棕榈学院

7 天 Python 进阶训练营讲义

第七讲

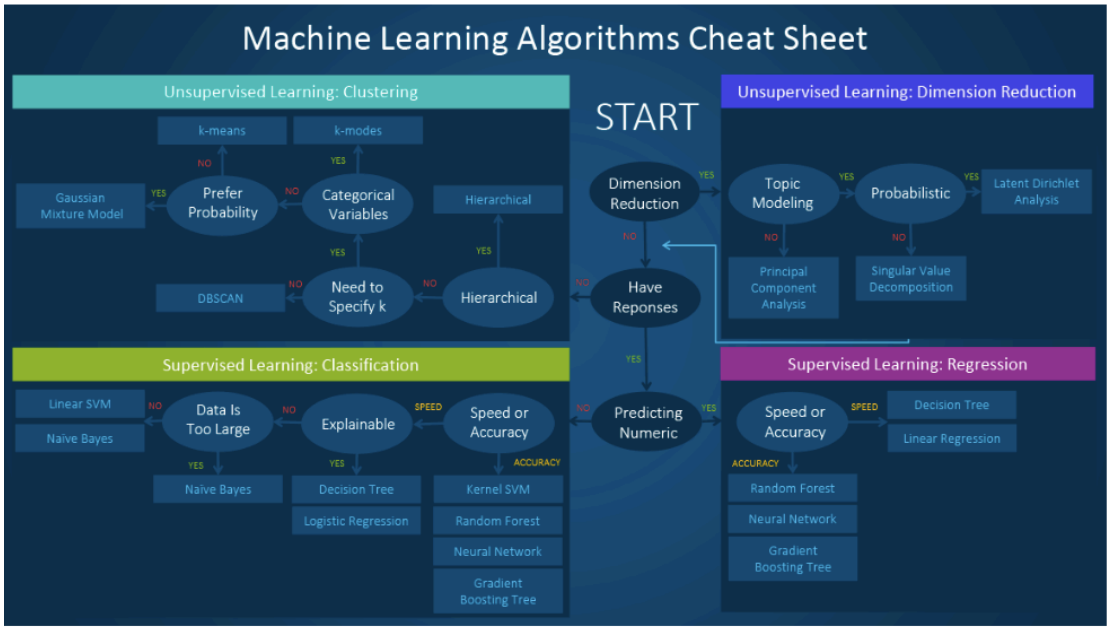
2019 年 1 月 14 日-2019 年 1 月 20 日

目录

在上节课我们完成了泰坦尼克号的所有 code 相关部分，从中已经得到了一些结果。在最后一节课和同学们分享一下如果想要继续学习 Python 的学习方向以及未来的发展规划。

