sharedParam:共有参数，此包共定义近二十个接口，全都以HasMaxIter形式命名

这些参数，都继承自org.apache.spark.ml.param.Param

位于：org.apache.spark.ml.param.shared.sharedParams.scala

作用：模型之间都有一些共有参数，如最大迭代次数、特征列。

1. 当我们想在自定义类里定义参数时，可以用这些“一般”的共有参数，继承这些，就可以覆写成员变量（相当于set）/还可以get。
2. 利用Param的参数有效性检验，设置不正确的参数还会报错

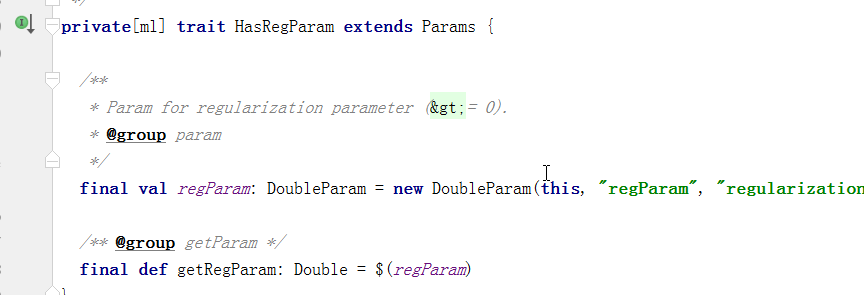
结构：

每个接口定义一个成员变量val， 定义一个相应的getXXX方法，如getMaxIter()

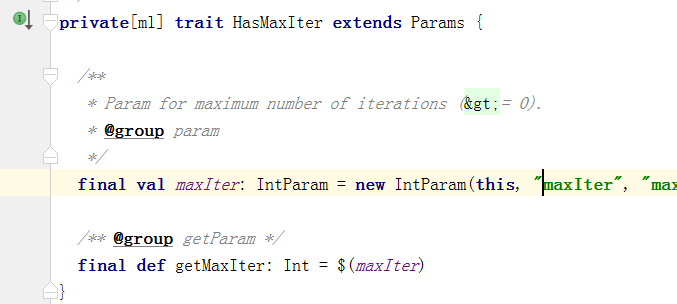
可以简单认为，这些接口，都是一个setter and getter

接口详细如下:

