深入浅出机器学习

-基于sklearn的应用实践

作者

2017-10-1

目 录

[第一章 Python语言入门 1](#_Toc25600)

[1.1 Python简介 1](#_Toc14890)

[1.1.1 Python的诞生 1](#_Toc19354)

[1.1.2 Python的提出背景 1](#_Toc20340)

[1.1.3 Python的应用 2](#_Toc32661)

[1.1.4 Python的缺点 2](#_Toc20374)

[1.2 安装配置Python运行环境 3](#_Toc17427)

[1.2.1 集成开发环境 3](#_Toc26640)

[1.2.2 Anaconda 4](#_Toc10992)

[1.3 Python基本语法 8](#_Toc31801)

[1.3.1数据类型——数字和字符串 8](#_Toc1868)

[1.3.2 数据类型——列表list 9](#_Toc17211)

[1.3.3数据类型——元组tuple 10](#_Toc10794)

[1.3.4 数据类型——字典 11](#_Toc20122)

[1.3.5基本语句 11](#_Toc6466)

[1.4 第三方包 13](#_Toc1618)

[1.4.1 NumPy简介 13](#_Toc14655)

[1.4.2 SciPy简介 14](#_Toc20114)

[1.4.3 Matplotlib简介 14](#_Toc28406)

[1.4.4 scikit learn简介 15](#_Toc27804)

[1.5 Python范例 15](#_Toc5987)

[第二章 基础知识 19](#_Toc11091)

[2.1 线性代数基本知识 19](#_Toc29799)

[2.1.1 向量与矩阵相关的基础知识 19](#_Toc1398)

[2.1.2 n维向量和向量空间 22](#_Toc23706)

[2.1.3图像解方程组 23](#_Toc2728)

[2.1.4 消元法解方程组 29](#_Toc7597)

[2.1.5 矩阵乘法和逆矩阵 31](#_Toc19843)

[2.1.6 特征分解 35](#_Toc23935)

[2.2 概率论基本知识 37](#_Toc20167)

[2.2.1 概率、条件概率 38](#_Toc22572)

[2.2.2 概率分布 38](#_Toc8564)

[2.2.3联合概率、边缘概率和条件概率 39](#_Toc23972)

[2.2.4独立性 41](#_Toc18343)

[2.2.5期望、方差和协方差 42](#_Toc14807)

[2.2.6常用概率分布 43](#_Toc28364)

[2.3 机器学习基本知识 45](#_Toc6351)

[2.3.1 何为机器学习 45](#_Toc15820)

[2.3.2 机器学习的主要任务 45](#_Toc11446)

[2.3.3 机器学习的基本步骤 45](#_Toc12677)

[第三章 scikit-learn基本知识 47](#_Toc11489)

[3.1 sklearn结构 47](#_Toc25121)

[3.2 加载范例数据库 47](#_Toc15958)

[第四章 预处理 48](#_Toc8201)

[4.1 标准化、归一化与正则化 48](#_Toc6507)

[4.1.1 离差标准化 52](#_Toc3460)

[4.1.2 标准差标准化 53](#_Toc7532)

[4.1.3 规范化 56](#_Toc16967)

[4.2 离散特征二值化 57](#_Toc19056)

[4.2.1 特征二值化（Binarization） 62](#_Toc12410)

[4.2.2 标签二值化（Label binarization） 63](#_Toc6844)

[4.2.3 类别特征编码 63](#_Toc32517)

[4.2.4 .标签编码（Label encoding） 63](#_Toc16492)

[4.3 特征中含异常值时 63](#_Toc27425)

[4.4 生成多项式特征 64](#_Toc12601)

[4.5 插补处理 64](#_Toc4984)

[4.6 独热编码 64](#_Toc31124)

[4.7 特征选择 64](#_Toc24934)

[4.8 降维 64](#_Toc22894)

[第五章 一般线性模型 65](#_Toc10514)

[5.1 线性模型概述 65](#_Toc96)

[5.2 线性判别分析 P60 67](#_Toc8055)

[第六章 非线性回归 68](#_Toc11292)

[6.1 Logistic Regression 68](#_Toc32155)

[6.2 多项式回归 72](#_Toc17223)

[第七章 k-邻近算法 75](#_Toc8828)

[7.1 k-邻近算法概述 75](#_Toc6181)

[7.1.1 基本思想 75](#_Toc20867)

[7.1.2 算法概述 75](#_Toc6601)

[7.2 算法范例：k-邻近算法应用范例 77](#_Toc22914)

[7.3 小结 77](#_Toc12200)

[第八章 支持向量机 78](#_Toc4215)

[8.1 概述 78](#_Toc6835)

[8.2 支持向量机算法 78](#_Toc12983)

[8.2.1 线性分类器 78](#_Toc20590)

[8.2.2 对偶问题 81](#_Toc26674)

[8.2.3 软间隔 82](#_Toc10143)

[8.2.4 核函数 83](#_Toc2269)

[8.3 算法范例：支持向量机分类应用范例 84](#_Toc11141)

[8.4 小结 84](#_Toc16790)

[第九章 决策树 85](#_Toc20922)

[9.1 简介 85](#_Toc1273)

[9.1.1 决策树表示法 85](#_Toc28714)

[9.2 决策树算法 85](#_Toc20475)

[9.3 决策树学习的适用问题 88](#_Toc31690)

[9.4 算法范例：决策树回归范例 88](#_Toc18507)

[9.5 小结 88](#_Toc7661)

[第十章 人工神经网络 89](#_Toc10951)

[10.1 简介 89](#_Toc23919)

[10.2 人工神经元结构 89](#_Toc29315)

[10.3 人工神经网络算法（BP算法） 90](#_Toc16911)

[10.4 人工神经网络算法适用的问题 95](#_Toc9534)

[10.5 算法范例：BP神经网络应用范例 95](#_Toc19368)

[10.6 小结 95](#_Toc22578)

[第十一章 AdaBoost算法 96](#_Toc9263)

[11.1 基本思想 96](#_Toc22684)

[11.2 算法流程 96](#_Toc27290)

[第十二章 随机森林 99](#_Toc32127)

[12.1 简介 99](#_Toc9765)

[12.2 随机森林基本原理 99](#_Toc5217)

[12.3 随机森林实现过程 100](#_Toc11405)

[第十三章 模型评价 101](#_Toc10523)

[13.1 模型评价指标 101](#_Toc23343)

[13.2 模型评价范例 101](#_Toc10477)

[13.3 模型选择范例 101](#_Toc16952)

# Python语言入门

## Python简介

### Python的诞生

Python是著名的“龟叔”Guido van Rossum在1989年圣诞节期间，为了打发无聊的圣诞节而编写的一个编程语言，而Python名称来自Guido挚爱的电视剧Monty Python’s Flying Circus。

### Python的提出背景

当你用一种语言开始作真正的软件开发时，你除了编写代码外，还需要很多基本的已经写好的现成的东西，来帮助你加快开发进度。比如说，要编写一个电子邮件客户端，如果先从最底层开始编写网络协议相关的代码，那估计一年半载也开发不出来。高级编程语言通常都会提供一个比较完善的基础代码库，让你能直接调用，比如，针对电子邮件协议的SMTP库，针对桌面环境的GUI库，在这些已有的代码库的基础上开发，一个电子邮件客户端几天就能开发出来。Python就为我们提供了非常完善的基础代码库，覆盖了网络、文件、GUI、数据库、文本等大量内容，被形象地称作“内置电池（batteries included）”。用Python开发，许多功能不必从零编写，直接使用现成的即可。除了内置的库外，Python还有大量的第三方库，也就是别人开发的，供你直接使用的东西。当然，如果你开发的代码通过很好的封装，也可以作为第三方库给别人使用。

许多大型网站就是用Python开发的，例如YouTube、Instagram，还有国内的豆瓣。很多大公司，包括Google、Yahoo等，甚至NASA（美国航空航天局）都大量地使用Python。

龟叔给Python的定位是“优雅”、“明确”、“简单”，所以Python程序看上去总是简单易懂，初学者学Python，不但入门容易，而且将来深入下去，可以编写那些非常非常复杂的程序。

总的来说，Python的哲学就是简单优雅，尽量写容易看明白的代码，尽量写少的代码。如果一个资深程序员向你炫耀他写的晦涩难懂、动不动就几万行的代码，你可以尽情地嘲笑他。

### Python的应用

既然Python功能这么强大，那适合开发哪些类型的应用呢？

Web开发

Python定义了WSGI标准应用接口协调http服务器与基于Python的Web程序之间的沟通。

GUI开发

用wxPython或者PyQt来开发跨平台的桌面软件。

操作系统

大多数Linux发布版以及NetBSD、OpenBSD和MacOS都集成了Python,Python标准库包含了多个调用作业系统功能的库。

多媒体

可用于计算机游戏三维场景制作。

### Python的缺点

任何编程语言都有缺点，Python也不例外。优点说过了，那Python有哪些缺点呢？

第一个缺点就是运行速度慢，和C程序相比非常慢，因为Python是解释型语言，你的代码在执行时会一行一行地翻译成CPU能理解的机器码，这个翻译过程非常耗时，所以很慢。而C程序是运行前直接编译成CPU能执行的机器码，所以非常快。

但是大量的应用程序不需要这么快的运行速度，因为用户根本感觉不出来。例如开发一个下载MP3的网络应用程序，C程序的运行时间需要0.001秒，而Python程序的运行时间需要0.1秒，慢了100倍，但由于网络更慢，需要等待1秒，你想，用户能感觉到1.001秒和1.1秒的区别吗？这就好比F1赛车和普通的出租车在北京三环路上行驶的道理一样，虽然F1赛车理论时速高达400公里，但由于三环路堵车的时速只有20公里，因此，作为乘客，你感觉的时速永远是20公里。

