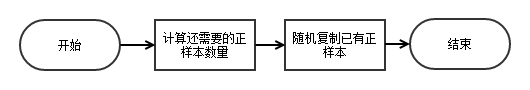
## 数据预处理

#### 平衡分类

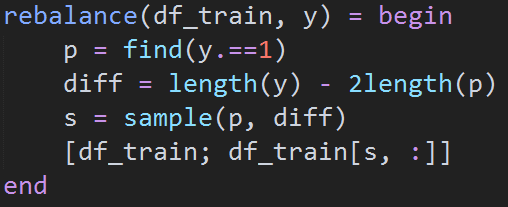
* 方法

1. 计算要使得样本平衡，还缺少的正样本数
2. 在已有正样本中随机抽样复制使数量平衡

* 流程



* 代码

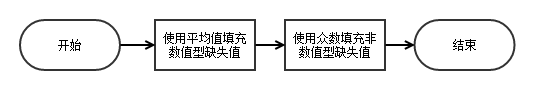


#### 填充缺失值

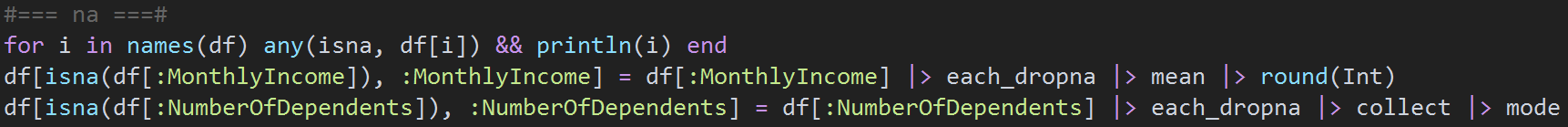
* 方法

1. 对数值型缺失值，使用平均值进行填充
2. 对非数值型缺失值，使用众数进行填充

* 流程



* 代码

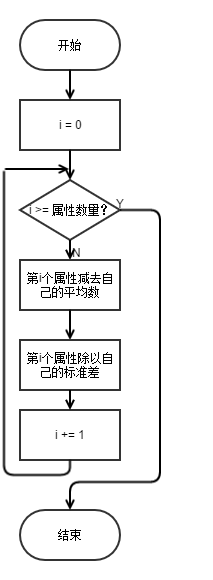


#### 正规化

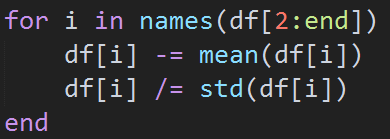
* 方法

1. 每个属性减去自己的平均值
2. 每个属性除以自己的标准差

* 流程



* 代码

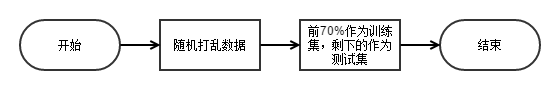


#### 训练-测试集划分

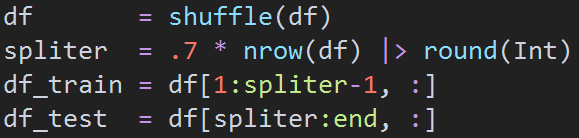