一、关于 Python 的学习

1. 非监督学习



再来回顾一下这张图：

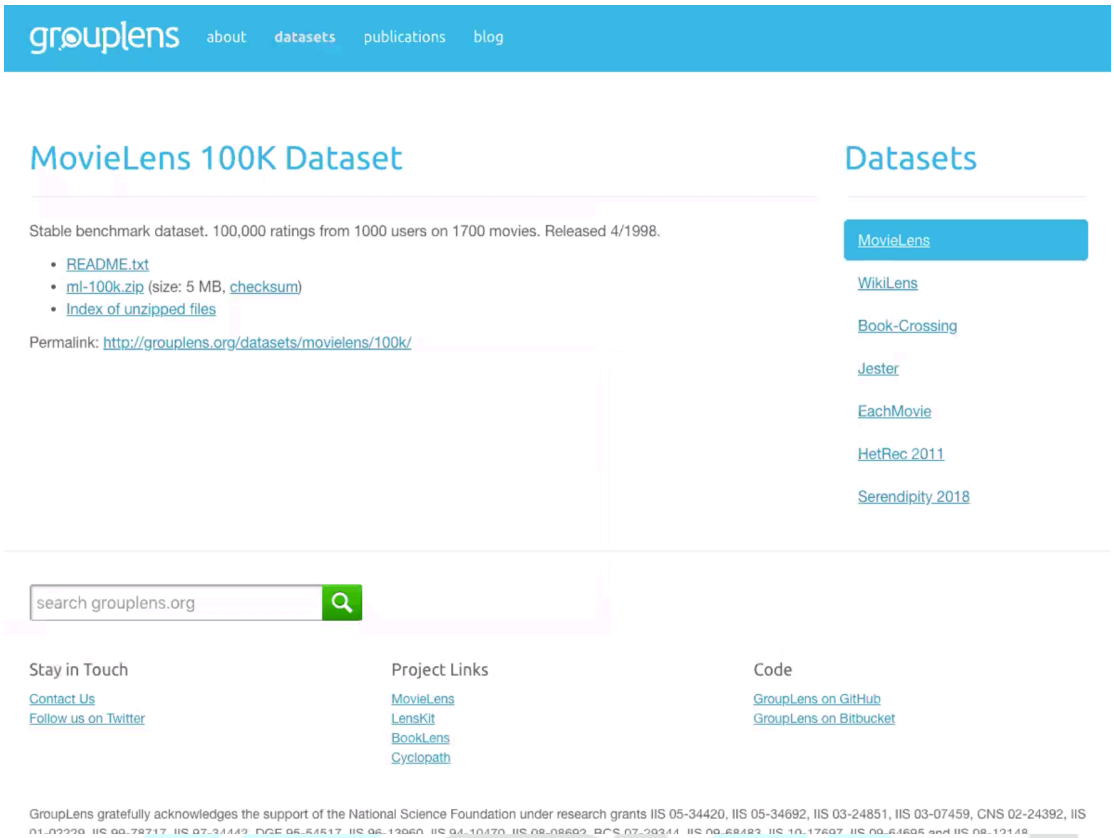
这一张图是机器学习算法的图，我们在第三讲曾对其做过讲解。分为两大部分：非监督学习与监督学习。我们做的泰坦尼克号就是监督学习中的一种。

而另一大类就是非监督学习，非监督学习并不是要预测结果。比如 Topic Modeling，即为没有确切结果，更多是根据数据特征对数据进行分类。

泰坦尼克号这个 project 所做的是用数字类的数据去处理问题，如果同学们感兴趣，也可以尝试去做一个非监督学习去处理文字（中英文均可，但中文相对处理起来更加困难。）。