第二个缺点就是代码不能加密。如果要发布你的Python程序，实际上就是发布源代码，这一点跟C语言不同，C语言不用发布源代码，只需要把编译后的机器码（也就是你在Windows上常见的xxx.exe文件）发布出去。要从机器码反推出C代码是不可能的，所以，凡是编译型的语言，都没有这个问题，而解释型的语言，则必须把源码发布出去。

这个缺点仅限于你要编写的软件需要卖给别人挣钱的时候。好消息是目前的互联网时代，靠卖软件授权的商业模式越来越少了，靠网站和移动应用卖服务的模式越来越多了，后一种模式不需要把源码给别人。再说了，现在如火如荼的开源运动和互联网自由开放的精神是一致的，互联网上有无数非常优秀的像Linux一样的开源代码，我们千万不要高估自己写的代码真的有非常大的“商业价值”。那些大公司的代码不愿意开放的更重要的原因是代码写得太烂了，一旦开源，就没人敢用他们的产品了。

## 安装配置Python运行环境

### 集成开发环境

集成开发环境(IDE，Integrated Development Environment )是用于提供程序开发环境的应用程序，一般包括代码编辑器、编译器、调试器和图形用户界面工具。集成了代码编写功能、分析功能、编译功能、调试功能等一体化的开发软件服务套。所有具备这一特性的软件或者软件套(组)都可以叫集成开发环境。如微软的Visual Studio系列，Borland的C++ Builder、Delphi系列等。该程序可以独立运行，也可以和其它程序并用。IDE多被用于开发HTML应用软件。例如，许多人在设计网站时使用IDE(如HomeSite、DreamWeaver等)，因为很多项任务会自动生成。

而Python IDE工具包括文本工具类IDE和集成工具类IDE，常见的文本工具类IDE包括IDLE和Sublime Text，常见的集成工具类IDE包括Pycharm和Spyder。下面分别简单介绍这四种IDE工具的特点：

1、IDLE具有如下特点：

* 自带
* 默认
* 常用
* 入门级

适用于Python初学者，这个功能简单直接，适合编写代码在300个之内。

2、Sublime Text是一种专业编程，适用于专业人员程序员，具有特点如下：

专为程序员开发的第三方专用编程工具

专业编程体验增强

多种编程风格

工具非注册免费试使用

3、Pycharm特点如下：

* 社区版免费
* 简单
* 适用于较复杂工程

### Anaconda

Anaconda是一个集成各类Python工具的集成平台。适用于科学计算和数据分析方面，也具有如下特点：

* 开源免费
* 支持近800个第三方库
* 包含多个主流工具
* 适合数据计算领域开发

显而易见，因为Anaconda是开源免费的，包括的第三方库也是非常多，并且也是非常好用的，同时也是个跨平台工具，适用于Win/Linux/OS X，系统，在国际上尤其是科学计算，数据分析领域应用非常普遍，所以我们在讲Python语言这个系类以Anaconda平台作为我们的开发工具。

为了更好的理解Anaconda，下面我们会介绍几个支撑Anaconda的工具。Anaconda来自于一个包管理和环境管理工具——conda.简单介绍一下conda,它是一个工具，用于包管理和环境管理；包管理与pip类似，管理Python第三方库；环境管理能够允许用户使用不同版本的Python，并能灵活切换。 Anaconda是一个基于conda这样一个管理工具的集合，它包括conda、某版本的Python以及一批第三方库。由于conda是一个第三方库，它将工具、第三方库、Python版本、conda都当做包来同等对待。

其中还有一个编写Python代码和调试代码工具Spyder，是一个非常优秀的第三方工具。

Anaconda还包含一个交互式编程环境——IPython。IPython是一个功能强大的交互式shell，适合进行交互式数据可视化和GUI相关应用。

所以掌握好Anaconda IDE几乎可以应对Python在科学计算和数据处理所有这方面的开发需求。

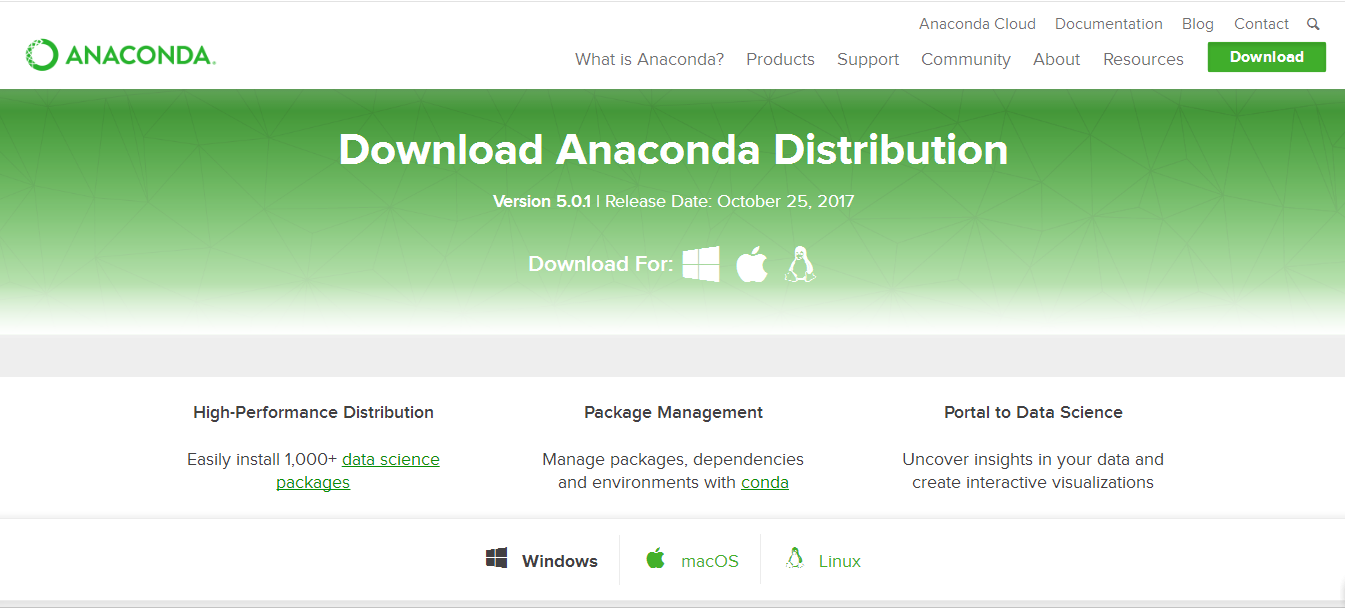
下面简单介绍Anaconda的安装过程：

我们安装Anaconda的过程可以由下面三个步骤来完成：

步骤一：下载的途径直接到Anaconda的官方网站，网址https://www.anaconda.com然后你会看到如下的一个界面:

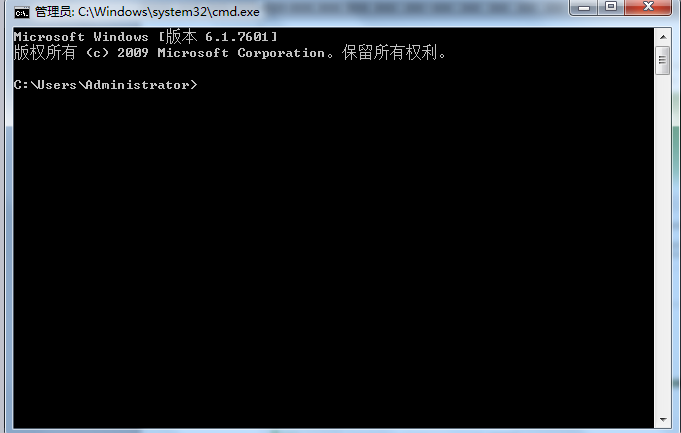


步骤二：然后点击“Download”，会弹出“Download Anaconda Distribution”的界面，直接根据你是Windows操作系统还是iOS系统，选定相对应的直接下载即可。