* 方法

1. 随机打乱数据
2. 将前70%作为训练集，剩下的作为测试集

* 流程



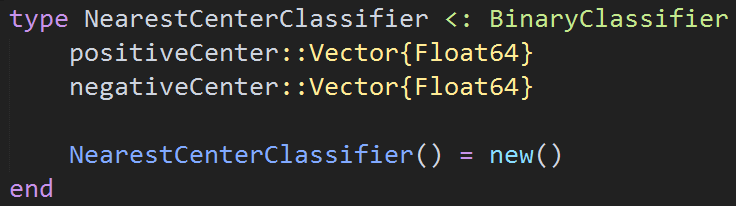
* 代码



## 最近中心分类

#### 模型参数

* 正样本中心：一个向量，表示正样本的中心点
* 负样本中心：一个向量，表示负样本的中心点

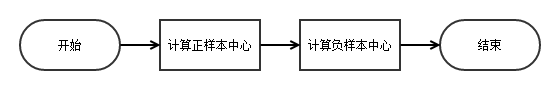


#### 模型训练

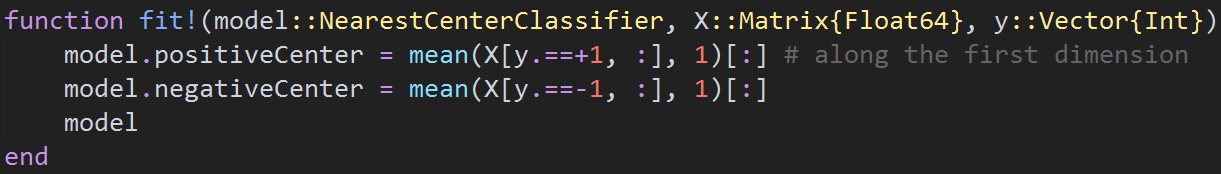
* 方法

对正负类别，分别将所有的样本在每个维度上取算术平均值得到中心点。

* 流程



* 代码

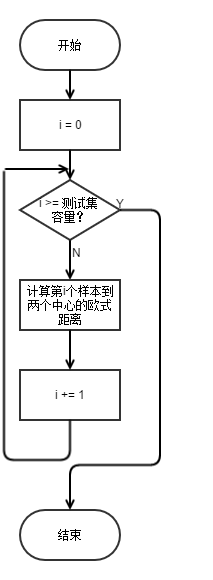


#### 预测

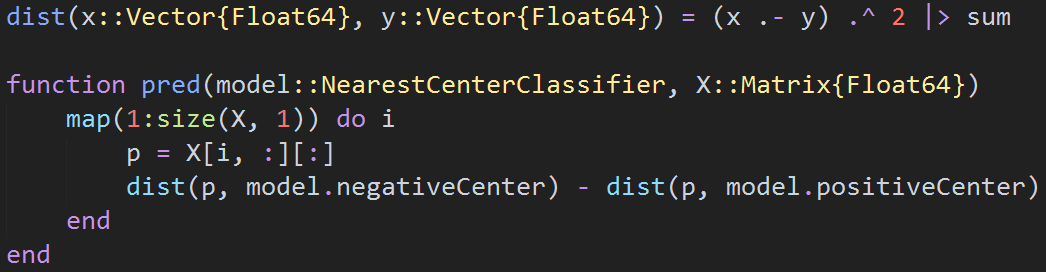
* 方法

对每个样本，计算它到正负中心的欧式距离，使用差值作为得分

* 流程



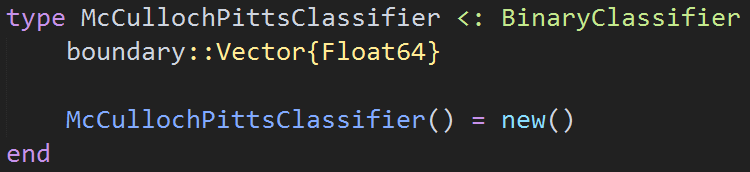
* 代码



## 感知机分类

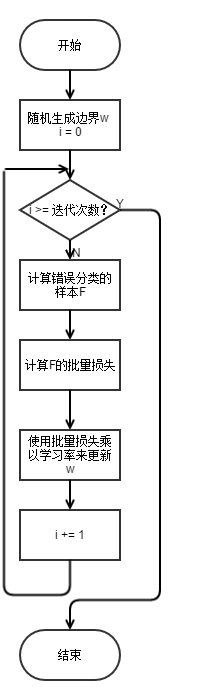