此外，推荐系统相关的内容，也是非监督学习可以去探索尝试的，同时，推荐系统也是机器学习非常常见的内容。

比如 GroupLens：有电影相关的数据，不同于全是文字，这一网站的数据都是用来做推荐系统的。



Movielens 是非常简单的一些数据，如果感兴趣可以去探索数据，会用到一些比如说协同过滤等知识去做推荐系统。

2. 深度学习

如果处理数字相关模型做得较好之后，可以进入到下一步：运用深度学习，进行比如图像识别等操作，再进阶可以进行语音识别（将语音识别为文字、不同语言的处理等）的操作。

学院君小提示⏰：

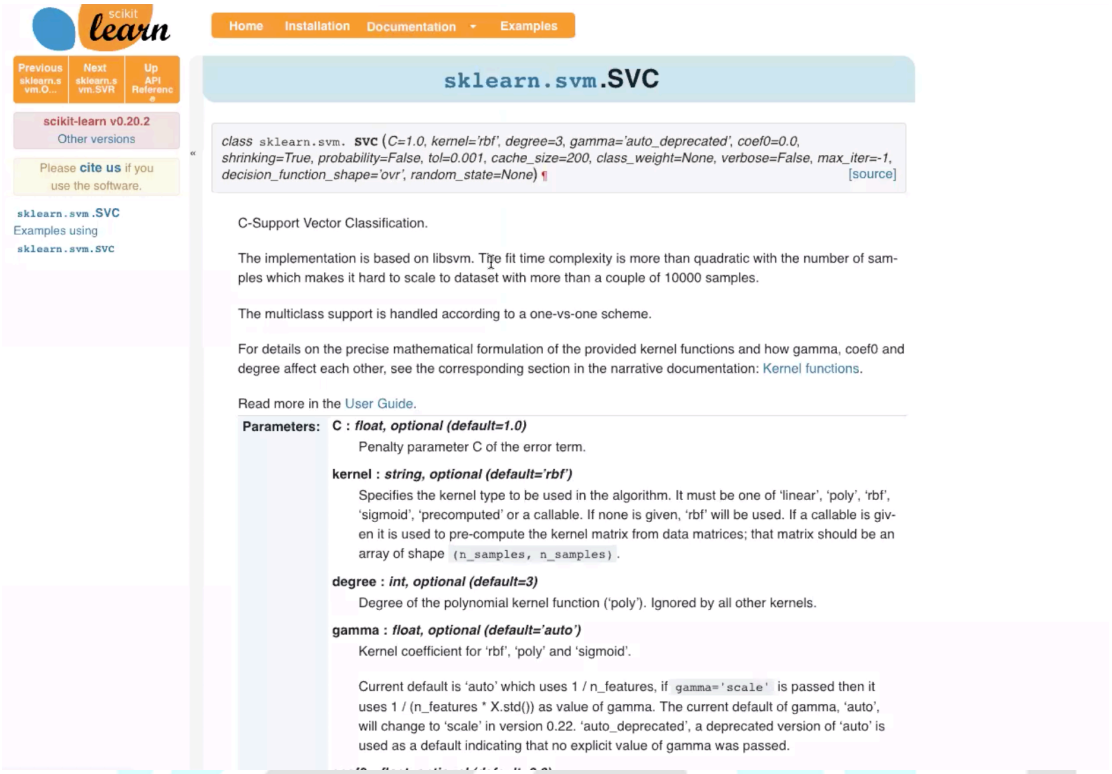
机器学习可以学习的内容还有很多，大家可以自行探索更多哦～

3. 关于 Python

Python 只是众多编程语言中的一种，相对其他语言理解起来更加容易，但能做的不是简单的机器学习的 project，它还可以进行爬虫。用 Python 可以自行用文本把文本从网站上爬虫抓取，自己再去处理文本。爬虫这一步就是我们这个 project 所没做的一步操作：获取数据。因为所需数据 Kaggle 已经给出，所以不需要再去进行获取数据这一步内容。

比如上节课展示给大家的 Kaggle 上的不同数据，在自己做模型的时候建议采用不同模型来尝试，比如 Kenel SVM，Neural Network，Naive Bayes 等（详见上图），看看结果会有什么不同。

而基本这些内容都可以在 sklearn 中找到：



sklearn 对于内部参数会有所介绍。但参数不用考虑太多，可以先将主流模型均过一遍，对其产生基础认识，对其优缺点产生更好的理解，这样以后在使用的时候才会知道使用哪个模型更为合适。