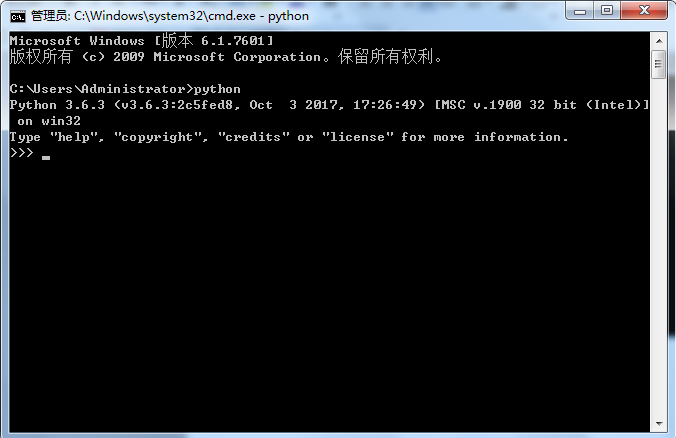


步骤三：测试Anaconda安装是否成功

在Windows下点开始，输入cmd命令，弹出如下界面：

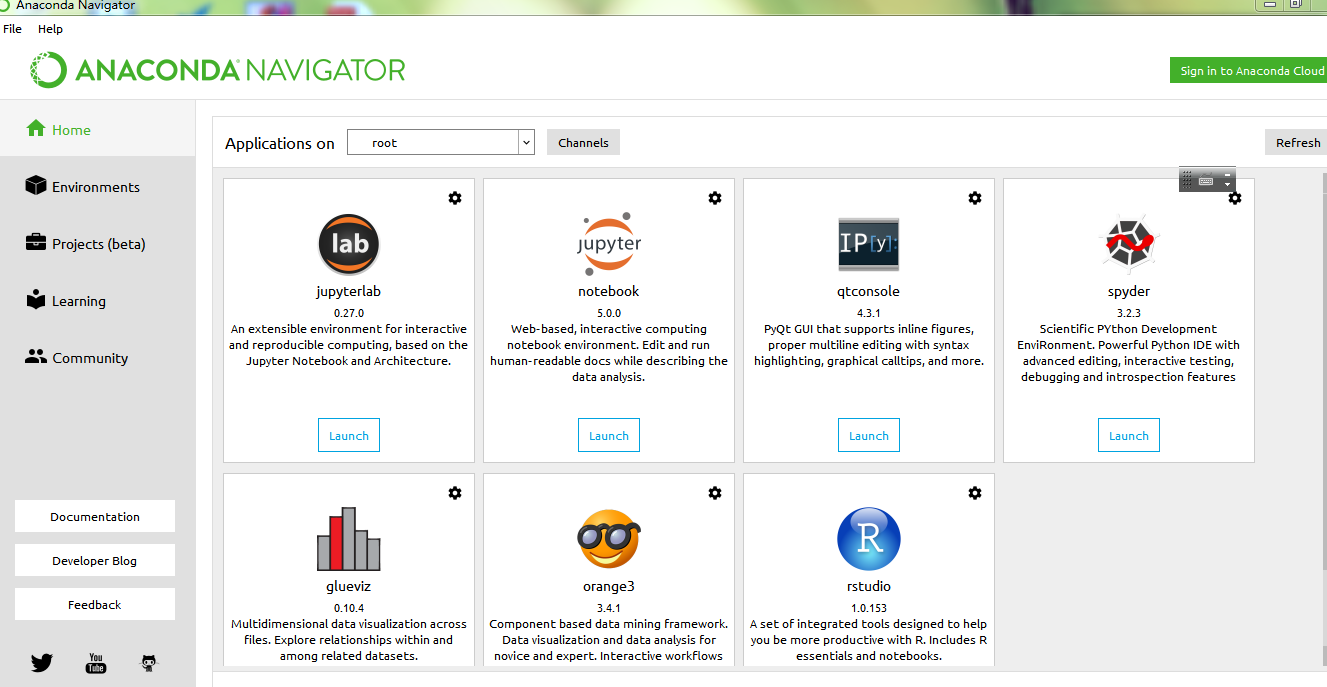


然后输入Python，弹出如下界面：



那就代表Anaconda安装成功了。

最后使用Anaconda IDE集成开发工具,点击开始，显示电脑所有程序中，点击Anaconda Navigator，来启动Anaconda,启动后界面如下：



## Python基本语法

Python是一种计算机编程语言。计算机编程语言和我们日常使用的自然语言有所不同，最大的区别就是，自然语言在不同的语境下有不同的理解，而计算机要根据编程语言执行任务，就必须保证编程语言写出的程序决不能有歧义，所以，任何一种编程语言都有自己的一套语法，编译器或者解释器就是负责把符合语法的程序代码转换成CPU能够执行的机器码，然后执行。Python也不例外。

任何计算机系统，其工作的实质都在于，将输入后的数据处理后再输出，即：

输入→处理→输出

Python的语法比较简单，采用缩进方式。缩进有利有弊。好处是强迫你写出格式化的代码，但没有规定缩进是几个空格还是Tab。按照约定俗成的管理，应该始终坚持使用4个空格的缩进。缩进的另一个好处是强迫你写出缩进较少的代码，你会倾向于把一段很长的代码拆分成若干函数，从而得到缩进较少的代码。

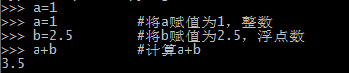
缩进的坏处就是“复制－粘贴”功能失效了，这是麻烦的地方。当你重构代码时，粘贴过去的代码必须重新检查缩进是否正确。此外，IDE很难像格式化Java代码那样格式化Python代码。

最后，请务必注意，Python程序是大小写敏感的，如果写错了大小写，程序会报错。

### 1.3.1数据类型——数字和字符串

计算机，顾名思义就是可以做数学计算的机器，因此，计算机程序理所当然地可以处理各种数值。但是，计算机能处理的远不止数值，还可以处理文本、图形、音频、视频、网页等各种各样的数据，不同的数据，需要定义不同的数据类型。数据类型的内容有很多，如字典、文件、列表和元组。

作为动态类型的语言，Python中使用数字无需先声明它的类型，如下所示。



在上面的例子中，我们分别将整数1和浮点数2.5赋值给了变量a和b，“=”就是赋值的意思，“+”为运算符，在Python中，除了加减乘除取余等运算符外，还有逻辑运算符如位移运算符等等。接下来直接输入公式即可完成计算。每句后面以“#”开头的文本时注释，它对于脚本的运行没有任何影响，目前可以完全被我们忽略掉——当然，在复杂脚本里面使用注释说明自己编程时的思路算法也是一个很好的习惯。

Python中的字符串是用来表示和储存文本的，通常由单引号、双引号或者三引号包围，可以包含数字、字母以及一些控制字符，如换行符（/n）、制表符（/t）、和回车（/r）等等。字符串的形式一般如下所示。

IMG_256

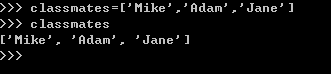
接下来用实例演示转义字符在字符串中的使用。



这个例子里面，print语句的作用是输出字符串，只有在使用print输出字符串时才能够解释字符串中的转义字符。需要注意的是，在Python3中，print的正确语法为print（），否则就会显示错误。

### 1.3.2 数据类型——列表list

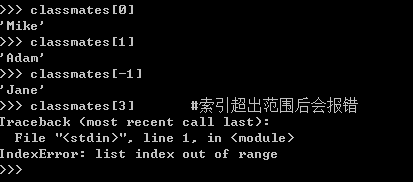
Python内置的一种数据类型是列表：list。List是一种有序的集合，可以随时添加和删除其中的元素。比如，列出班里所有同学的名字，就可以用一个list来表示：



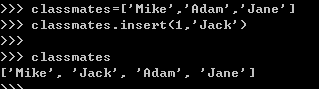
在上面的例子里，变量classmates就是一个list。我们还可以用len()函数来获取list元素的个数。

1509348425(1)

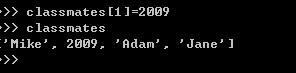
用索引来访问list中每一个位置的元素，记得索引是从0开始的。当索引超出了范围时，Python会报一个IndexError错误，所以，要确保索引不要越界，记得最后一个元素的索引是len(classmates) - 1。如果要取最后一个元素，除了计算索引位置外，还可以用-1做索引，直接获取最后一个元素。



List是一个可变的有序表，所以，可以往list中追加元素到末尾，也可以把元素插入到指定的位置，要删除指定位置的元素，用pop（）方法，



要把某个元素替换成别的元素，可以直接赋值给对应的索引位置，list里面的元素的数据类型也可以不同，比如下图的这个例子。

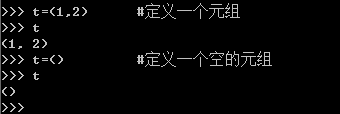


不过因为list里面可以包含不同的数据类型，这也造成它运行起来可能会比较慢，为了弥补这个缺陷，我们将来会引用数组这一个概念。

### 1.3.3数据类型——元组tuple

接下来我们来讨论另一种有序列表，叫做元组：tuple。tuple和list非常类似，但是tuple一旦初始化就不能修改，比如同样是列出同学的名字，现在，classmates这个tuple不能变了，它也没insert()这样的方法。其他获取元素的方法和list是一样的，你可以正常地使用classmates[0]，classmates[-1]，但不能赋值成另外的元素。

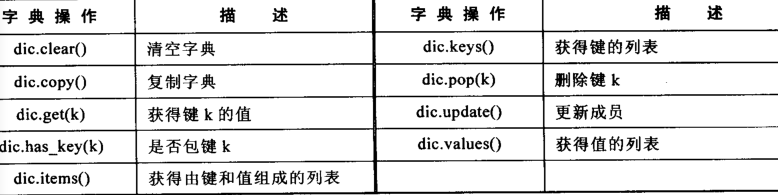
不可变的tuple有什么意义？因为tuple不可变，所以代码更安全。如果可能，能用tuple代替list就尽量用tuple。所以当你定义一个tuple的时候，tuple的元素就必须被确定下来。



需要注意的是，要定义一个只有1个元素的tuple，如果你这么定义：t=(1),定义的不是tuple，是1这个数！这是因为括号()既可以表示tuple，又可以表示数学公式中的小括号，这就产生了歧义，因此，Python规定，这种情况下，按小括号进行计算，计算结果自然是1。所以，只有1个元素的tuple定义时必须加一个逗号，来消除歧义：t=(1,)。而Python在显示只有1个元素的tuple时，也会加一个逗号，以免你误解成数学计算意义上的括号。

### 1.3.4 数据类型——字典

字典是Python中比较特别的一类数据类型，以大括号“{}”包围的数据集合。字典与列表的最大不同在于，字典是无序的，在字典中是通过键来访问成员的。字典也是可变的，可以包含任何其他类型，不过，字典中的成员位置只是象征性的，并不能通过其位置来访问该成员。字典中的成员是以“键：值”的形式来声明的，常用字典操作都在下面这张表里：