#### 模型参数

* 边界：一个向量，表示划分两类样本的一个超平面

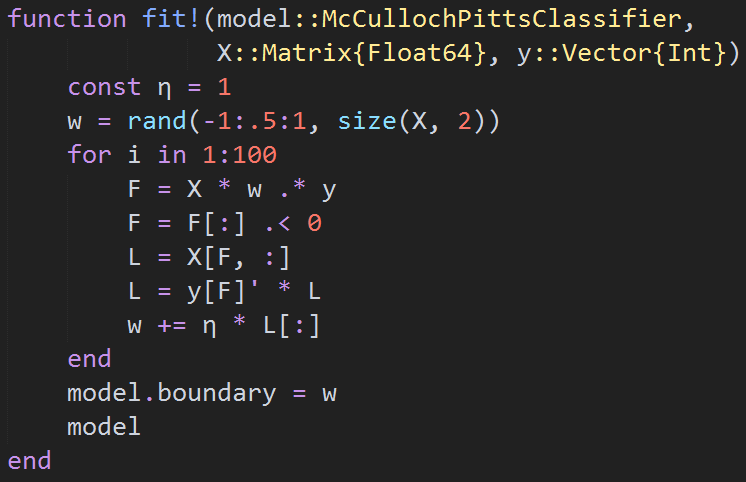


#### 模型训练

* 方法
  1. 随机生成初始边界
  2. 每轮学习计算错误分类样本的惩罚，乘以学习率后加在边界上
* 流程



* 代码

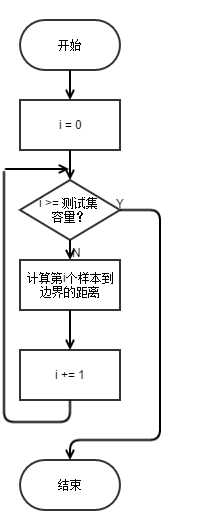


#### 预测

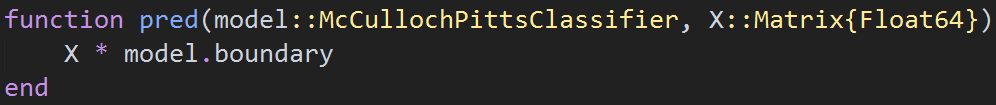
* 方法

对每个样本，计算它到超平面的距离(点乘法向量，有符号)作为得分

* 流程



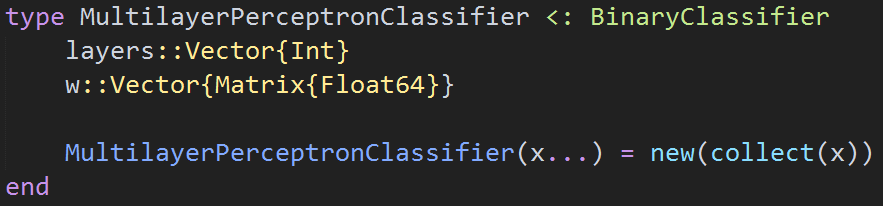
* 代码



## 多层感知机分类

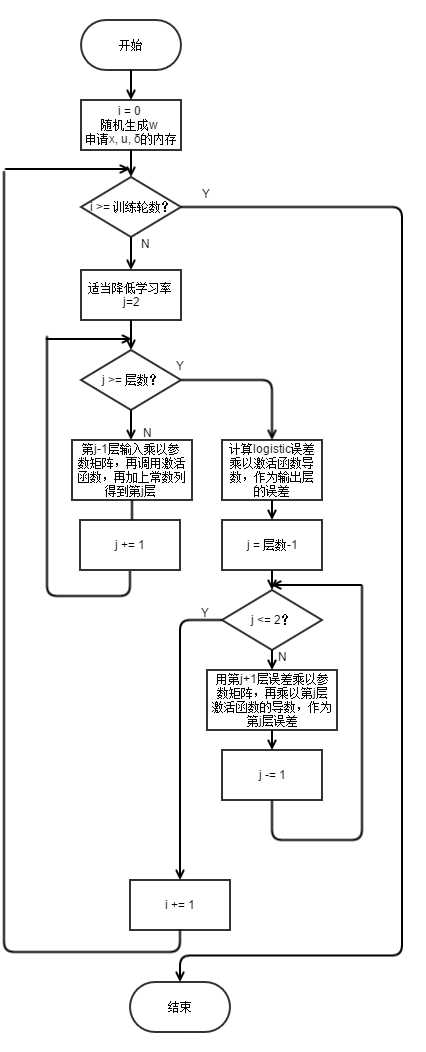
#### 模型参数

* layers：一个向量，表示神经网络的结构，比如[11, 15, 5, 1]表示11个输入，第一个隐藏层有15个节点、第二个隐藏层有5个节点、1个输出的神经网络。
* w：一个矩阵数组，第i个元素表示从第i层到第i+1层的参数矩阵。