4. 做项目的关键点

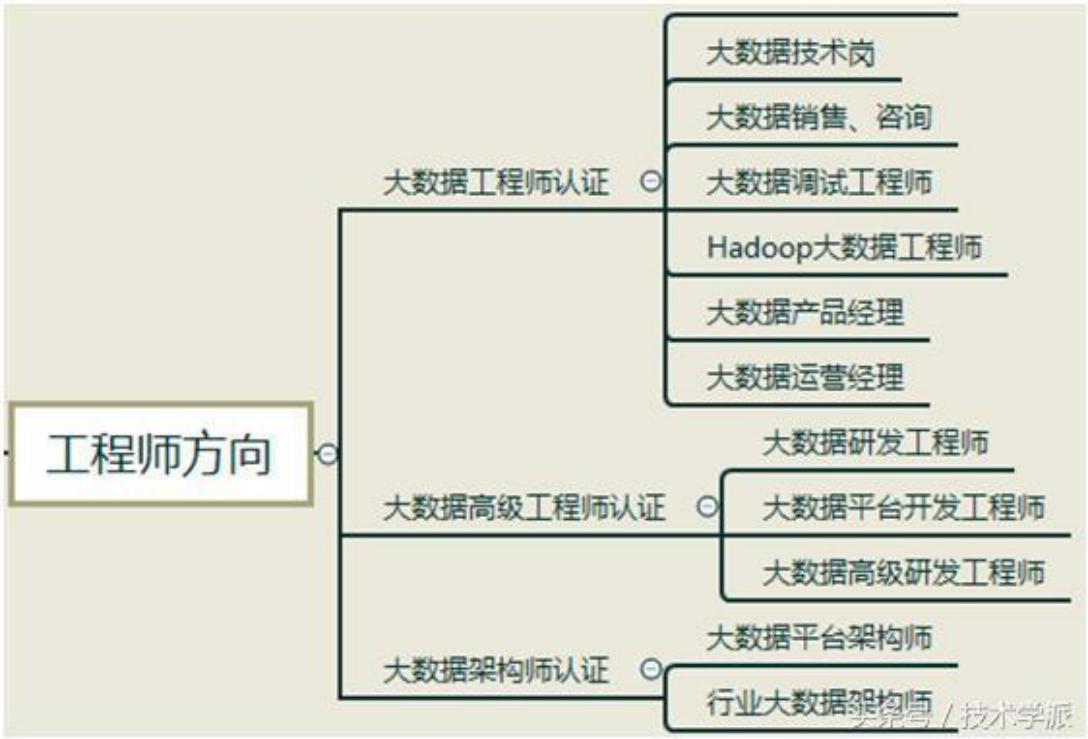
在做不同项目找到自己喜欢的数据去做预测的时候，不需要追求非常高的精确度，更多的是去认识、理解这些数据，要更多的去思考这些数据为什么会对预测有用，这些数据是怎样来影响预测的。总的来说，要更多的去理解这个问题，找到数据和结果之间的关系，而不是去追求最终百分之零点几或一点几精确度的提高。

二、Python 工程师可从事的多领域编程工作

1. 分类

（1）爬虫开发
（2）人工智能
（3）数据分析
（4）python 开发
（5）算法工程师
（6）系统运维工程师
（7）搜索引擎工程师
（8）测试自动化

2. Data Engineer 的工作日常



Yiya 导师在一家公司做 app，类似喜马拉雅 app，可以在 app 上听广播，担任职位即为大数据工程师(Data Engineer),大数据工程师所做的工作大概分为三大部分：

第一部分：保证数据的完善。

因为用户用 app 收听广播，进行各种操作之时，数据会以不同形式导入到最终数据库中，反馈到系统中，大数据工程师所做的是获取数据、对其进行清理，最终导入

到大的数据库中。相当于数据管道的一部分，要保证数据的管道不要出错，因为每时每刻都会有数据进入，一出错就不能接收到部分来自用户的信息。

第二部分：

（1）写 sql 导出数据。

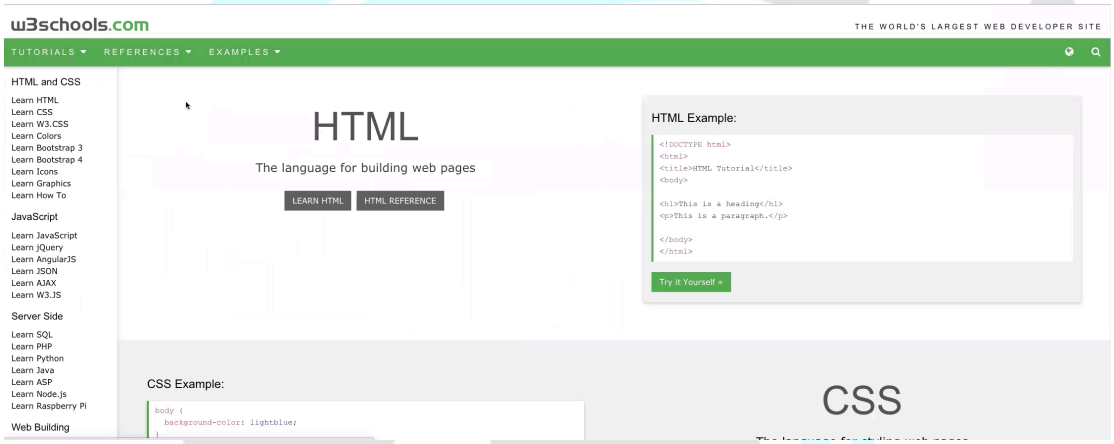
给其他部门如市场部门、广告部门等，他们要了解有关数据的情况时，就由数据工程师写 sql，从数据库中导出数据，提供给这些部门。因为如果数据量很大的话，不能全部导入画图，否则非常慢，所以用 sql 提取数据至关重要。

