《成人死亡率预测》作业说明

一、数据来源

数据来源于世界卫生组织从各个国家所收集到的真实数据，由于年份跨度大、数据来源广，因此不可避免地存在一些数据项的缺失和错误。

为了处理从真实场景中直接获取而存在噪音的数据，可以通过一些数据清洗和特征工程的方式来对含噪数据预处理，从而获得更好的模型性能。此外，也可以尝试使用不同的回归模型（如决策树和随机森林等），将不同模型结果组合起来，构建更加鲁邦的算法。在第二次授课过程中，我们介绍了一些常用的数据清洗算法，同学们可参考。于是，对含噪数据进行清洗，可提升模型的性能。

二、数据预处理

下面介绍几种数据预处理方法：

关于缺失值处理：

数据中可能会包含一些缺失的数据项，可以如下借助 sklearn 中的 SimpleImputer 来填充缺失值：

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean', missing_values=np.nan)
imputer = imputer.fit(data)
data = imputer.transform(data)
```

上述代码是一个填充缺失值经典的例子。针对数据项中的空值（也就是 np.nan），使用对应列的均值来进行填充。当然也可以改变 strategy 参数的传入值来改变填充方式（'median'是中位数填充，'most_frequent'是众数填充，'constant'用自定义常量进行填充）。

关于异常值检测和处理：

另外，针对含噪数据的预处理这一问题，也有较多的方法来解决，下面介绍比较简单的例子：

先使用箱线图可视化数据的分布（plt.boxplot），来观察是否存在离群点。然后，借助 Tukey's method（即图基法）计算出数据集的四分之一分位数（Q1）和四分之三分位数（Q3），从而计算出四分位数间距（IQR），然后将小于 Q1 - 1.5IQR 或者大于 Q3 + 1.5IQR 的数据点认定为疑似异常值。

```
# 假设 data 是一个一维 numpy array
q75, q25 = np.percentile(data, [75 ,25])
iqr = q75 - q25
min_val = q25 - (iqr*1.5)
max_val = q75 + (iqr*1.5)
len_0 = len(np.where(data>max_val)[0])
len_1 = len(np.where(data<min_val)[0])
limit_0 = len_0 / len(data)
limit_1 = len_1 / len(data)
```

于是，可以定位出异常数据的位置。为了进一步处理异常值，一种简单的方法是借助 winsorize 来做缩尾处理，也就是将超出变量特定百分位范围的数值替换为其特定百分位数值的方法。具体操作可参考如下代码：

```
from scipy.stats.mstats import winsorize
new_data = winsorize(data,(limit_0, limit_1))
```

假设原始 data 为[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]，limit_0 为 0.1，limit_1 为 0.2。这样取值为低于 10%（即 1）和高于 20%（即 9 和 10）的数据都会分别被 2 和 8 替换，于是原始数据被处理为 [2, 2, 3, 4, 5, 6, 7, 8, 8, 8]。

只要结合上面所说的异常检测方法，就可以对异常值做一定程度的修正。

其他

另外，在数据处理过程中也可以删除一些不相关的列（比如在作业中可以通过 data.drop(["Country"], axis=1)）来删除掉国家这个字符串类型的属性。也可以引入一些 data scale 或者标准化的技巧来预处理数据，限定值在 0 到 1 之间，最简单的方法可以借助 sklearn 中的 MinMaxScaler：

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
data_norm = scaler.fit_transform(data)
```

三、评分说明

平台上显示的分数是同学们所提交模型在测试数据运行结果 **r2 score 乘以 100 的数值**，并不是本次作业的分数。

本次作业的具体赋分规则如下（**按十分制记**）：

1. 提交模型且成功运行，且提交文档完整，即可获得 6 分基础分。
2. **r2 score 大于 0.50（平台显示分数大于 50），即可再获得 2 分，即总分 8 分。**
3. 在原模型的基础上，尝试不同的回归模型或者调试模型参数，也可以参考第二节的数据清洗方法，对数据噪音进行处理，来优化数据预处理过程，尽可能地提高 r2 score。**最终 r2 score 大于 0.55（平台显示分数大于 55），即可获得 9 分；最终 r2 score 大于 0.57（平台显示分数大于 57），即可获得 10 分。**