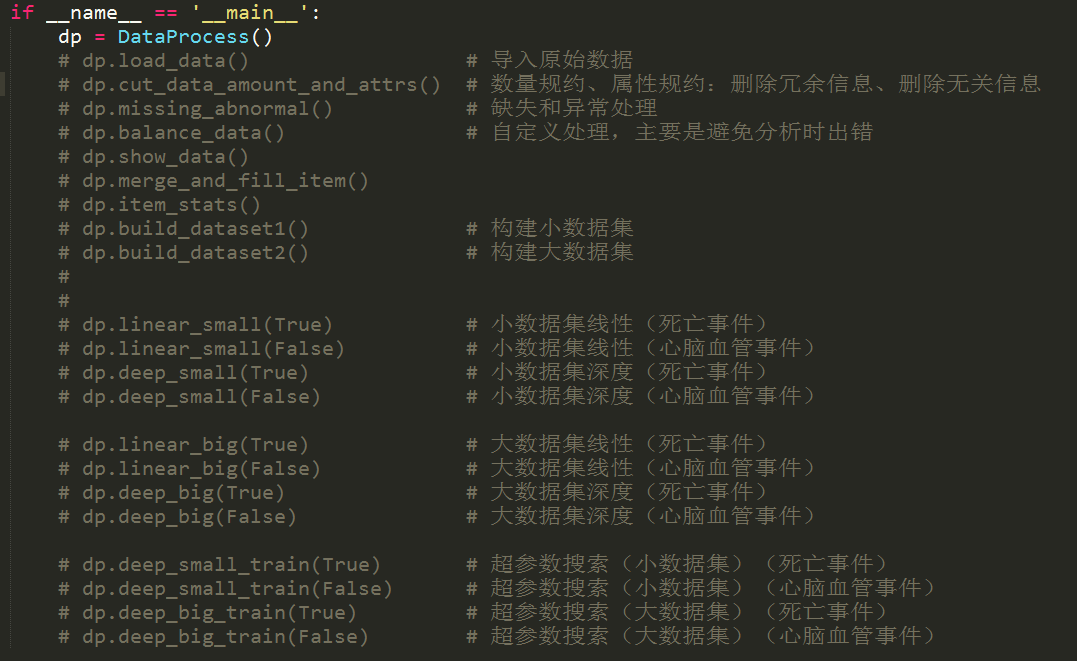


## 2.2 数据处理

dataprocess.py是本项目的最重要的文件，需着重理解。文件所处如下的结构。



dataprocess.py代码结构如下所示，主函数显示了dataprocess.py的逻辑处理结构。



（1）cut\_data\_amount\_and\_attrs()

这个函数的主要功能是通过merge连接patient.csv和LGalldata.xls两个文件，并保证其他文件的patient\_id与patient.csv的一致。

（2）missing\_abnormal()

这个函数分别调用了三个处理缺失和异常的函数。建议对patient.csv、item.csv、record.csv三个文件分别执行该函数。在对item.csv、record.csv处理中，对缺失值采用了先前向填充再后向填充的方法。

（3）balance\_data()

该函数考虑了分类标签为字符串的字段的分类值数量全部为0的情况，也就是用字符串代替的分类变量操作失败，以避免missing\_abnormal()的处理错误，同时保证不分析记录数过少的病人。

（4）show\_data()（新增函数）

这个函数可查看.csv文件的每个字段为空的情况。

（5）merge\_and\_fill\_item()（新增函数）

根据第一版的处理过程的文件结构，对item.csv添加patient\_id和tx\_date两个字段，并使用先前向填充再后向填充的方法对item进行缺失值填充。

（6）item\_stats()

这个函数是统计子项，同时把record.csv、item.csv连接起来。get\_value()计算item部分字段的均值、方差、极差和总变化幅度。注意保存为.h5文件。这个函数最后得到recorditem.csv文件。

（注：分类属性（如科别、治疗项目等）参与时间序列分析只需计算1个总变化幅度属性）

（7）build\_dataset1()

这个函数对patient.csv和recorditem.csv连接起来，最终得到的文件是每个病人对应一条记录，也就是小数据集。（即如果有27人，则有27条记录）

为了进行时间序列分析，对数值属性的字段计算均值、方差、极差和总变化幅度，和时间序列分析的p、d、q，即该属性计算7个参与时间序列分析的值。get\_value()计算recorditem部分字段。同时，定义哪些属性可进行时间序列，通过get\_arima()得到该属性p、d、q值。最后，将get\_value()和get\_arima()得到的统计量特征和时间特征值赋给patient.csv，得到一个新的数据集dataset1。

【疑问】

* 类似'vein\_change'和'membrane\_change'，在（6）中计算的得到的数值，为什么在这里还要计算该值的均值、方差、极差和总变化幅度。
* get\_value()和get\_arima()的属性不完全一样。

（8）build\_dataset2()

这个函数先进行了一些字段的map操作，再将patient.csv和patient\_date.csv连接（merge），起来，之后再与recorditem.csv连接起来，并将生存时间的单位转换为月。同时，该函数还区分数值变量和分类变量，最后得到dataset2，也就是大数据集。

（9）linear\_small(True)

