Python机器学习快速上手入门指南

张璇，陈一帅

北京交通大学电子信息工程学院网络智能实验室

**一、介绍**

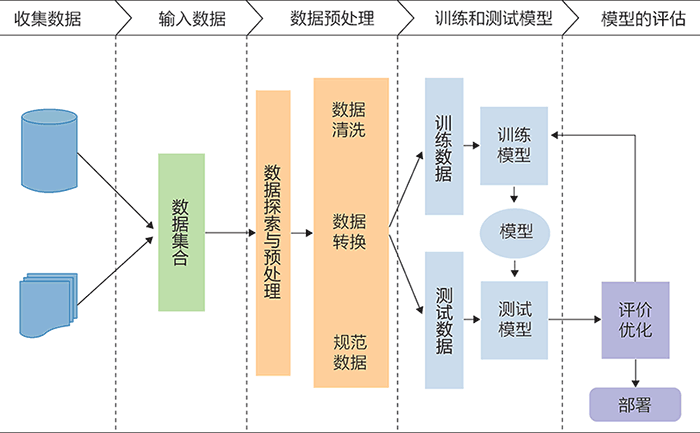
本文假设您有一定的编程经验，比如Java，C++都可以。本文将提供一些信息，帮助您快速上手Python和基于Python的机器学习，编出您的第一个Python机器学习程序。

为什么要用Python呢？使用Python最大的好处是有很多现成库，调用这些库中很多现成的函数，可以用几行简单的代码实现丰富的功能。而且相对于Java和C++，Python非常简单易学，容易上手。

**二、步骤**

使用Python完成一个机器学习项目，要完成哪些**步骤呢？**

请大家首先对机器学习/深度学习的完整流程有一个概念。请大家看《Python机器学习基础》分享的一张图（<https://www.92python.com/view/373.html>）。这张图我们也分享在下面。每一个机器学习项目都可以用下图中的流程概括，流程中的每一个步骤也就是我们上课学习的内容。



**三、每一个步中，使用的python库**

那么，在上图所示的每一步中，我们可以用什么Python库编程呢？

请大家看这个网页：《【机器学习实战】科学处理鸢尾花数据集》<https://blog.csdn.net/weixin_40431584/article/details/105433366>

该网页是一个简单的例子，用Python对一个鸢尾花数据集完成上述机器学习的完整流程。

在这个例子中，数据处理用了pandas库，数据可视化用了matplotlib、seaborn库，训练决策树模型用了sklearn库，评估也用了sklearn库。

你会发现，其实每一步骤只需要几行代码就能够实现的。

  下面是常用的python库：

1. numpy、pandas——数据处理
2. sklearn、tensorflow、pytorch——机器学习/深度学习
3. pyspark ——大数据

**四、在库中找到需要的函数**

要实现每一步的具体功能，如何在库中找到需要的函数呢？

关于函数最完整的介绍，功能和参数，尽量看官网的官方文档，因为机器学习领域更新特别快，这些库的更新也特别快。官网上的文档是最新的。很多库的文档都有我国的志愿者翻译的中文版。请一定找到它们。