### 1.3.5基本语句

语句是Python脚本的基础，它能帮助我们根据一定的条件来执行脚本中的不同语句，从而完成不同的功能。

首先，我们来看条件语句。计算机之所以能做很多自动化的任务，因为它可以自己做条件判断。，在Python程序中，它是if语句是基本的的条件测试语句，用来判断可能遇到的不同情况，以及在不同情况下应该进行的下一步操作。它的基本形式如下：

If <条件>: #设为条件A

<语句> #要用缩进来表示语句处于if语句之中，设为语句A

elif <条件>： #设为条件B

<语句> #设为语句B

else：

<语句> #设为语句C

用简单的话来说就是，首先判断输入的数据是否满足条件A，如果满足就执行条件A，不满足的话就接着判断是否符合条件B的要求，如果满足了就执行语句B，否则就执行语句C。

第二个要介绍的是循环语句。要计算1+2+3，我们可以直接写表达式：

IMG_256

要计算1+2+3+...+10，勉强也能写出来。但是，要计算1+2+3+...+10000，直接写表达式就不可能了。为了让计算机能计算成千上万次的重复运算，我们就需要循环语句。

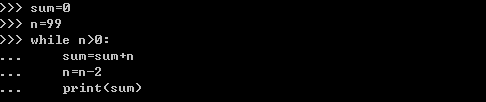
Python的循环有两种，一种是for...in循环，依次把list或tuple中的每个元素迭代出来，如下面的例子所示。

IMG_256

执行这段代码，会依次打印names的每一个元素：

IMG_256

第二种循环是while循环，只要条件满足，就不断循环，条件不满足时退出循环。比如我们要计算100以内所有奇数之和，可以用while循环实现：



上面这段例子就是在循环内部变量n不断自减，直到变为-1时，不再满足while条件，循环退出。

while语句与for语句不同的是，while语句中只有在测试条件为假时才会停止，所以在while的语句块中一定要包含改变条件的语句，以保证循环能够结束，避免死循环的出现。

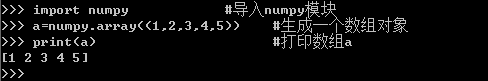
## 第三方包

Python本身就内置了很多非常有用的模块，只要安装完毕，这些模块就可以立刻使用。除了内建的模块外，Python还有大量的第三方模块。基本上，所有的第三方模块都会在PyPI - the Python Package Index上注册，只要找到对应的模块名字，即可用pip安装。而由于前面本书已经推荐大家安装了Anaconda，一些常用第三方库已经包括在里面，省去了我们很多需要自行安装的环节。

科学计算是计算机应用的主要内容之一，也是机器学习的很重要的部分，NumPy和SciPy是Python中用以实现科学计算的模块包，Matplotlib是绘制2D图形的一种模块，Sklearn包含有完善的文档和丰富的机器学习算法，这也是Python语言如此好用的原因之一。接下来我们主要介绍一下这几个第三方包。

### 1.4.1 NumPy简介

NumPy提供了Python没有提供的数组对象，使用了NumPy的数组对象就可以创建类似于C语言中的数组。使用NumPy时要先导入NumPy模块，如下图所示。

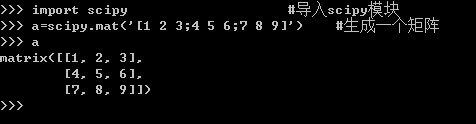


上面这个例子里，

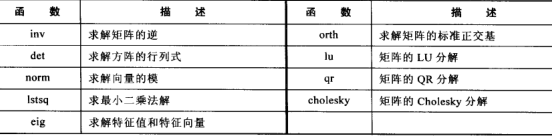
NumPy主要为后面的SciPy和sklearn提供了基础的数据结构。

### 1.4.2 SciPy简介

SciPy模块依赖于NumPy，但是SciPy提供了更多的数学工具。使用SciPy不仅可以进行矩阵运算，还可以用来求解线性方程组、进行积分运算、优化等。SciPy的功能非常接近Matlab。SciPy简单的运用如下图所示：



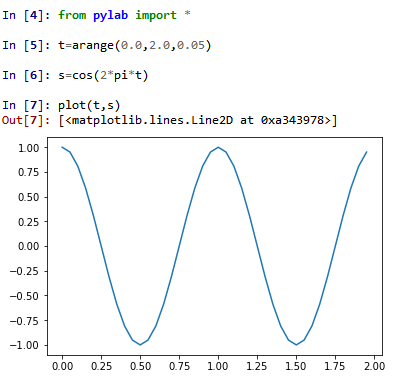
使用SciPy可以完成矩阵分解、线性方程组求解和多项式求根数学运算。SciPy中的函数名与Matlab中的函数名大部分都相同，用法也差不多。熟悉Matlab的用户可以很快的熟悉这个第三方包。下图是一些常用的函数，以后在机器学习中很多也都会被用到：



### 1.4.3 Matplotlib简介

在Python中，我们可以使用Matplotlib来绘制函数图形，它依赖于NumPy与SciPy。和SciPy一样，其中的大部分函数都和Matlab中的函数名相同。它可以绘制多种形式的图形包括普通的线图、直方图、饼图、散点图以及误差线图等。

使用Matplotlib绘制图形主要使用它的plot函数，Matplotlib还包含了一些用于设置x、y轴标签文本以及图形标题的函数。同时它还支持一部分Tex排版命令，可以较好地显示数学公式。接下来我们将在IPython交互式命令行中使用Matplotlib绘制cos图形。



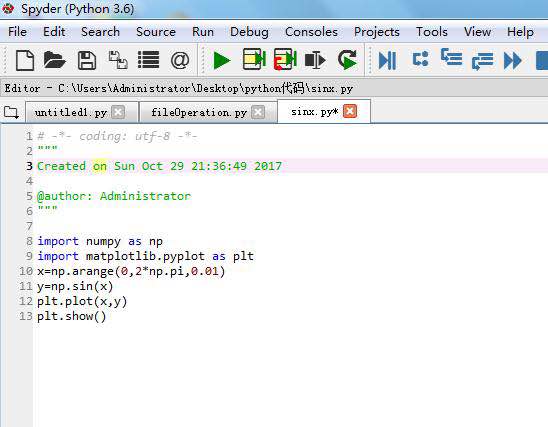
### 1.4.4 scikit learn简介

Sklearn包括了很多种机器学习的类型，比如说监督学习和非监督学习，在监督学习中，又可以分为分类学习和线性回归学习。当我们想训练我们的电脑来识别一些图片，比如猫狗猪牛等等时，这个叫做分类学习。当我们想预测房价高低时，用到的就是线性回归学习。非监督学习就是指，我们不告诉计算机你看到了什么东西，让计算机自己琢磨出来这些数据有什么不同，再画一条线区分开来这些数据的类型。从上面的大概了解可以看出来Sklearn是非常实用的。

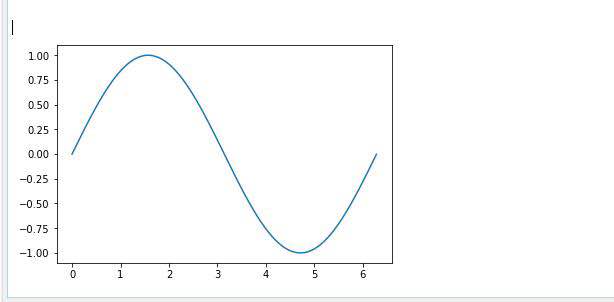
## Python范例

前面我们已经把Python安装到自己的电脑上了，并且把Python的基本语句也已经做了详细的介绍，那我们下面来实践下，本书给出了两个例子供大家参考。

打开Anaconda的代码编辑工具Spyder,在左边编辑区输入如下代码，我们来画出一个sin（x）的图像。



输入如上的代码后，点击上方Run的命令，可以得到如下的一个sinx图形



现在我们来分析一下这个脚本的内容。



首先，在第一行，import作为调用函数，因为要画正弦曲线要先设定一下x的取值范围，从0到2π，这个要用到numpy模块，用import调用函数，来导入模块numpy并以np作为别名

第二步由于是要画图，所以需要用到matplotlib.pyplot模块中plot方法



第三步Arange函数是numpy包中一个函数，np.pi表示π，这个表示x从0到2π，以0.01步进。



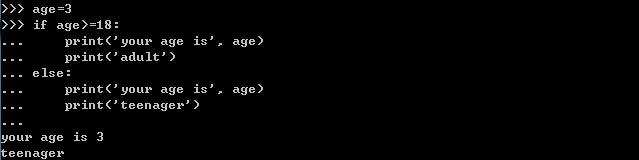
第四步就是定义下y的值



第五步就是画图呈现x轴y轴的意思



最后一步就是显示图的意思。

接下来我们利用Python来运行一个简单的if语句，打开开始菜单，输入cmd命令，弹出提示符窗口，敲入Python后，然后输入下面简单的if语句。

同样的，我们来分析一下这个脚本的内容：



这个意思就是说年龄是为3，如果年龄是大于等于18，那么就输出两行，一行是输出你的年龄，写出年龄，另外一行是输出成年人。根据Python的缩进规则，如果if语句判断是true，就把缩进的两行print语句执行了，否则，什么也不做。

也可以给if添加一个else语句，意思是，如果if判断是False,不要执行if的内容，去把else执行了,如下面的一个程序：



这个if语句进行了一个条件判断，因为之前输出的年龄是3，因为是3是小于18，那就执行else后的语句，然后同样输出两行，



一行是输出你的年龄是3，另外输出是青年人这一行。

参考资料：

1. 《征服PYTHON：语言基础与典型应用》.孙广磊
2. 廖雪峰的官方网站 <https://www.liaoxuefeng.com/>
3. 《基于R语言》