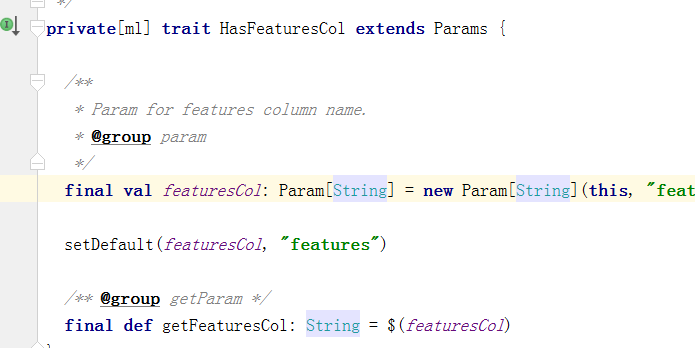
1. 带正则项——参数>0.0，定义损失函数时有用，特别是回归



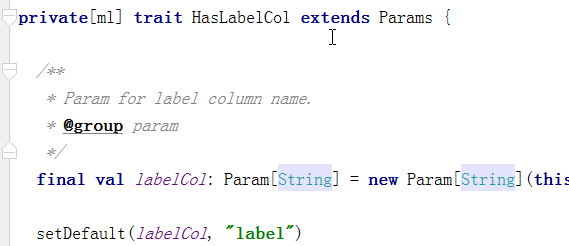
2最大迭代次数————通常用于“优化算子”，如随机梯度下降，Int型



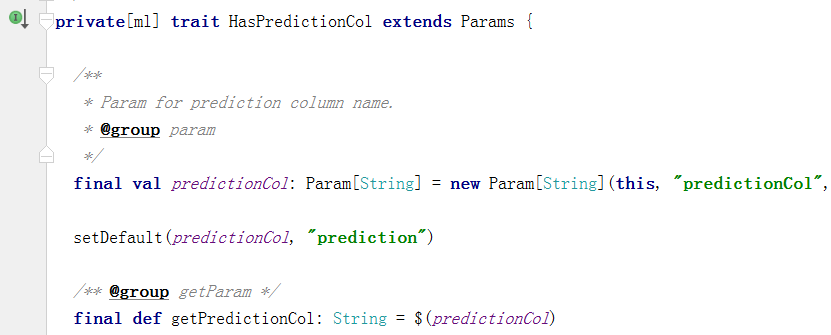
3特征列——String类型



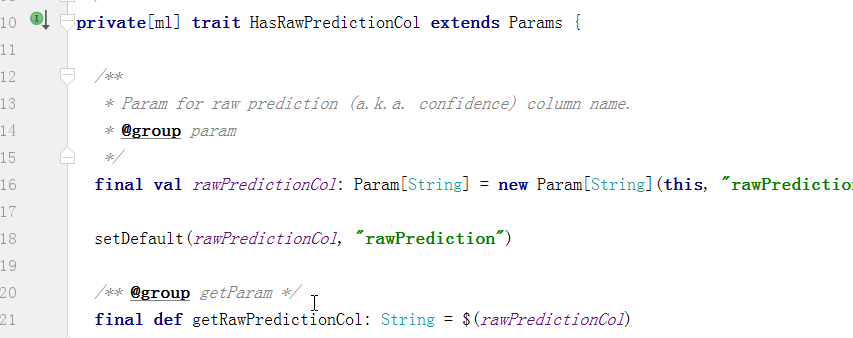
4 标签列——String类型



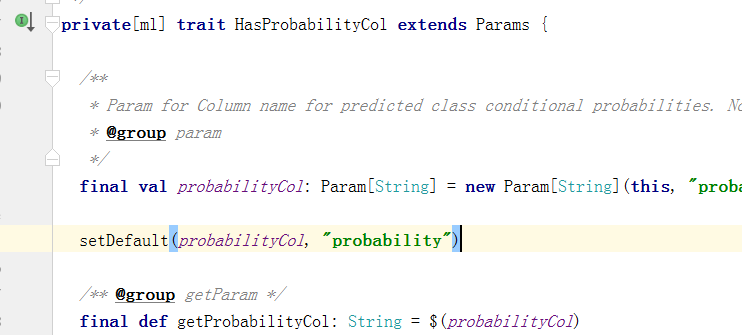
5预测列——当模型训练好，进行预测时，可以指定某一列名，将其作为新的列，该列存放“预测结果”————显然输入数据，应该就是“待预测数据的特征样本”



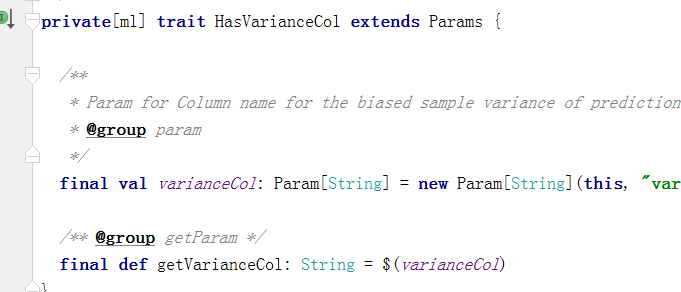
6.原预测列——用途？



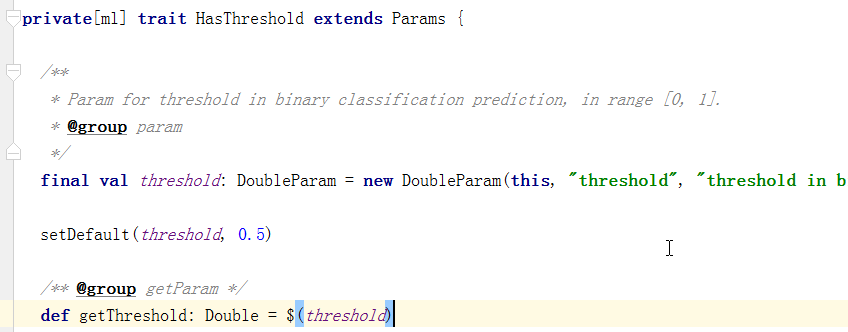
7 概率列——一般是预测结果的置信区间



8有偏样本方差列——也是用于存储预测结果的”方差”，输入String即可



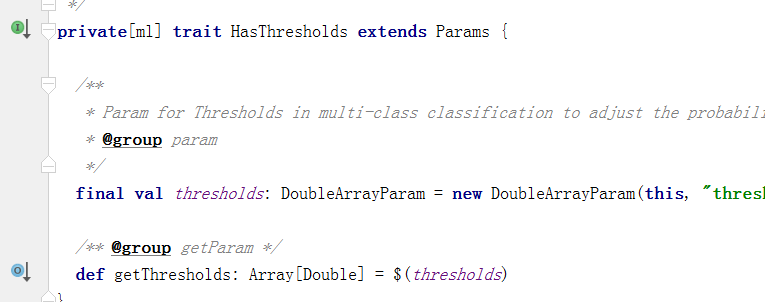
9. 阀值——Double型，[0,1]之间， 一般用于二分类预测，作为终止条件