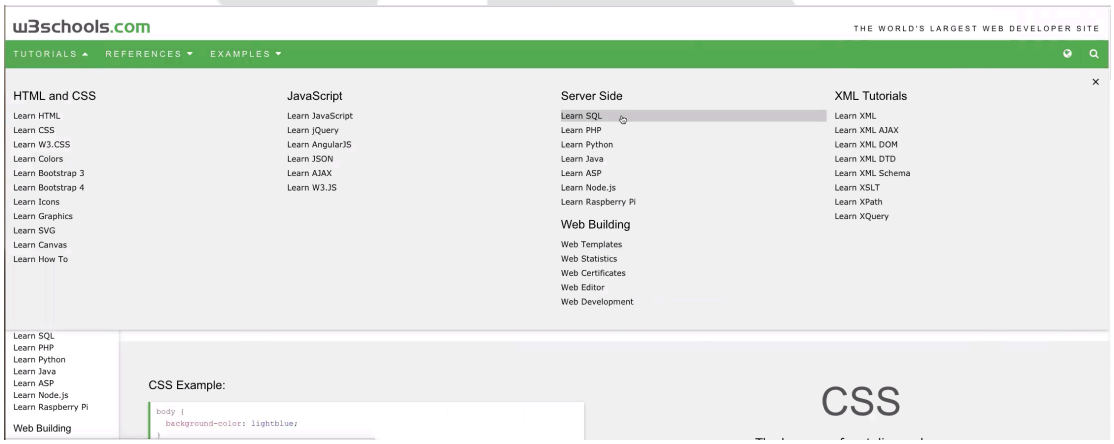
学院君小提示⏰：

sql 是非常重要的知识点，如果有机会可以多去学习一下～

会有很多不同的 sql，比如 mysql，redshift 等，但这些都是大同小异，语言其实是非常相似的，只要去精通一个，其他上手很快。

学习 sql 推荐网站：w3schools





（2）对数据制作图表。

图表是对数据的分析，很多图表很像我们所做的 project 中所做的分析。画图这一部分，虽然看起来会比较简单，但其实是非常重要的一部分，因为去做数据分析的时候，分析数据后更多是要展示数据，如果表格可以很清晰表达一些东西，效果会比较好。