# 基础知识

## 线性代数基本知识

线性代数作为数学的一个分支，广泛用于科学和工程中。同样地，线性代数对于理解和从事机器学习算法相关工作也是很有必要的，比如说，在机器学习某种事件特征我们常常会描绘成特征向量，那么线性代数所涉及到的矩阵的运算和理论方法都必将会应用进来。因此，在我们开始介绍机器学习之前，我们集中探讨一些必备的线性代数知识，如果你已经很熟悉线性代数，那么你可以轻松地跳过本章。

### 向量与矩阵相关的基础知识

1、向量

把数排成一列就是向量。比如和和等，当需要强调分量个数的时候，我们分别称上面的例子为2维向量、3维向量和4维向量。我们在理解2维向量的时候可以用二维坐标去理解，3维向量我们可以用3维空间中的一点来表示。同样的，我们可以用有向线段去理解向量，其中线段的起点就是原点O，终点是该向量对应的位置，与此同时，我们就可以通过图形来理解向量的加法和数量乘法。我们可以把向量看作空间中的点，每个元素是不同的坐标轴上的点。

2、向量空间

我们可以把向量看作空间中的点，每个元素是不同的坐标轴上的点。对于这样一个附加了加法与数量乘法运算的世界，我们就称之为向量空间，也有人称之为线性空间。

3、矩阵

把数排列成长方形就是矩阵。比如或或等。需要强调矩阵规模（大小，高和宽，行数和列数）的时候，我们分别称以上例子为2×2矩阵、2×3矩阵和5×3矩阵。这个时候当行数和列数相同的时候，我们称之为方阵。我们通常会赋予矩阵粗体的大写变量名称，比如 A。矩阵A中第i行第j列的值，称之为A的（i，j）元素。为了简单起见，2×3矩阵可以写成A。

这里我们可以把矩阵理解为映射，例如：将n维向量x乘以m×n矩阵A，能得到m维向量y=Ax。也就是说，指定了矩阵A，就确定了从向量到另外一个向量的映射。这里是从n维空间到m维空间的移动。

4、逆矩阵

对于方阵A，它的逆映射对应的矩阵称之为A的逆矩阵，记为A。任意的向量x，若有Ax=y,则有Ay=x成立。反之，对于任意的向量y，若有Ay=x,则有Ax=y。更为直接一点说，如果有这样的一个映射，对于已知的目标点y,总能还原到出发点x,那么这个映射所对应的矩阵就是A。

5、矩阵的转置

对于矩阵A，将其行列互换得到的矩阵，就称为A的转置矩阵，记为A。例如：A= → A=其中，第一行变成了第一列，第二行变成了第二列。

1. 行列式

行列式，记作 det(A)，是一个将方阵 A 映射到实数的函数。行列式等于矩阵特征值的乘积。行列式的绝对值可以用来衡量矩阵相乘后空间扩大或者缩小了多少。如果行列式是 0, 那么空间至少沿着某一维完全收缩了，使其失去了所有的体积。如果行列式是 1, 那么矩阵相乘没有改变空间体积。

7、矩阵的运算规则

矩阵乘法是矩阵运算中最重要的操作之一。两个矩阵 A 和 B 的矩阵乘积 是第三个矩阵 C。为了使乘法定义良好，矩阵 A 的列数必须和矩阵 B 的行数相等。如果矩阵 A 的形状是 m×n，矩阵 B 的形状是 n×p，那么矩阵 C 的形状是 m × p。矩阵乘积可以作用于两个或多个并在一起的矩阵。例如：

C = AB.

矩阵乘积运算有许多有用的性质，从而使矩阵的数学分析更加方便。比如，矩阵乘积服从分配律：A(B + C) = AB + AC.

矩阵乘积也服从结合律: A(BC) = (AB)C

但是不同于标量乘积，矩阵乘积并不满足交换律（AB = BA 的情况并非总是满足）。然而，两个向量的点积 (dot product) 满足交换律：

x ⊤ y = y ⊤ x

矩阵乘积的转置有着简单的形式：

(AB) ⊤ = B ⊤ A ⊤

由于本书的重点不是线性代数，我们并不试图展示矩阵乘积的所有重要性质，但读者应该知道矩阵乘积还有很多有用的性质。

8、线性相关、秩以及特征值、特征向量的知识

（1）线性相关

如果一组向量中的任意一个向量都不能表示成其他向量的线性组合，那么这组向量被称为线性无关。如果某个向量是一组向量中某些向量的线性组合，那么我们将这个向量加入到这组向量后不会增加这组向量的所构成的空间。用数学语言来说的话，如果对于数u,...u，当成立时有u=...=u=0，则称a...a为线性无关的。

（2）秩有关的概念

令A是m×n矩阵，也就是说，我们考虑把n维向量x变成m维向量y=Ax的这一映射。这里我们把A维数命名为矩阵A 的秩。也就是非零行的行数就是矩阵的秩。记为rankA。当rankA=n（秩与原空间（定义域）的维数相等）=A是单射；rankA=m（秩与目标空间（值域）的维数相等）=A是满射。

以上论述从直观上看也是显然的，如果n维原空间在映射之后的目标空间中同样保持着n维，则一定没有遇到“压缩”的情况，另外，在映射到目标空间时，如果有着和整个目标空间相同的m维，那么一定也就覆盖到了空间全体。

（3）特征值与特征向量

一般而言，对于方阵A，满足：

Ap=p （p≠0）

的数和向量P分为矩阵的特征值和特征向量。

从几何意义上来讲，特征向量乘上A之后，除了长度会有伸缩变化，方向不发生变化，这里的长度变化倍率，便是特征值。

[https://mp.weixin.qq.com/s?\_\_biz=MzI4MDYzNzg4Mw==&mid=2247485066&idx=1&sn=7ee345dcd76e4d90d884eda7a09f148f&chksm=ebb43e5edcc3b748d1d1cc0ca29787aae4406f4682ff3b1ef245d78bd7c6b83c0ccda806b677&scene=21#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzI4MDYzNzg4Mw==&mid=2247485066&idx=1&sn=7ee345dcd76e4d90d884eda7a09f148f&chksm=ebb43e5edcc3b748d1d1cc0ca29787aae4406f4682ff3b1ef245d78bd7c6b83c0ccda806b677&scene=21" \l "wechat_redirect)

### n维向量和向量空间

1、n维向量

n个数a，a，...，a构成的有序数组称为n维向量，记作（a，a，...，a)或（a，a，...，a）⊤，分别称为n维行向量或n维列向量，也就是1×n或n×1矩阵，数a称为向量的第i个分量。

2、向量空间

设v是n维向量的非空集合，且v中向量对于加法及数乘这两种运算封闭，则称v是向量空间。在这样的一个世界里，更加强调方向性，一般我们表示向量都用方向箭头来表示。其实，向量空间是我们生活的现实空间的一个缩影，也是对现实空间的抽象化。

如果W是向量空间V的一个非空子集，如W对V的加法及数乘两运算都封闭，则称W是V的子空间。

3、基及维数

设V是向量空间，若V中有r个向量线性无关，且V中任一向量都可由这r个向量来线性表出，则称这r个向量为向量空间V的一组基，r为向量空间V的维数，并称V为r维向量空间。

4、极大线性无关组

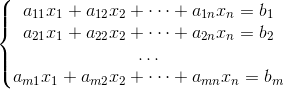
在向量组a，a，...，a中，如存在一个部分组a,a,...,a线性无关，且再添加进组中任一个向量a,向量组a,a,...,a，a一定线性相关，则称a,a,...,a是向量组a，a，...，a的一个极大线性无关组。

向量组a，a，...，a的极大线性无关组所含向量的个数r称为该向量组的秩，记为r(a，a，...，a)=r。

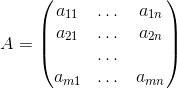
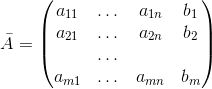
[https://mp.weixin.qq.com/s?\_\_biz=MzI4MDYzNzg4Mw==&mid=2247485175&idx=1&sn=b18c1df20be69abd403ee97f1e7ff5b3&chksm=ebb43e23dcc3b735c8c8029bdb5d6a22bb0745f8a5a3d7dc025c84457ac2c75ae21047e405f3&scene=21#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzI4MDYzNzg4Mw==&mid=2247485175&idx=1&sn=b18c1df20be69abd403ee97f1e7ff5b3&chksm=ebb43e23dcc3b735c8c8029bdb5d6a22bb0745f8a5a3d7dc025c84457ac2c75ae21047e405f3&scene=21" \l "wechat_redirect)

### 2.1.3图像解方程组

### 在第一节中，我们已经介绍了线性代数有关的基础知识，了解了向量，也知道了空间的概念，培养了把矩阵看成是映射的思想。那么，从本节开始，我们一起来学习线性代数的具体知识。首先我们从解方程谈起，学习线性代数的应用之一就是求解复杂方程问题，本节核心之一也是从空间，向量的角度来求解方程。

以下是线性代数方程组的一般形式：  
为m个方程n个未知数的方程组。

其中：

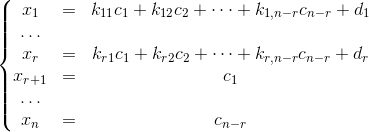
称称为系数矩阵，称为增广矩阵。

* 方程组与它的增广矩阵互相唯一决定。

判断方程组是否有解的情况：

1. 当方程个数等于未知数个数，则若≠0,方程组有唯一解；
2. 若=0且r（A）≠r(****),方程组有无解
3. 若=0且r（A）=r()，方程组由无穷多个解