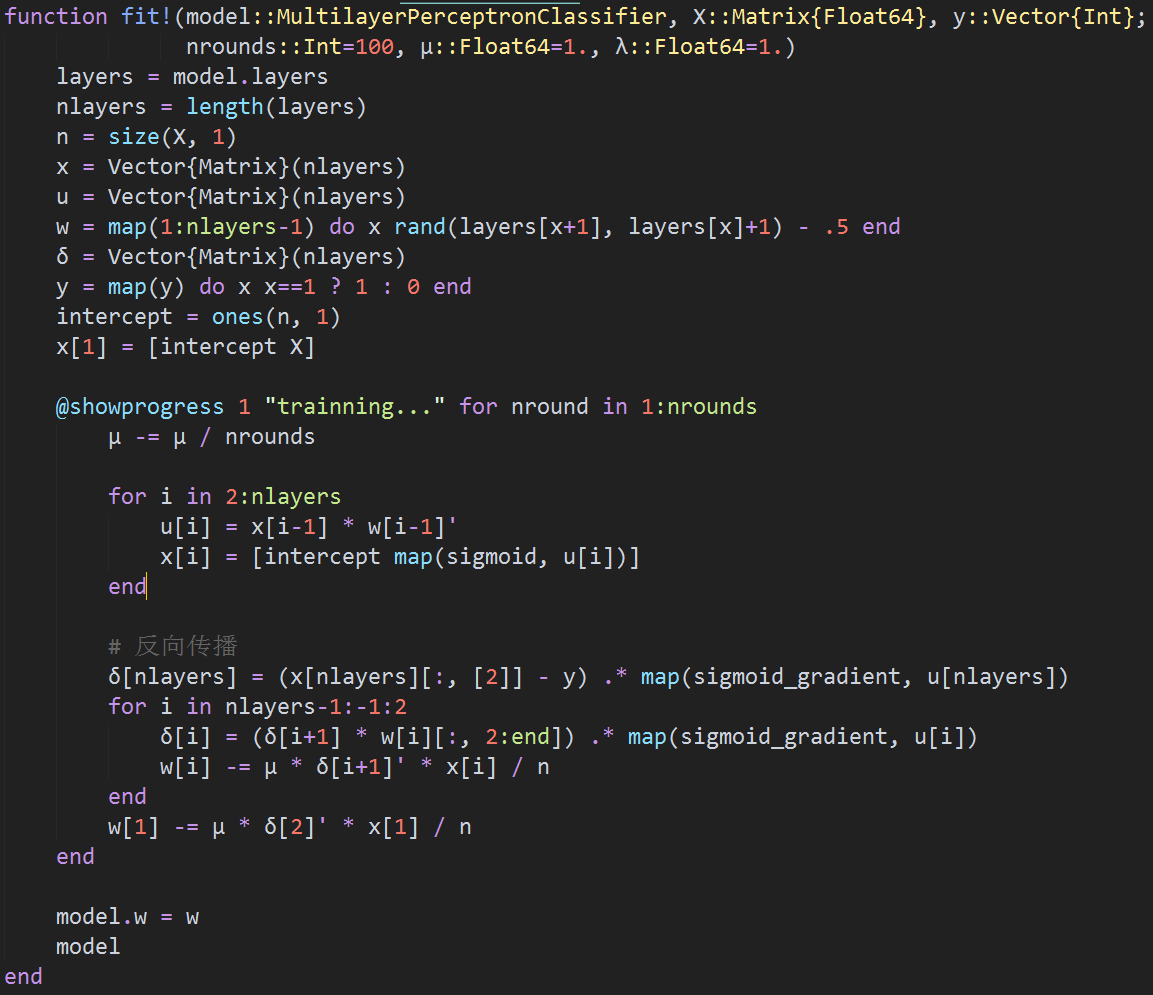


#### 模型训练

* 方法
  1. 随机生成初始w
  2. 每层加上一列常数项1，然后乘以参数矩阵，再调用激活函数，得到下一层，直到计算到最终结果
  3. 结果与真实值计算logistic误差
  4. 误差乘以参数矩阵，再乘以上一层的激活函数导数，得到上一层的误差，直到计算到输入层的误差
  5. 误差乘以学习率，用这个值更新w
  6. 返回第2步，直到达到指定的学习轮数
* 流程