【总结】Presentation 是工作中非常重要的一部分，而制作图表是 Presentation 中非常重要的一部分。

第三部分：与推荐系统相关，向用户推荐与其喜爱节目相关的其他节目。

对于一个时长较长的广播节目来说，这是一个比较复杂的问题。主要有三种推荐方法可以使用：

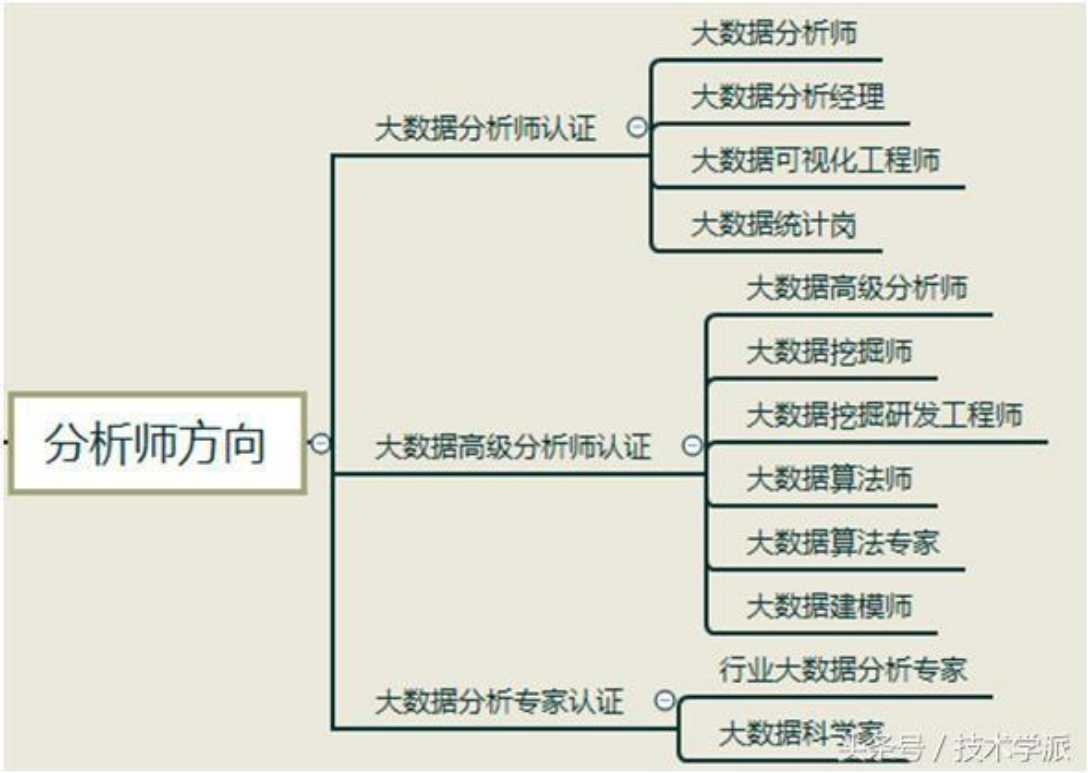
一种推荐方法是基于节目进行推荐。要推荐节目首先要了解用户为什么收听这一节目，是喜欢主题、喜欢主持人或其他情况，接下来方能为其推荐相关类似节目。这是一种推荐方法。

另一种推荐方法为基于用户进行推荐。比如用户 A 收听若干个节目，发现用户 B 和用户 A 有很多相似点，那么如果 A 收听了新节目，可以将其推荐给 B。这就是两种比较主流的推荐方法。

第三种推荐方法为将前两种方法结合起来的更加高级的方法。

如果同学们想做相关的推荐的内容，完全可以到前一部分所述的 Grouplens 网站中的 Movielens 数据，它的数据分类非常简单：用户、电影、评分，几乎不用去做数据清理，即可以直接尝试三种不同模型：基于用户推荐、基于电影推荐、用户推荐与电影推荐相结合。

3. 热门岗位：Data Analyst



现在学习 Python、sql 的知识，在国内外有关的非常热门的岗位叫做数据分析师(Data Analyst)，在很多科技公司、金融公司都招聘数据分析师这一岗位。数据分析师在大多数情况下并不用太多去做模型这一部分,更多的是要去对数据进行分析，即画图、理解和思考等，再展现给市场运营部门等。数据展示、数据分析和模型是非常不一样的部分。

学院君小提示⏰：

在这里也建议大家在做 project 的时候多做一些图，可以用 Matplot，也可以用 seaborn，做了图之后多去思考图中的意义。不仅在 Python 中可以做图，如果想做更加漂亮，大型的图的话，可以去了解 tableau 这个软件哦～

三、作业

到今天课程结束，我们的 7 天 Python 进阶训练营就临近尾声啦～请大家谈谈自己学习本次课程的感想，报名学习课程的原因是什么呢？还有什么有关 Python 的知识点是想要学习的呢？以及其他想要表达给学院的内容都可以写一写哦❤

四、第六讲作业答案

答案示例：

病者癌症为良性，检测结果为恶性，则为 False Positive，假阳性

病者癌症为恶性，检测结果为良性，则为 False Negative，假阴性

这种情况就是 False Negative 特别不好 我们需要绝对去降低，因为这是属于“漏诊”的情况，风险极大。当然，不论是误诊还是漏诊行为，都是对患者生命极大的不负责任。