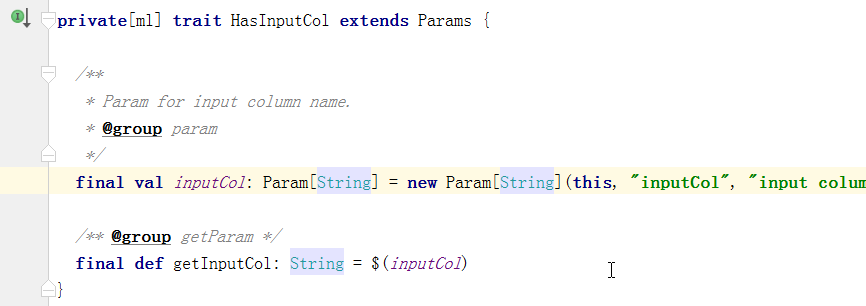
10.多个阀值——Array[Double],均是[0,1]之间

说明：

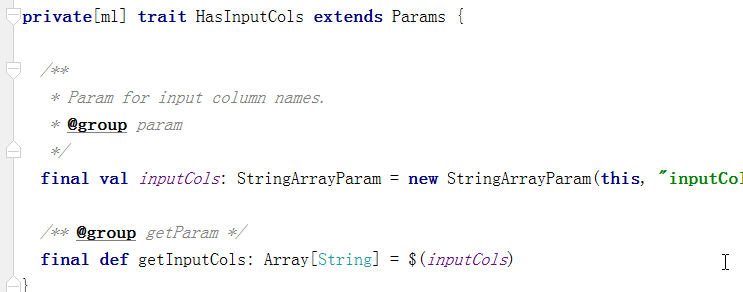
Param for Thresholds in multi-class classification to adjust the probability of predicting each class. Array must have length equal to the number of classes, with values &gt; 0 excepting that at most one value may be 0. The class with largest value p/t is predicted, where p is the original probability of that class and t is the class's threshold.