接下来讨论方程组通解的情况：

1. 通解：设方程组有无穷组解，则有个自由未知量，令得  
   其中称为方程组的通解。  
   通解也可写成向量式。

2.特解：通解中的某个具体的解。

3.齐次线性方程组（右边常数项全为0）

设A是m×n矩阵，齐次方程组Ax=0有非零解的充要条件是r(A)<n，亦是A的列向量线性无关。

特别地，=0也是齐次线性方程组有非零解的充要条件。

当方程个数<未知数的个数同样为其有非零解的充要条件。

角度一：从行图像与列图像的角度解方程

1） 二维的行图像

我们首先通过一个例子来从行图像角度求解方程

〖例1〗求解方程 2x-y=0

-x+2y=3

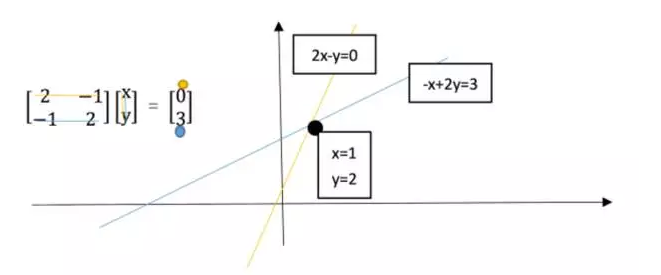
我们首先把方程写成矩阵的形式：

系数矩阵 未知向量 向量

接下来我们通过行图像来求解这个方程：

所谓行图像，就是在系数矩阵上，一次取一行构成方程，在坐标系上作图。和我们在初等数学中学习的作图求解方程的过程无异。



2）二维的列图像

从列图像的角度我们来解决这个方程组 2x-y=0

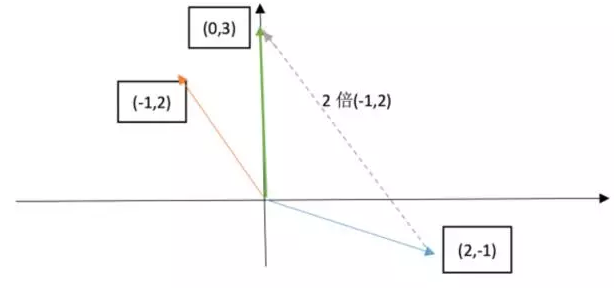
-x+2y=3

这一次我们来求解这个方程组，我们把从列提取，使用的矩阵为：

x+y=

如图，我们使用列向量构成系数矩阵，将问题化为向量和向量的线性组合，使其组合的结果为。

接下来我们使用列图像求解此方程：



即寻找合适的 x，y 使得 x 倍的(2,-1) + y 倍的(-1,2)得到最终的向量(0,3)。很明显能看出来，1 倍(2,-1) + 2 倍(-1,2)即满足条件。反映在图像上，明显结果正确。

我们再想一想，仅仅针对这个x+y=方程，我们任意取x和Y我们能的得到的是什么呢，可想而知，我们得到的是铺满整个平面的任意方向的向量。

3）高维行图像

我们把方程维数进行推广，从三维开始，

2x-y=0

-x+2y-z=-1,如果我们继续用行图像去解方程，我们会得到一个非常

-3y+4z=4

复杂的图像，如果绘制行图像，很明显这是一个三个平面相交得到一点，我们想直接看出这个点的性质可谓是难上加难。

比较靠谱的思路是先联立其中两个平面，使其相 交于一条直线，在研究这条直线与平面相交于哪个点，最后得到点坐标即为方程的解。这个求解过程对于三维来说或许还算合理，那四维呢？五维甚至更高维数呢？直观上很难直接绘制更高维数的图像，这种行图像受到的限制也越来越多。所以我们考虑用列图像来理解。

4）高维列图像

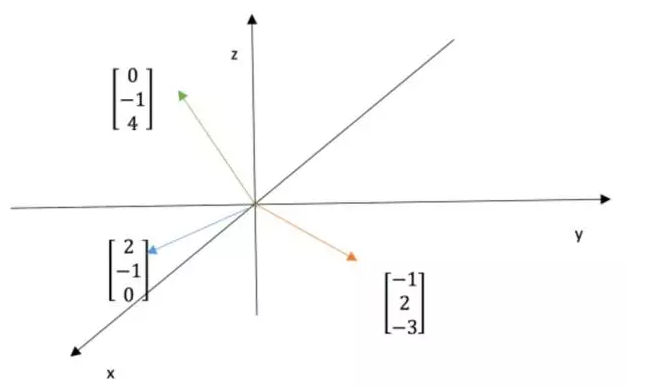
还用那个方程组 2x-y=0

-x+2y-z=-1，我们采用列图像的思路去求解这个方

-3y+4z=4

程组。上面的方程组可以化为：x+y+z=

左侧是线性组合，右侧是合适的线性组合组成的结果，这样一来思路就清晰多了，“寻找线性组合”成为了解题关键。



很明显这道题是一个特例，我们只需要取 x = 0，y = 0，z = 1。就得到了结果，这在行图像之中并不明显。

当然，之所以我们更推荐使用列图像求解方程， 是因为这是一种更系统的求解方法，即寻找线性组合，而不用绘制每个行方程的图像之后寻找那个很难看出来的点。

另外一个优势在于，如果我们改变最后的结果 b，例如本题中，

x+y+z=

那么我们 2 −1 1 0 −3 4 −3 就重新寻找一个线性组合就够了，但是如果我们使用的是行图像呢？那意味着我 们要完全重画三个平面图像，就简便性来讲，两种方法高下立判。

[https://mp.weixin.qq.com/s?\_\_biz=MzI4MDYzNzg4Mw==&mid=2247485066&idx=1&sn=7ee345dcd76e4d90d884eda7a09f148f&chksm=ebb43e5edcc3b748d1d1cc0ca29787aae4406f4682ff3b1ef245d78bd7c6b83c0ccda806b677&scene=21#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzI4MDYzNzg4Mw==&mid=2247485066&idx=1&sn=7ee345dcd76e4d90d884eda7a09f148f&chksm=ebb43e5edcc3b748d1d1cc0ca29787aae4406f4682ff3b1ef245d78bd7c6b83c0ccda806b677&scene=21" \l "wechat_redirect)

http://www.cnblogs.com/CQBZOIer-zyy/p/5722418.html

### 2.1.4 消元法解方程组

1、消元法

对于一些“好”的系数矩阵（可逆矩阵）A 来说，我们可以使用消元法来求解方程 Ax = b，我们还是从一个例子谈起。

求解方程 x+2y+z=2

3x+8y+z=12 还是使用矩阵来计算，写成矩阵运算的形式：

4y+z=2

=

所谓矩阵的消元法，与我们初等数学中学习的解二元一次方程组的消元法其实师出同门，都是通过将不同行的方程进行消元运算来简化方程，最后能得到简化的方程组，只不过这里我们把系数单独抽出来进行运算，寻找一种矩阵情况下的普遍规律而已。

消元针对的对象是系数矩阵A**：**，首先注意，左上角的1是消元法关键，我们称之为主元1，接下来就是我们熟悉的“将一行乘倍数加到另一行”的行化简的方法将第一列中除了主元之外的元素全变为0.

第一步目标达成，我们在第一列中只留下了主元1，很好，接下来我们可以认为第一行与第一列已经“完工”了，再看去掉第一行第一列之后右下角剩下的部分：，同样，我们将左上角的2视为主元，消元第一列，使其列上（不包括第一行中元素），除此主元2之外皆为0.

这个时候第三行就只剩下5，我们直接将其处理为主元即可。由于A矩阵可逆，经过消元处理得到的上三角矩阵U=中有三个主元，至此，消元结束，得到的U即为我们想要化简的形式。

注意：并不是所有的A矩阵都可以消元处理，需要注意在我们消元过程中，如果主元位置（左上角）为0，那么意味着这个主元不可取，需要进行“换行”处理：首先看它的下一行对应位置是不是0，如果不是，就将这两行位置互换，将非零数视为主元，如果是，就看下一行，以此类推，若其下面一行都看到了，仍然没有非零数的话，那就意味着这个矩阵不可逆，消元法求出的解不唯一。

2、回带求解

其实回带求解应该和消元法同时进行，只不过本课中以及一些软件工作原理中它们是先后进行的，所以我们这里分开讨论，先介绍增广矩阵：

上文的例子中，矩阵形式=其增广矩阵的形式为：一下子就看出来了，就是把系数矩阵 A 和向量 b 拼接成一个矩阵就行了。然后像我们之前说的那样消元，但是这次要带着增广的b的那一列一起进行：

   带回方程Ax=b,变为： x+2y+z=2

x+2y+z=2

2y-2z=6 从下向上开始求解，很容易求出x.y.z的值了。

5z=-10

[https://mp.weixin.qq.com/s?\_\_biz=MzI4MDYzNzg4Mw==&mid=2247485175&idx=1&sn=b18c1df20be69abd403ee97f1e7ff5b3&chksm=ebb43e23dcc3b735c8c8029bdb5d6a22bb0745f8a5a3d7dc025c84457ac2c75ae21047e405f3&scene=21#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzI4MDYzNzg4Mw==&mid=2247485175&idx=1&sn=b18c1df20be69abd403ee97f1e7ff5b3&chksm=ebb43e23dcc3b735c8c8029bdb5d6a22bb0745f8a5a3d7dc025c84457ac2c75ae21047e405f3&scene=21" \l "wechat_redirect)

