**金融学院 金融学 翁嘉译 1120240628**

**python里面相关的数据处理模块及功能**

**1、NumPy**

NumPy是Python开源的数值计算扩展工具，它提供了Python对多维数组的支持，能够支持高级的维度数组与矩阵运算。此外，针对数组运算也提供了大量的数学函数库。NumPy是大部分Python科学计算的基础，它具有以下功能：

1. 快速高效的多维数据对象ndarray。
2. 高性能科学计算和数据分析的基础包。
3. 多维数组（矩阵）具有矢量运算能力，快速、节省空间。
4. 矩阵运算。无需循环，可完成类似Matlab中的矢量运算：线性代数、随机数生成以及傅里叶变换功能。

**2、Pandas**

Pandas库可用于快捷的数据处理、聚合和可视化，尤其可以灵活处理缺失数据及数据分组。它可根据需要帮助组织各种参数的数据，是基于Numpy底层数据结构，让python成为类似Excel，R等统计学软件，主要就是Pandas的功劳。Pandas在Python中实现了各种数据的计算，分组计算，添加删除，排序，筛选，抽样等都能工作。使Pandas成为数据科学家中最受欢迎的库。

Pandas主要包含两种数据结构：Series与DataFrame。Series是一种类似于以为数组的对象，它由一组数据以及与之相关的数据标签组成，仅有一组数据即可产生最简单的Series。Series可以构成二维的DataFrame。行为记录值，列为观测值。Python能直接处理数据，而Pandas几乎可以像SQL（Structured Query Language 结构化查询语言）那样对数据进行控制。

**3、Scipy**

科学及工程软件库，包含线性代数、优化、集成、统计等功能。它的主要功能基于Numpy库，因此其中数组操作大量应用了Numpy库，Scipy是一个用于数学、科学、工程领域的常用软件包，可以处理插值、积分、优化、图像处理、常微分方程数值解的求解、信号处理等问题。它用于有效计算Numpy矩阵，使Numpy和Scipy协同工作，高效解决问题。

SciPy是一组专门解决科学计算中各种标准问题域的包的集合，包含的功能有最优化、线性代数、积分、插值、拟合、特殊函数、快速傅里叶变换、信号处理和图像处理、常微分方程求解和其他科学与工程中常用的计算等，这些对数据分析和挖掘十分有用。Scipy是一款方便、易于使用、专门为科学和工程设计的Python包，它包括统计、优化、整合、线性代数模块、傅里叶变换、信号和图像处理、常微分方程求解器等。Scipy依赖于Numpy，并提供许多对用户友好的和有效的数值例程，如数值积分和优化。

**4、matplotlib**

matplotlib是python中优秀的数据可视化的包，Python中最常用的可视化库，由此可与MatLab、Mathematica等科学工具相提并论。它可以进行多种基本图形的可视化操作，并且包含标签、网格、图例等多种进行实体格式化的工具。根据命名就可以看到，它其实是一个matlib的plot库，也就是利用Python将matlib的绘图功能实现了一遍。matplotlib是Python编程语言及其数值数学扩展包。NumPy的可视化操作界面。它为利用通用的图形用户界面工具包，如Tkinter，wxPython，Qt或GTK+向应用程序嵌入式绘图提供了应用程序接口（API）。Matplotlib能够对数据和记过进行可视化，快速理解数据。

**5、Seaborn**

Seaborn主要用于统计模型的可视化，比如热力图等，可以对数据进行概述的同时描绘整体分布。Seaborn基于Matplotlib实现，并高度依赖于后者。

**6、Bokeh**

Bokeh库旨在进行交互式可视化，并且不依赖于Matplotlib，主要通过浏览器以“数据驱动文档”（Data-Driven Documents，d3.js）的形式演示。

**7、Basemap**

Basemap库通过使用Matplotlib中的坐标，为Matplotlib提供了简易地图的支持。Folium库在Basemap的基础上构建，可以制作交互式的Web地图，这类似于使用Bokeh创建的JavaScript小部件。