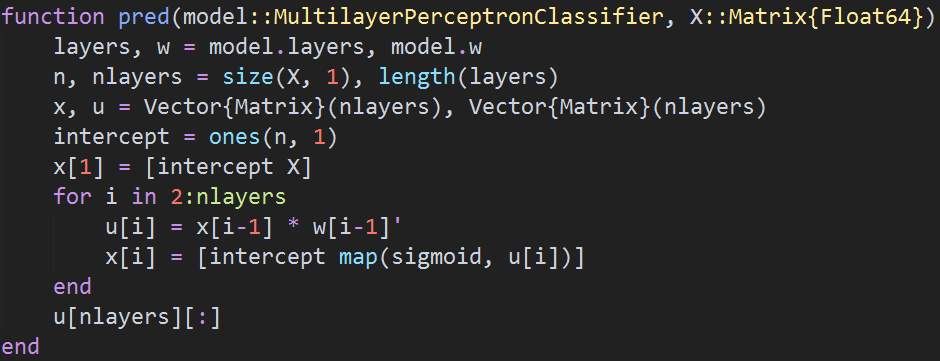


* 代码



#### 预测

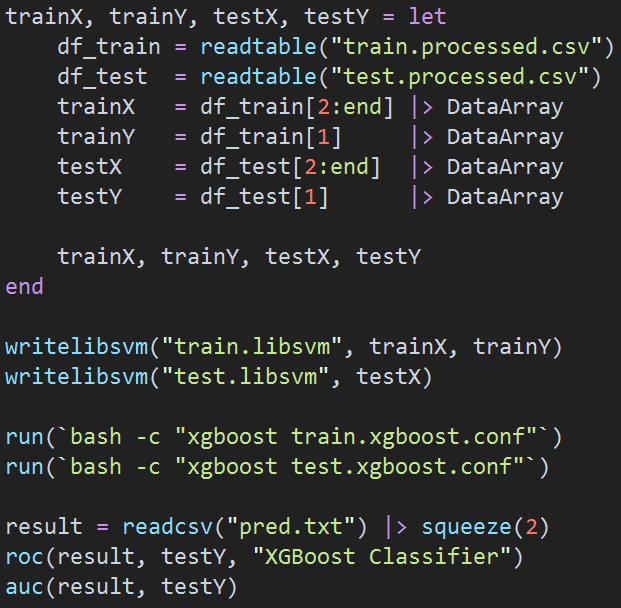
预测其实就是训练的前半部分，输出层结果即为样本得分。



## GBDT分类

GBDT是基于决策树的一种分类方法，通过随机分裂叶子节点来降低基尼熵。与单棵决策树不同，GBDT迭代建立新的决策树，其训练目标为之前的决策树分类的残差。

我们采用了开源的XGBoost软件来实现GBDT分类



## 分类效果评价

#### ROC图绘制

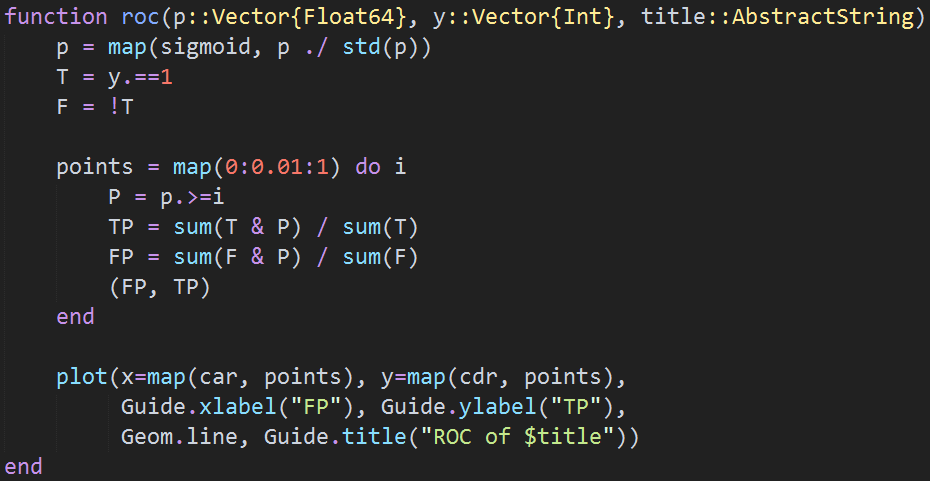
* 方法

1. 使用sigmoid函数将分类器输出的得分转化为概率
2. 对分类阈值的不同取值，计算分类阳性的样本实际为正和负的比率
3. 使用这一组比率描点画曲线