该函数为小数据集（dataset1）线性（死亡事件），分为以下步骤：对标称变量（即分类表里）进行哑变量处理为数值属性、死亡事件和心血管事件的选择、cox比例风险函数（CoxPHFitter()）挑选符合统计学意义的属性、10000次.632 bootstrap来检验预测模型的准确性，防止模型过拟合，并计算concordance\_index、2000次10折交叉验证来验证模型的稳定。

【注意】

* 在cox.fit()时，得到的p值均大于0.05，即无统计学意义。由于病人数27个远小于属性数379导致的。
* Bootstrap用于小数据集、而2000次10折交叉验证用于大数据集。

（10）linear\_small(False)

该函数为小数据集（dataset1）线性（心脑血管事件），具体步骤与（9）一致。

（11）deep\_small(True)

该函数为小数据集（dataset1）深度（死亡事件），分为以下步骤：对标称变量（即分类变量）进行哑变量处理为数值属性、死亡事件和心血管事件的选择、数据集分train\_data、valid\_data 以及dataset1，神经网络搭建（DeepSurv()），并计算concordance\_index、保存相应的数据。

（12）deep \_small(False)

该函数为小数据集（dataset1）深度（心脑血管事件），具体步骤与（11）一致。

（13）linear\_big(True)

该函数是大数据集（dataset2）线性（死亡事件），分为以下步骤：对标称变量（即分类表里）进行哑变量处理为数值属性、死亡事件和心血管事件的选择、cox比例风险函数（CoxPHFitter()）挑选符合统计学意义的属性、10000次.632 bootstrap来检验预测模型的准确性，防止模型过拟合，并计算concordance\_index、2000次10折交叉验证来验证模型的稳定。

（14）linear\_big(False)

该函数为大数据集（dataset2）线性（心脑血管事件），具体步骤与（13）一致。

（15）deep\_big(True)

该函数为大数据集（dataset2）深度（死亡事件），分为以下步骤：对标称变量（即分类变量）进行哑变量处理为数值属性、死亡事件和心血管事件的选择、数据集分train\_data、valid\_data 、test\_data以及dataset2，神经网络搭建（DeepSurv()），并计算concordance\_index、保存相应的数据。

（16）deep \_big(False)

该函数为大数据集（dataset2）深度（心脑血管事件），具体步骤与（15）一致。

（17）deep\_small\_train(True)

小数据集（dataset1）深度（死亡事件）训练，使用网格搜索训练。神经网络层数1-3，神经元在[30, 50, 100, 150, 200]取值，学习率在[1e-4, 1e-5, 1e-6]取值，将数据集分为train\_data、valid\_data和all\_set分别进行训练、验证，并计算get\_concordance\_index。

（18）deep\_small\_train(False)

小数据集（dataset1）深度（心脑血管事件）训练，与（17）步骤一致。

（19）deep\_big\_train(True)

大数据集（dataset2）深度（死亡事件）训练，使用网格搜索训练。神经网络层数1-3，神经元在[30, 50, 100, 150, 200]取值，学习率在[1e-4, 1e-5, 1e-6]取值，batch\_size在[5000, 2000, 1000, 500, 256, 100, 50, 20, 10, 5, 2, 1]取值，将数据集分为train\_data、valid\_data、test\_data和all\_set，训练完成后，并计算get\_concordance\_index。

（20）deep\_big\_train(False)

big数据集（dataset1）深度（心脑血管事件）训练，与（19）步骤一致。

## 2.3 数据可视化

数据可视化datavisual.py文件结构如下所示。这部分没有难点，看代码基本就能理解。



# 3 出现的问题

由于岭南院区（LFalldata.xlsx）的病人在清理完成之后仅有27个人，导致dataset1的记录数过少，而病人的属性数达到379个。当属性数远大于记录数时，不适用做统计分析，而在实验的过程也验证了这个说法，即在使用风险比例函数时，无一个属性具有统计意义，得到了错误的结果，因此舍弃这个做法。

# 参考文献

[1] Banerjee, T, Kim, SJ, Astor, B. Vascular access type, inflammatory markers, and mortality in incident hemodialysis patients: the Choices for Healthy Outcomes in Caring for End-Stage Renal Disease (CHOICE) Study[J]. American journal of kidney diseases : the official journal of the National Kidney Foundation, 2014, 64(6): 954-961.

[2] Holme, I, Fellstrom, BC, Jardin, AG. Prognostic model for total mortality in patients with haemodialysis from the Assessments of Survival and Cardiovascular Events (AURORA) study[J]. Journal of internal medicine, 2012, 271(5): 463-471.

[3] Wagner, M, Ansell, D, Kent, DM. Predicting mortality in incident dialysis patients: an analysis of the United Kingdom Renal Registry[J]. American journal of kidney diseases : the official journal of the National Kidney Foundation, 2011, 57(6): 894-902.

[4] Hwang, SJ, Yang, WC, Lin, MY. Impact of the clinical conditions at dialysis initiation on mortality in incident haemodialysis patients: a national cohort study in Taiwan[J]. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association, 2010, 25(8): 2616-2624.

[5] Cox D.R. Regression models and life tables[J]. J.Roy Statist Soc, 1972, (34): 187-220.

[6] Jared, L, Katzman, Uri, Shaham. Deep Survival: A Deep Cox Proportional Hazards Network[J]. arXiv, 2016, (1606): 1-12.