机器学习基础

所谓“机器学习”就是利用计算机将纷繁复杂的数据处理成有用的信息，这样就可以发掘出数据带来的意义以及隐藏在数据背后的规律。现如今，“机器学习”和“大数据”可以说是 IT 行业中最热点的两个词汇，而无论是“机器学习”还是“大数据”最终要解决的问题本质上是一样的，用最为直白的话来说就是用现有的数据去预测将来的状况。

按照问题的“输入”和“输出”，我们可以将用计算机解决的问题分为四大类：

1. 输入的信息是精确的，要求输出最优解。
2. 输入的信息是精确的，无法找到最优解。
3. 输入的信息是模糊的，要求输出最优解。
4. 输入的信息是模糊的，无法找到最优解。

在上面的四大类问题中，第 1 类问题是计算机最擅长解决的，这类问题其实就是“数值计算”和“逻辑推理”方面的问题，而传统意义上的人工智能也就是利用逻辑推理来解决问题（如早期的“人机对弈”）。一直以来，我们都习惯于将计算机称为“电脑”，而基于“冯诺依曼”体系结构的“电脑”实际上只是实现了“人脑”理性思维这部分的功能，而且在这一点上“电脑”通常是优于“人脑”的，而“人脑”在处理输入模糊信息时表现出来的强大的处理能力，在今天看来也不是“电脑”可以完全企及的。所以我们研究人工智能也好，研究机器学习也好，是希望输入模糊信息时，计算机能够给出满意的甚至是最优的答案。

至此，我们可以给“机器学习”下一个定义：机器学习是一门专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身性能的学科。机器学习目前已经广泛的应用到生产生活的各个领域，以下列举了一些经典的场景：

1. 搜索引擎：根据搜索和使用习惯，优化下一次搜索的结果。
2. 电商网站：自动推荐你可能感兴趣的商品。
3. 贷款申请：通过你最近的金融活动信息进行综合评定。
4. 图像识别：自动识别图片中有没有不和谐的内容。

机器学习可以分为监督学习和非监督学习。监督学习是从给定的训练数据集中学习得到一个函数，当新的数据到来时，可以根据这个函数预测结果，监督学习的训练集包括输入和输出，也可以说是特征和目标。监督学习的目标是由人来标注的，而非监督学习的数据没有类别信息，训练集也没有人为标注结果，通过无监督学习可以减少数据特征的维度，以便我们可以使用二维或三维图形更加直观地展示数据信息 。

实现机器学习的一般步骤：

1. 数据收集
2. 数据准备
3. 数据分析
4. 训练算法
5. 测试算法
6. 应用算法

NumPy 与 Pandas 之间的区别?

Numpy 与 Pandas 的区别

Numpy
Numpy 是以矩阵为基础的数学计算模块，纯数学。
Scipy
Scipy 基于 Numpy，科学计算库，有一些高阶抽象和物理模型。比方说做个傅立叶变换，这是纯数学的，用 Numpy；做个滤波器，这属于信号处理模型了，在 Scipy 里找。
Pandas
Pandas 提供了一套名为 DataFrame 的数据结构，比较契合统计分析中的表结构，并且提供了计算接口，可用 Numpy 或其它方式进行计算。
总结
NumPy 中的 ndarray 用于处理多维数值型数组，重点在于进行数值运算，无索引
*Pandas 中的 Series 类似于 DataFrame 的子集，DataFrame 中的每一列都可以看作是一个 Series，有索引，方便进行数据的查询，筛选，所以 Pandas 重点在于进行数据分析在数学与统计方法上，NumPy 中的 ndarray 只能进行数值型统计，而 Pandas 中的 DataFrame 既可以进行数值型，也可以进行非数值型统计。基于可以容纳不同的数据类型而定
1.NumPy
数值型，重点在于进行矩阵运算

N 维数组容器，Numpy 是以矩阵为基础的数学计算模块。

Numpy 专门针对 ndarray 的操作和运算进行了设计，所以数组的存储效率和输入输出性能远优于 Python 中的嵌套列表，数组越大，Numpy 的优势就越明显。Numpy 系统是 Python 的一种开源的数值计算扩展。这种工具可用来存储和处理大型矩阵，比 Python 自身的嵌套列表（nested list structure)结构要高效的多（该结构也可以用来表示矩阵（matrix））。
ndarray
所有元素的类型相同，存储元素时内存可以连续；Python 里 list 中的元素类型任意，只能通过寻址方式找到下一个元素
ndarray 矩阵结构与 matlab 或者 C++或者 fortran 都很不一样，没有行优先或者列优先的概念
ndarray 支持并行化运算（向量化运算）,类似于 Matlab
Numpy 底层使用 C 语言编写，内部解除了 GIL（全局解释器锁），其对数组的操作速度不受 Python 解释器的限制，效率远高于纯 Python 代码
2.Pandas
多数据类型，重点在于进行数据分析

pansdas 是基于 Numpy 的一种工具,该工具是为了解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。pandas 提供了大量快速便捷地处理数据的函数和方法。使 Python 成为强大而高效的数据分析环境的重要因素之一。

Series

参看书：Series 是一种类似于一维数组的对象，它由一组数据(各种 NumPy 数据类型)以及一组与之相关的数据标签(即索引)组成。**小规模数据**
类似于一维数组，索引对象的数据类型是一致的
有了索引标签，方便实际问题中进行信息提取与筛选
python 字典类型的数据可以直接给 Series 对象
Series 可以运用 ndarray 或字典的几乎所有索引操作和函数，融合了字典和 ndarray 的优点。

属性	说明
values	获取数组
index	获取索引

name　values 的 name
index.name　索引的 name

DataFrame

DataFrame 就是按照 column 和 index 组织起来的数据集合，类似于 excel 表格，也类似于基本的 database 结构。DataFrame 是一个表格型的数据结构，它含有一组有序的列，每列可以是不同的值类型（数值、字符串、布尔值等）。DataFrame 既有行索引也有列索引，它可以被看做由 Series 组成的字典（共用同一个索引）。

1DataFrame 范例

	year	state	pop	debt
one	2000	Ohio	1.5	16.5
two	2001	Ohio	1.7	16.5
three	2002	Ohio	3.6	16.5
four	2001	Nevada	2.4	16.5
five	2002	Nevada	2.9	16.5
six	2003	Nevada	3.2	16.5

Series 类似于 DataFrame 的子集，从上表可以看出，每一列都对应这一个 Series

matplotlib三层架构：

1，Backend：Canvas类专门用于调配计算资源，用于底层的图形绘制

2，Artist

fig：画板

axes：坐标系 可以有多个

3，scripting：

绘制的实际内容：

1，刻度

2，线