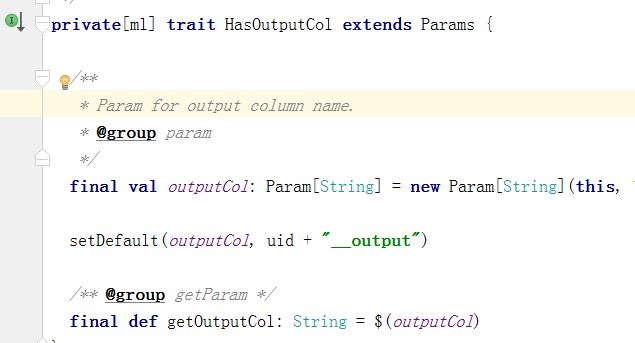
11.输入列——程序内部得到参数后，可以从原data切分得到指定的该列



12输入多列————指定哪几列作为输入

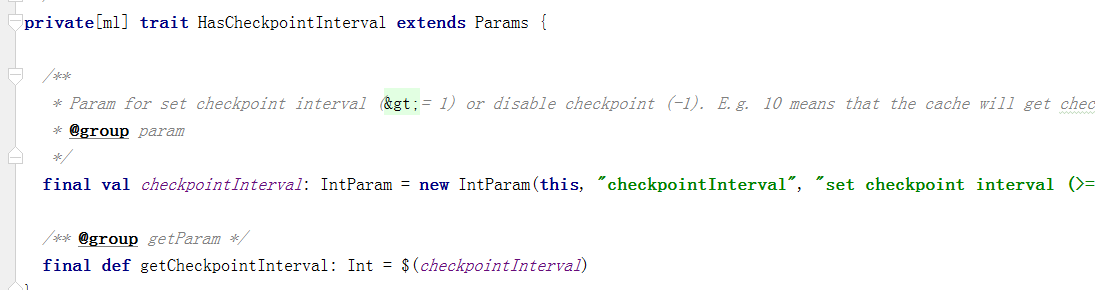


13.输入列——指定哪一列

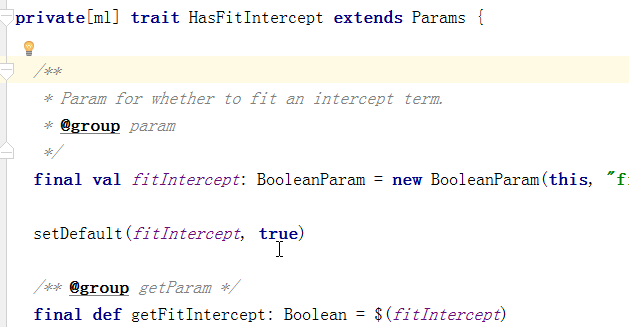


14 断点——用于中间数据缓存，Int型，范围， =-1或者>1————表示每迭代N次，缓存一下数据

说明：Param for set checkpoint interval (&gt;= 1) or disable checkpoint (-1). E.g. 10 means that the cache will get checkpointed every 10 iterations.

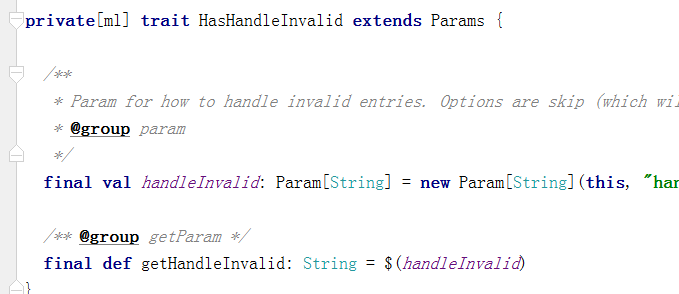


15 截距项——Boolean型，常用于回归

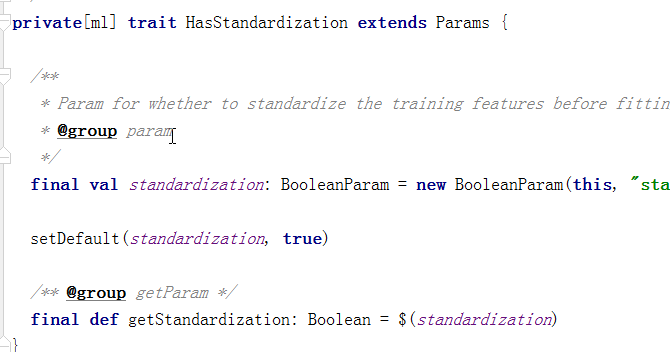


16 如何处理无效样本（df中的行或一行特征向量）——String类型，取值:skip或者error————skip:遇到无效项之间忽略该样本，不用与训练；error:遇到无效项直接报错。

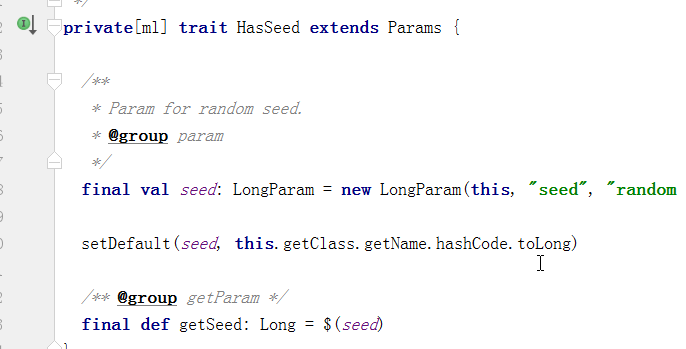
说明：Param for how to handle invalid entries. Options are skip (which will filter out rows with bad values), or error (which will throw an error). More options may be added later.