**8、Plotly**

Poltly库是一个基于Web、用于数据可视化的工具集，通过一系列API接口用于Python等编程语言的实现，Plotly网站上展示了许多强大的图形。

**9、NetworkX**

NetworkX库用于复杂社会网络的分析，它可以处理标准数据格式及非标准数据格式，这使其具备高效性和可扩展性。

**10、Scikit-learn**

Scikit-Learn的主要功能分为六个部分：分类、回归、聚类、数据降维、模型选择、数据预处理。Scikit-learn是完成python大数据机器学习的包，基于BSD开源许可证，它是python数据分析中非常重要的一个模块，它是一个基于NumPy和SciPy构建的开源机器学习工具包。它具有常用的ML算法，可用于预处理，分类，回归以及聚类。算法包括[支持向量机]（Support vector machines），Ridge回归，网格搜索算法（Grid Search algorithm），k均值聚类等等。另外还有样本数据集。API易学易用。在几乎所有平台上的良好性能，它在学术和商业用途中都很受欢迎。SciKit-Learn为常见的机器学习算法提供了简洁而一致的界面，使得机器学习更容易应用于生产系统。它结合了高质量的代码和应用文档，具备易用性和高性能，事实上已成为使用Python进行机器学习的行业标准。

Scikit-Learn自带一些经典的数据集，比如用于分类的iris和digits数据集，还有用于回归分析的数据集。该数据集是一种字典结构，数据存储在.data成员中，输出标签存储在.target成员中。Scikit-Learn建立在Scipy之上，提供了一套常用的机器学习算法，通过一个统一的接口来使用，Scikit-Learn有助于在数据集上实现流行的算法。

**11、TensorFlow**

TensorFlow库由谷歌工作人员开发，是一个数据流图计算的开源库，在机器学习方面的应用表现优越。主要特点是多层节点系统，可以在大型数据集上快速训练人工神经网络，这使得谷歌的语音识别和图像识别成为可能。

**12、Theano**

Theano库主要用于机器学习的需要，它定义了类似于Numpy的多维数组，以及数据操作和表达，在低级别操作上与Numpy结合紧密。Theano库经过了编译，优化了GPU和CPU的使用，在数据密集型计算方面性能更高，结果也更精确。

**13、Keras**

Keras库是用Python编写、用于高级界面中神经网络构建的开源库，具有可扩展性和高度模块化。它使用Theano或TensorFlow作为后端。它设计简约，旨在通过构建紧凑型系统进行快速简单的实验。它的基本思想是基于层次的，其他一切都是围绕它们建立的。数据准备为张量形式，第一层负责张量输入，最后一层负责输出，模型置于两者之间。Keras是人工神经网络和深度学习模型，基于Theano之上，依赖于Numpy和Scipy，利用它可以搭建普通的神经网络和各种深度学习模型，如语言处理、图像识别、自编码器、循环神经网络、递归审计网络、卷积神经网络等。

**14、Scrapy**

Scrapy是专门为爬虫而生的工具，具有URL读取、HTML解析、存储数据等功能，可以使用Twisted异步网络库来处理网络通讯，架构清晰，且包含了各种中间件接口，可以灵活的完成各种需求。

**15、Gensim**

Gensim是用来做文本主题模型的库，常用于处理语言方面的任务，支持TF-IDF、LSA、LDA和Word2Vec在内的多种主题模型算法，支持流式训练，并提供了诸如相似度计算、信息检索等一些常用任务的API接口。

**16、Statsmodels**

Statsmodels库主要用于统计分析，通过应用统计模型估计方法进行数据探索，并进行推断分析。它使用的模型包括线性回归模型、广义线性模型、离散选择模型、时间序列分析模型等，还提供了大量统计分析方面的绘图功能。