* 流程



* 代码



#### AUC计算

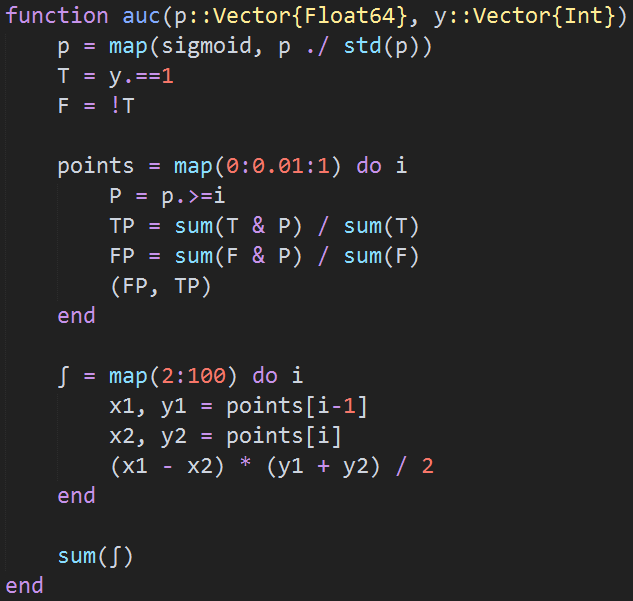
* 方法

1. 同ROC的前两步
2. 划分成小梯形计算面积

* 流程



* 代码



## 模型可视化

#### 决策树可视化

* 方法

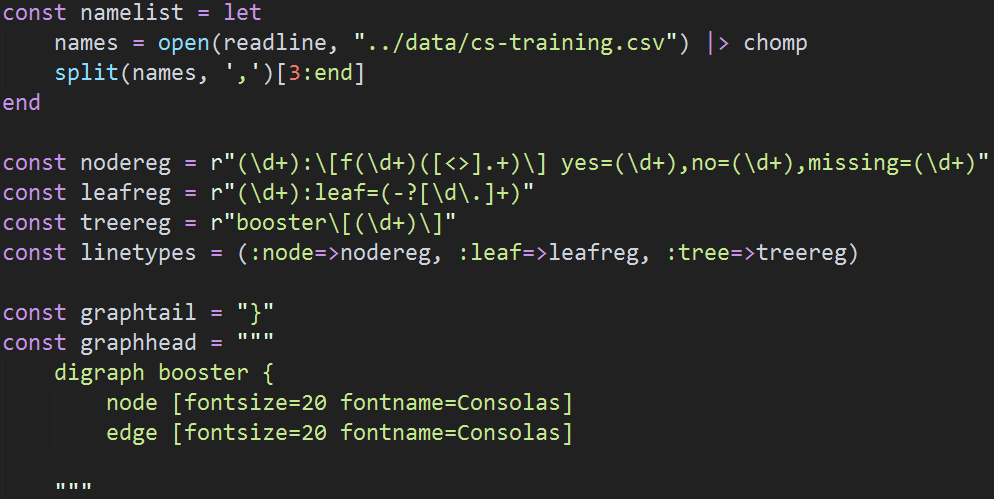
1. 使用正则表达式解析XGBoost生成的决策树
2. 生成Dot代码
3. 使用开源软件Graphviz将Dot代码编译成svg
4. 编写一个网页展示这些图片

* 流程

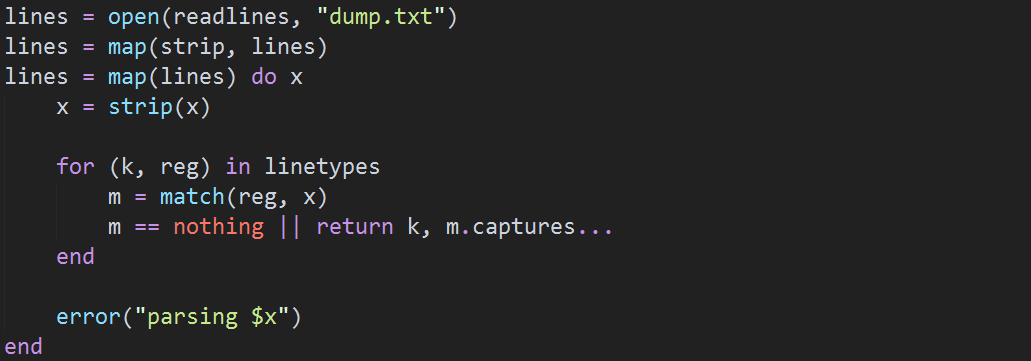


* 代码

-正则表达式和Dot模板



-解析XGBoost导出的决策树



-生成Dot代码



-调用Graphviz和JLT，生成svg以及网页