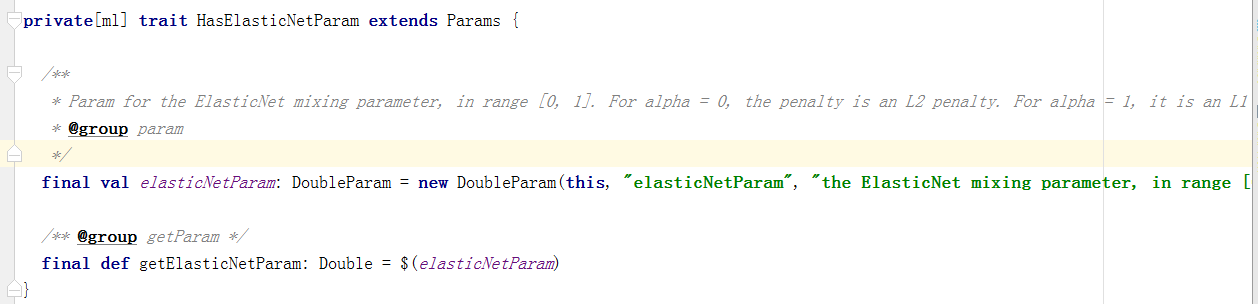
17.是否标准化——Boolean值



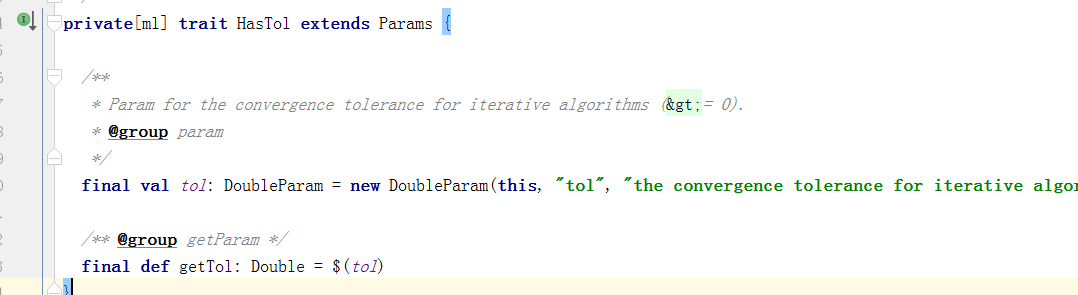
18.随机种子——Long类型，用于生成“伪随机数”，但设定种子后，随机数也就固定了



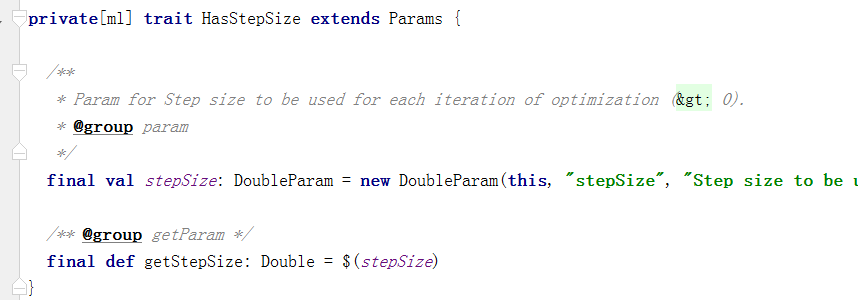
19.L1和L2的弹性系数——Double,介于【0,1】，为0时是纯L2，为1时是纯L1



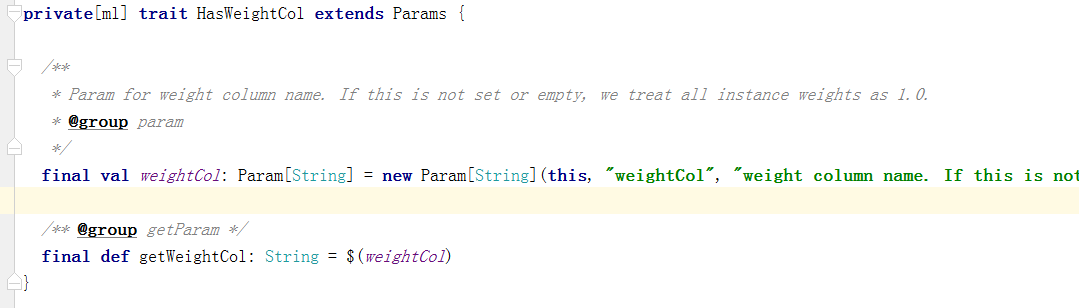
20. 精度——Double,介于[0,1]