**17、NLTK**

Natural Language Toolkit的简称，用于符号和统计自然语言处理等一般任务。NLTK包括许多功能，比如文本标记、分类、名称实体标识、建立显示句子间及句子间依赖性的语料树、词干、语义推理等。所有的构建块允许为不同的任务构建复杂的研究系统，比如情感分析、自动汇总等。

**18、Pattern**

Pattern库将Scrapy库和NLTK库结合在一个大型库中，被设计为用于Web挖掘、自然语言处理、机器学习和网络分析的解决方案。其内部工具有网络爬虫；用于谷歌、推特和维基百科的API；解析树、情感分析等文本分析算法。

**19、Gensim**

Gensim是一个Python的开源库，为更高效地处理大文本数据而设计，实现了向量空间建模和主题建模。Gensim主要用于原始的、非结构化的数字文本处理，实现了层次Dirichlet过程（HDP）、潜在语义分析（LSA）、隐含Dirichlet分布（LDA）以及tf-idf、随机投影、word2vec、document2vec等算法，以分析语料集中重复文本出现的模式。所有算法都是无监督的，只需要输入语料库。

Gensim是用来做文本主题模型的库，常用于处理语言方面的任务，支持TF-IDF、LSA、LDA和Word2Vec在内的多种主题模型算法，支持流式训练，并提供了诸如相似度计算、信息检索等一些常用任务的API接口。

**20、plotnine**

R语言数据可视化的强大之处在于ggplot2，而plotnine相当于是python版的ggplot2，语法与R语言的ggplot2基本一致，无论是从语法简洁性、作图灵活性、美观度等方面，相对于传统可视化模块，plotnine均有不错的表现。plotnine是将ggplot2移植到python上，在python上完全重现ggplot2的功能。

**21、OpenCV**

OpenCV（open source computer vision library）是一个基于BSD许可（开源）发行的跨平台计算机视觉库。它轻量级而且高效，由一系列C函数和少量C++类构成，同时提供了Python、Ruby、MATLAB等语言的接口，实现了图像处理和计算机视觉方面的很多通用算法。OpenCV用C++语言编写，它的主要接口也是C++语言，但是依然保留了大量的C语言接口。

在计算机视觉项目的开发中，OpenCV作为较大众的开源库，拥有了丰富的常用图像处理函数库，能够快速的实现一些图像处理和识别的任务。此外，OpenCV还提供了Java、python、cuda等的使用接口、机器学习的基础算法调用，从而使得图像处理和图像分析变得更加易于上手，让开发人员更多的精力花在算法的设计上。

**22、IPython**

IPython是Python科学计算标准工具集的组成部分，它可以把很多东西联系到一起，有点类似一个增强版的Python shell。目的是为了提高编程，测试和调试Python代码的速度，而且不用print，回车就能打印。

**23、GraphLabGreate**

GraphLabGreate是一个Python库，由C++引擎支持，可以快速构建大型高性能数据产品。可以在计算机上以交互的速度分析以T为计量单位的数据量，在单一平台上可以分析表格数据、曲线、文字、图像。

**24、PuLP**

PuLP是一个用Python编写的线性编程模型。它能产生线性文件，能调用高度优化的求解器，GLPK，COINCLP/CBC，CPLEX，和GUROBI，来求解这些线性问题。

**25、Spark**

Spark提供了弹性分布数据集（RDD），一个按照聚类的节点进行分区的元素的集合，它可以在并行计算中使用。RDDs可以从一个Hadoop文件系统中的文件（或者其他的Hadoop支持的文件系统的文件）来创建，或者是驱动程序中其他的已经存在的标量数据集合，把它进行变换。

Spark可以实现在并行操作中变量的共享。默认情况下，当Spark在并行情况下运行一个函数作为一组不同节点上的任务时，它把每一个函数中用到的变量拷贝一份送到每一任务。有时，一个变量需要被许多任务和驱动程序共享。Spark支持两种方式的共享变量：广播变量，它可以用来在所有的节点上缓存数据。另一种方式是累加器，这是一种只能用作执行加法的变量，例如在计数器中和加法运算中。