### 2.1.5 矩阵乘法和逆矩阵

第一节我们已经介绍过了向量与矩阵之间的乘法，这节我们来介绍一下矩阵之间的乘法运算以及引进求逆的方法。

1、矩阵乘法最常见的求解方式

首先来了解矩阵之间进行乘法运算时，我们是如何求解单个元素的呢？

？ ？ = ？

A \* B = C

现在我们取C中的一个元素，就比如C中的一个元素C：

这个下标的意义就是C。它的计算也和其密切相关。即A中C对应行向量与B中C对应列向量的数量积的值即为C的值。

如图:

=

A B C

C=

推广：C=(A中第i行向量）（B中第j列向量）=

另外我们这里还要注意一下矩阵规格的问题，因为我们矩阵乘法运算是使用前一个矩阵的行向量点乘后一个矩阵的列向量，所以从向量角度看，它们必须有相同的分量。故，A的列数必须与B的行数相同，结果C矩阵的规格是A的行数，B的列数。

2、列组合与行组合方式

1）列组合的角度：还记得我们之前学习过矩阵与列向量的乘积，得到一个列向量：

 =A A A A=C

A B

B的各列，用来线性组合A的列向量

之后，每一个类似于A这样的A矩阵与向量乘积，都可以转化为矩阵A的列向量线性组合。最后将得到的列向量组合在中，即为结果C。这种方法的关键就是将右侧矩阵B看做列向量组合，将问题转化为矩阵与向量的乘法问题，也表明了矩阵C就是矩阵A中各列向量的线性组合，而B其实在告诉我们，要以什么样的方式组合A中的列向量。

2）行组合的角度：同理，我们还学习过行向量与矩阵的乘法，按照形式，这次将矩阵A看做行向量组合就行了：

==C

之后每一个类似于B这样的行向量与矩阵之积，会得到一个行向量，各行向量最后构成了C中的各行。这里注意，C中各行都来自于B形式求得的行向量，所以是矩阵B各行的线性组合组成了C的各行。

3、分块做乘法

分块乘法就是宏观上的矩阵乘法，比如现在有一个 50\*50 的矩阵与 50\*50 矩阵相乘，一个一个进行运算很麻烦，尤其是如果矩阵在某一区域上有一定的性质， 那么我们可以将其分块，如：

=

其中的A与B都是划分之后的一块块矩阵，如：得到的各个部分，而C=AB+AB,和矩阵乘法的计算步骤一样，只是这里的AB，AB都是矩阵之间的乘法而已。只要A与B分块相互匹配，就可以用这样的分块乘法求解。

4、逆矩阵求解

之前我们在基础知识的部分已经介绍过了有关的逆矩阵的知识。对于一个方阵A,如果A可逆，就有这样一个A使 AA=E=AA。

如：，如果通过行列式的角度，这个矩阵对应的行列式为0，所以矩阵不可逆，但是我们这里不从行列式的角度，我们换一个角度。我们看到这个矩阵中含有两个列向量和这两个向量，它们之间互成倍数，也就是说这两个向量之一对其线性组合无意义，那么这个A不可能有逆矩阵，换句话说：

若存在非零向量x，使得Ax=0,那么A就不可能有逆矩阵

这个是为什么呢？其实，原因很简单，如果A有逆矩阵，在Ax=0这个等式两端同时乘以A，就有 AAx=E=零向量，但是E不可能是零向量，自相矛盾，所以A没有逆矩阵。

再看看矩阵，由于两个列向量线性相关，一定有一个x，使得Ax=0。

下面将介绍求解逆矩阵的方法，其实求逆矩阵就是解方程组的过程。

例：求解的逆矩阵。

=E=

A A

从列向量的角度看来，得到两个方程：

= =直接解这个方程组就好了。

但是这个对于低阶矩阵还好，对于高阶矩阵计算量未免太大了，所以这里介绍高斯—若尔当 方法。还是用上面的例子，两个方程；

= =

这个方法就是可以同时处理两个方程组，即使用增广矩阵联系两个方程增广矩阵：



接下来进行行变换，将直线左侧消为单位矩阵E，此时右侧矩阵就是逆矩阵。

逆矩阵为： 

检验一下：AA=E 正确

所以说，高斯—若尔当 方法在求逆矩阵的过程是非常普遍使用的。

接下我们来论证它的合理性：

我们知道，上面这个过程，对进行消元处理使其变为单位矩阵E,就相当于左乘一大推消元矩阵，假设为P，就有P=E,那么P肯定就是逆矩阵A，再看直线右侧，单位矩阵经历了与同样的消元过程，最后相当于结果就是PE=A,那么直线右侧就是其逆矩阵就对了。

[https://mp.weixin.qq.com/s?\_\_biz=MzI4MDYzNzg4Mw==&mid=2247485310&idx=1&sn=72a7b0d6142c3b8cad385ff9ea239f01&chksm=ebb43faadcc3b6bce80291ab19dfd6342842a2ec658a48197d93c18b288c3033af8da0f0aba5&scene=21#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzI4MDYzNzg4Mw==&mid=2247485310&idx=1&sn=72a7b0d6142c3b8cad385ff9ea239f01&chksm=ebb43faadcc3b6bce80291ab19dfd6342842a2ec658a48197d93c18b288c3033af8da0f0aba5&scene=21" \l "wechat_redirect)

## 2.1.6 特征分解

许多数学对象可以通过将它们分解成多个组成部分，或者找到它们的一些属性而更好地理解，这些属性是通用的，而不是由我们选择表示它们的方式引起的。

例如，整数可以分解为质数。我们可以用十进制或二进制等不同方式表示整数12，但质因数分解永远是对的 12 = 2 × 3 × 3。从这个表示中我们可以获得一些有用的信息，比如 12 不能被 5 整除，或者 12 的倍数可以被 3 整除。

正如我们可以通过分解质因数来发现整数的一些内在性质，我们也可以通过分解矩阵来发现矩阵表示成数组元素时不明显的函数性质。特征分解是使用最广的矩阵分解之一，即我们将矩阵分解成一组特征向量和特征值。

方阵 A 的特征向量是指与 A 相乘后相当于对该向量进行放缩的非零向量 P：

AP = P

数值 被称为这个特征向量对应的特征值。（类似地，我们也可以定义左奇异向量p⊤A = p⊤ ，但是通常我们更关注右奇异向量。

如果p是A的特征向量，那么任何放缩后的向量 sp（≠0）也是A的特征向量。此外，sp和 p有相同的特征值。基于这个原因，通常我们只考虑单位特征向量。

假设矩阵 A 有 n 个线性无关的特征向量 {v (1) ,...,v (n) }，对应着特征值

{λ ,...,λ}。我们将特征向量连接一个矩阵，使得每一列是一个特征向量：V =

[v (1) ,...,v (n) ]. 类似地，我们也可以将特征值连接成一个向量 λ = [λ 1 ,...,λ n ] ⊤ 。因此 A 的特征分解可以记作

A = Vdiag(λ)V

我们已经看到了构建具有特定特征值和特征向量的矩阵，能够使我们在目标方向上延伸空间。然而，我们也常常希望将矩阵分解 (decompose) 成特征值和特征向量。这样可以帮助我们分析矩阵的特定性质，就像质因数分解有有助于我们理解整数。

不是每一个矩阵都可以分解成特征值和特征向量。在某些情况下，特征分解会涉及到复数，而非实数。幸运的是，在本书中我们通常只需要探讨一类有简单分解的矩阵。具体地，每个实对称矩阵都可以分解成实特征向量和实特征值：

A = QΛQ⊤

其中 Q 是 A 的特征向量组成的正交矩阵，Λ 是对角矩阵。特征值 Λi,i 对应的特征向量是矩阵 Q 的第 i 列，记作 Q i 。因为 Q 是正交矩阵，我们可以将 A 看作是沿方向 v (i) 延展 λi 倍的空间。

虽然任意一个实对称矩阵 A 都有特征分解，但是特征分解可能并不唯一。如果两个或多个特征向量拥有相同的特征值，那么这组特征向量生成子空间中，任意一组正交向量都是该特征值对应的特征向量。因此，我们可以使用任意一组正交向量作为 Q 的特征向量。按照惯例，我们通常按降序排列 Λ 的条目。在该约定下，特征分解唯一当且仅当所有的特征值都是唯一的。

矩阵的特征分解给了我们很多关于矩阵的有用信息。矩阵是奇异的当且仅当含有零特征值。实对称矩阵的特征分解也可以用于优化二次方程 f(x) = x⊤Ax，其中限制 =1。当 x 等于 A 的某个特征向量时，f 将返回对应的特征值。在限制条件下，函数 f 的最大值是最大特征值，最小值是最小特征值。

所有特征值都是正数的矩阵被称为正定；所有特征值都是非负数的矩阵被称为半正定。同样地，所有特征值都是负数的矩阵被称为负定 ；所有特征值都是非正数的矩阵被称为半负定。半正定矩阵受到关注是因为它们保证 ∀x,x⊤Ax ≥ 0。此外，正定矩阵还保证 x ⊤Ax = 0 ⇒ x = 0。

https://github.com/exacity/deeplearningbook-chinese

## 概率论基本知识

概率论是用于表示不确定性陈述的数学框架。它不仅提供了量化不确定性的方法，也提供了用于导出新的不确定性陈述的公理。在人工智能领域，我们主要以两种方式来使用概率论。首先，概率法则告诉我们AI系统应该如何推理，所以我们设计一些算法来计算或者近似由概率论导出的表达式。其次，我们可以用概率和统计从理论上分析我们提出的AI系统的行为。