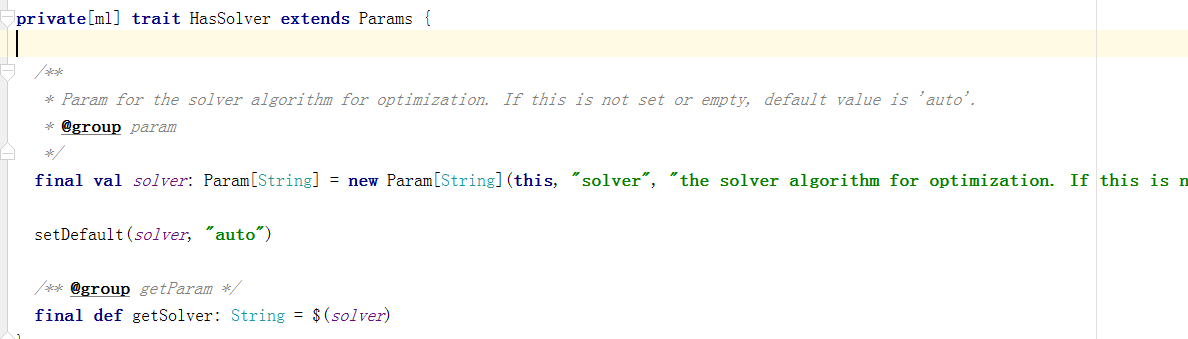
21 步长——Double类型，应该是指学习率



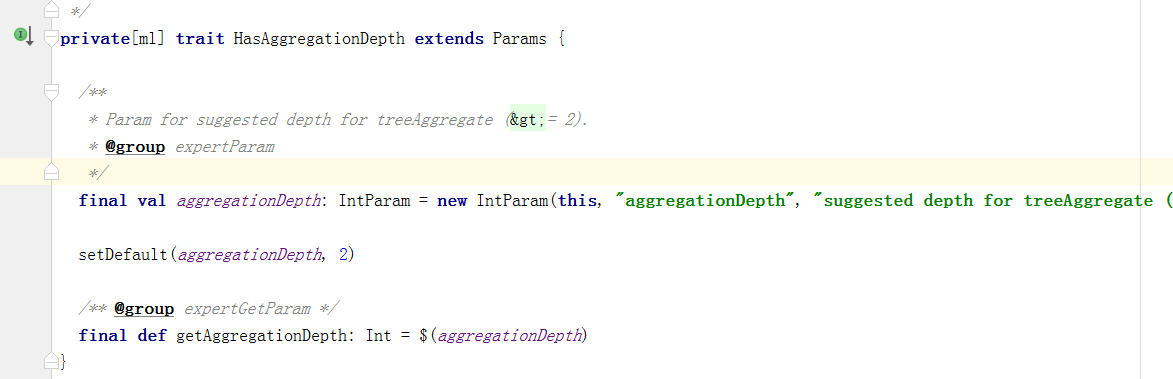
22 权重列——String类型，指定哪一列为权重列，不设置则所有特征列权重均为1.0



23 优化算子类型——String，为空则为“auto”, 应该可以设置如mini-batch, 随机梯度下降

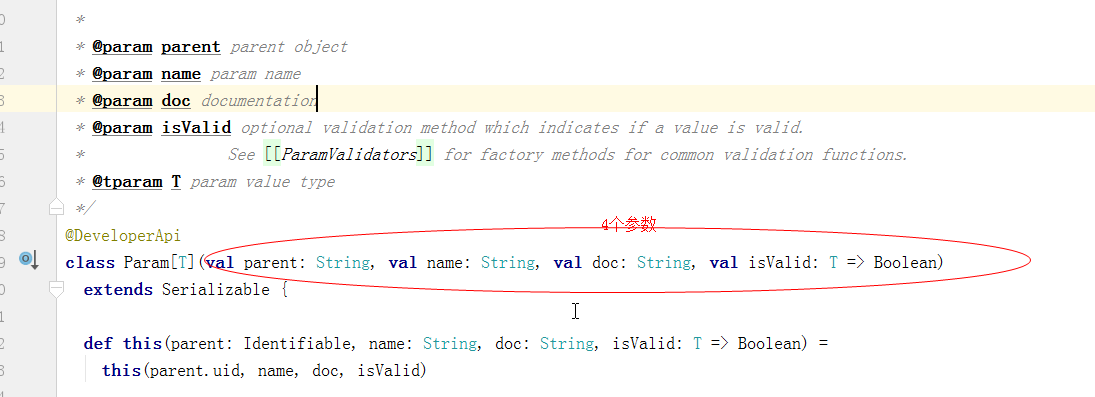


24 树的深度——使用treeAggregate时有用，像线性回归就有这用到这个加总（RDD的加总函数）



如何自定义一个Param呢

就是采用Param的主构造器，传入参数即可，就会得到Param对象



Param.setDefault(param:Param[T], value: Any) 这个方法很有作用——即用户如果自己不输入的话，默认是一个什么值

所以，完全可以自己随便定一个参数，extends Param即可，得到Param的子对象，从而可以overide其成员变量val（因此成员变量可以不用var定义）和方法def

Param还有很多其它方法，可以慢慢研究其作用：最主要是set和get相关及valiation相关

注意，源码如下：

和一般类的定义方式都不一样，都带val，表示这些构造参数，都是这个类里的成员变量，继承其时，可以覆写