然后可以百度/CSDN/github/… 看看大家是怎么用的。

举个例子：心电图类型识别案例，要使用决策树模型进行分类。那么，调用sklearn库的什么决策树模型函数呢？请按下面的步骤

1、打开sklearn官方文档：

中文（<http://www.scikitlearn.com.cn/>）

英文（<https://scikit-learn.org/stable/modules/classes.html>）

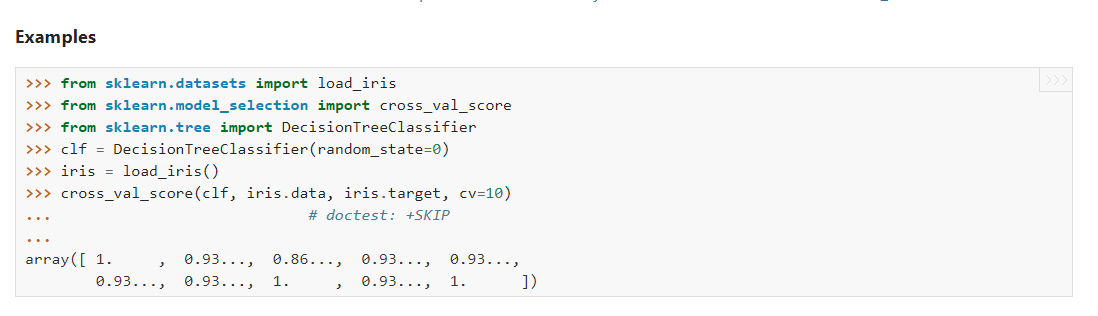
2、找到决策树DecisionTreeClassifier函数的说明

中文（<http://www.scikitlearn.com.cn/0.21.3/11/>）

英文（<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier>）

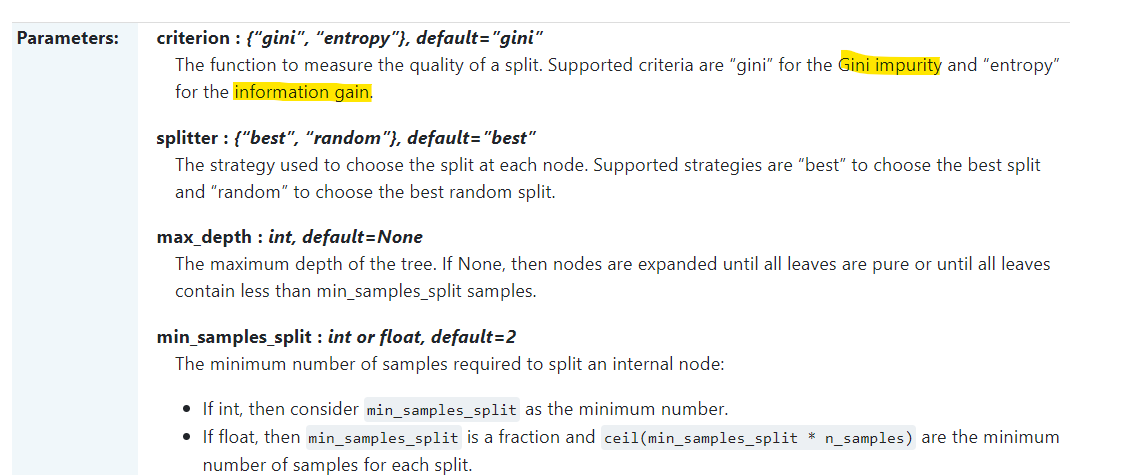
官方文档中有最详细的介绍：

特别是它会举例子告诉你这个函数怎么用，如下图中的代码：



上图中的代码中给出了使用这个函数的一般方法。

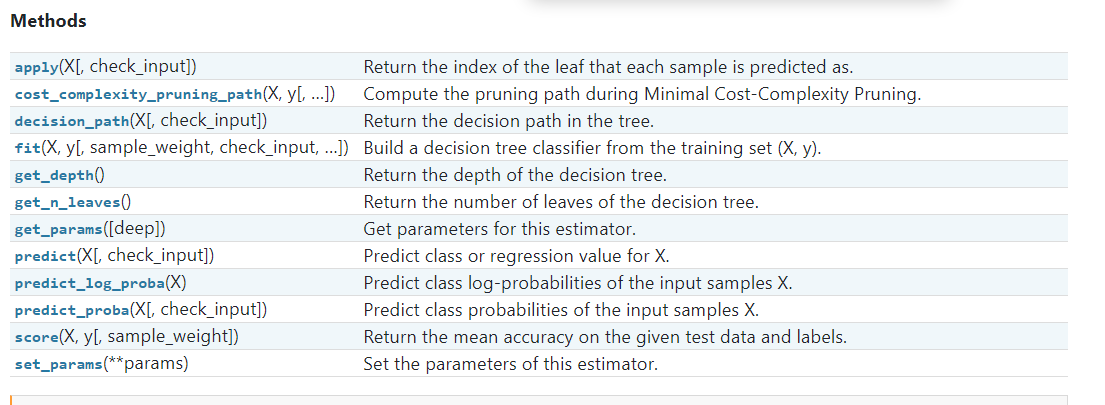
这个函数的参数怎么设置呢？文档里也有详细的说明，如下图所示：



上图中给出了该函数各个参数的详细说明。其中，有些参数上课讲算法时提到过，比如gini impurity（基尼不纯度）、entropy（熵）、information gain（信息增益）。

设置参数最简单的方法就是：什么参数都不改，就用默认值就好。大多数参数都有默认值（default项）。请放心，默认值一般来说是经过精心挑选的，在大多数情况下都有很好的性能。

都有些什么函数可以调用呢？文档里有详细的说明，如下图所示：



上图就给出了决策树模型最全的函数方法和它们的功能。

也请大家善用搜索引擎。百度输入关键词（python、sklearn、决策树、…），看看别人对于各种不同的数据集是怎么用的

**五、机器学习模型的训练**

训练机器学习模型的代码其实很简单。比如上面例子中对鸢尾花数据集使用的决策树就很简单，就几行代码

1. train\_test\_split函数划分测试训练集
2. fit函数训练模型
3. score函数输出结果准确度。

如下图所示：



**六、Python机器学习环境的安装和使用**

使用python，推荐安装anaconda（一个集成库，里面放了很多很多的常用库），也可以就在华为云平台上跑代码就行~