从信息论到机器学习，从模式识别到数据挖掘，概率与统计的概念和原理活跃于计算机科学的各个领域，发挥着重要的作用。它是众多科学和工程学科的基本工具。

为什么在机器学习中会用到概率呢？这是因为，机器学习必须始终处理不确定量，有时也可能需要处理随机（非确定性）量。不确定性和随机性可能来自多个方面。研究人员至少从20世纪80年代开始就对使用概率论来量化不确定性提出了令人信服的论据。几乎所有的活动都需要能够在不确定性存在时进行推理。概率可以被看作是用于处理不确定性的逻辑扩展。

### 概率、条件概率

概率是对随机事件发生的可能性的度量，一般以一个在0到1之间的实数表示一个事件发生的可能性的大小。越接近1，该时间更可能发生；越接近0，则该事件更不可能发生。这是一个客观论证，而不是主观验证。

随机变量，通俗来讲，它是一种会随机改变的不确定量。就其本身而言，一个随机变量只是对可能的状态的描述；它必须伴随着一个概率分布来指定每个状态的可能性。

随机变量可以是离散的或者连续的。离散型随机变量拥有有限或者可数无限多的状态。注意，这些状态不一定非要是整数，它们也可能只是一些被命名的状态并且没有数值。连续型随机变量都伴随着实数值。

### 概率分布

概率分布也是概率论里的一个基本概念。概率分布是用来描述随机变量或一簇随机变量在每一个可能取到的状态的可能性大小。我们描述在描述概率分布时，首先要确定这个随机变量是离散的还是连续的。

比如掷骰子或者投硬币之类的就是离散型问题。离散型变量的概率分布可以用概率分布律函数来描述。我们通常用P来表示概率分布律函数。通常，每一个随机变量都会有一个不同的概率分布律函数，并且读者必须根据随机变量来推断所使用的概率分布律函数，而不是根据函数的名称来推断，例如：P(x)通常和P(y)不一样。

概率分布律函数将随机变量能够取得的每个状态映射到随机变量该状态的概率。X=x的概率用P(x)来表示，概率为1表示X=x是确定的，概率为0表示X=x是不可能发生的。我们把变量x遵循的分布表示为：X~P(x)。

一个函数P如果想要成为随机变量x的概率分布函数，那么它必须满足下面这几个条件：

* P的定义域必须是x所有可能状态的集合。
* 不可能发生事件的概率为0，并且没有比这概率更低的状态了。类似的，能够确保一定发生事件的概率为1，而且没有比这概率更高的状态了。
* .我们把这条性质称之为归一性。如果没有这条性质，那么当我们计算很多事件的其中之一发生的概率时可能会得到大于1的概率。

当我们研究的对象是连续型随机变量时，我们用概率密度函数而不是概率分布律函数来描述它的概率分布。一个函数p如果想要成为概率密度函数，必须满足下面几个条件：

* P的定义域必须是x所有可能状态的集合。
* .注意，我们并不要求P(x)。
* 。

概率密度函数p(x)并没有直接对特定的状态给出概率，相对的，它给出了落在一定面积无限小的区域内的概率。我们可以对概率密度函数求积分来获得点集的真实分布律。特别地，x落在集合S中的概率可以通过p(x)对这个集合求积分来得到。在单变量的例子中，x落在区间[a,b]的概率是。

接下来我们来认识一下累积分布函数（或者简称为分布函数），它的数学定义如下，,其中X是一个实数值随机变量。我们可以通过累积分布函数与概率密度函数的图像来表示实数值的概率分布。如果条件允许，我们还可以用数式替代图像来表示分布。它们的形式由具体的分布决定。而且，在很多实际应用中，概率密度函数比累积分布函数应用范围更广。

### 2.2.3联合概率、边缘概率和条件概率

对于现代的概率统计来说，分析多个随机变量之间的相互关系是一个关键。“在星期五购买一次性纸尿布的顾客很可能也会买啤酒”这类的观点是不是看起来很熟悉？为了讨论这类问题，我们必须分析多个随机变量之间的相互关系。我们的目标首先是要理解联合概率、边缘概率与条件概率这三个概念。它们是讨论随机变量之间关系的基本道具。最近活跃于各领域的贝叶斯公式也是这组概念的一种应用。此外，独立性的定义也基于这组概念。在日常的概率问题中，我们在使用独立性时无需特地声明，不过由于本书的要求更高，因此必须明确定义独立性的含义。在分析随机变量之间的关系时，某一变量是否独立往往是解决问题的关键。

假设有随机变量X与Y，此时P(X=a,Y=b)用于表示X=a且Y=b的概率。这类包含多个条件且所有条件同时成立的概率称为联合概率。要注意，联合概率并不是其中某个条件的成立概率，而是所有条件同时成立的概率。与之对应地，P(X=a)或P(Y=b)这类仅与单个随机变量有关的概率称为边缘概率。联合概率的一览表称为联合分布，边缘概率的一览表称为边缘分布。

联合概率与边缘概率的关系如下。





接下来，我们来引入条件概率这个概念。在实际生活中，许多有价值的变量都能以条件概率这一概念来描述。举个例子，“包含‘免费’这一单词的邮件很可能是广告”，这种A条件下事件B的概率称为条件概率。下面是条件概率的定义。



我们从面积的角度看，可以将理解为“区域a中b的比例”，代表的是“区域a中b的面积”，而P(X=a)代表的是“区域a的总面积”，那么就是“区域a中b的面积除以区域a的总面积”了。

条件概率经常是理工科一些问题的焦点。这是因为在研究理工科问题时，我们常会采用控制变量法分析变量之间的关系，讨论变量X取特定值时变量Y的取值情况。如果没有误差，我们可以用函数Y=f(X)来表示它们的关系。不过在现实中，我们很难去除所有影响因素，测到绝对精确的值。于是，即使X的测定值不变，Y的测定值也会发生细微变化。此时，我们就需要研究在X为某个特定值时Y的概率分布。也就是说，我们将研究条件概率，而不是函数Y=f(X)。

也是正因如此，要在实际问题中运用概率理论，我们必须掌握条件概率的计算方法。接下来我们来熟悉一下条件概率中常见的基本关系式。

我们可以像下面这样，通过边缘分布和条件分布来表示联合分布。



我们也可以想了解条件概率的定义一样从面积的角度来理解这个式子，这里就不一一赘述。

现在，我们将应用条件概率来解决逆问题。简单地说，逆问题是指那些需要从结果反推原因的问题。通常，原因X无法被直接观察、测量，此时，我们常会通过其结果Y来反推X。比如，在文字识别问题时，需要根据扫描仪读取的图像数据Y来推测用户书写的文字X。请注意，X相同，Y也可能会不同。由于绝大多数情况都存在噪声与误差，因此我们不能简单地使用函数Y=f(X)来模拟问题。为此，我们需要借助概率来处理这些噪声与误差，通过随机变量X,Y来描述这两者之间的相互关系。即在已知时来计算。由此，如果我们还知道P(X=a)，就可以用贝叶斯公式来实现这一目的：



贝叶斯公式可以从条件概率的定义直接推导出来，但我们最好记住这个公式以及它的名字，因为在之后的算法中会经常用到。

### 2.2.4独立性

现在，我们来讨论多个随机变量之间关系中更为根本的问题。如果一个问题当中存在多个随机变量，我们首先会怀疑这些随机变量之间是否真的存在关联。因此，独立性的概念是很多应用问题中的关键。

独立的定义有多种等价表述方式。在讨论概率问题是，独立性是一种很重要的基础概念，因此大家需要熟悉各种不同的表达方式。请大家记住，以下表述的含义全都是相同的。

* X与Y独立
* 条件概率与条件无关：
* 添加或去除条件不影响概率：
* 联合概率是边缘概率的乘积：

### 2.2.5期望、方差和协方差

函数f(x)关于某分布P(X)的期望（expectation）或者期望值是指，当x由P产生时，f作用于x的平均值。对于离散型随机变量，可以通过求和得到：



对于连续型随机变量可以通过求积分得到：



方差（variance）衡量的是当我们对x依据它的概率分布进行采样时，随机变量x的函数值会呈现多大的差异：



当方差很小时，f(x)的值形成的簇比较接近它们的期望值。方差的平方根被称为标准差。

协方差在某种意义上给出了两个变量线性相关性的强度以及这些变量的尺度：



协方差的绝对值如果很大则意味着变量值变化很大并且他们同时距离各自的均值很远。如果协方差是正的，那么两个变量都倾向于同时取得相对较大的值。如果协方差是负的，那么其中一个变量倾向于取得相对较大的值的同时，另一个变量倾向于取得相对较小的值，反之亦然。其他的衡量指标如相关系数将每个变量的贡献归一化，为了只衡量变量的相关性，而不受变量大小的分别影响。

协方差和相关性是有联系的，但实际上是不同的概念。它们是有联系的，因为两个变量如果相互独立，那么它们的协方差为零，如果两个变量的协方差不为零，那么它们一定是相关的。然而，独立性又是和协方差完全不同的性质。两个变量如果协方差为零，它们之间一定没有线性关系。独立性的要求比协方差为零更高，因为独立性在此之上还排除了非线性的关系。两个变量可能相互依赖，但同时它们协方差为零。

### 2.2.6常用概率分布

伯努利分布，又名两点分布或0-1分布。这是一种最简单的离散型分布。假设随机变量X为投一枚硬币的结果，类似于“投硬币”的试验就是伯努利试验。如果试验E是一个伯努利试验，将E独立重复地进行n次，则称这一串重复的独立试验为n重伯努利分布。进行一次伯努利试验，成功(X=1)概率为p()，失败(X=0)概率为1-p，则称随机变量X服从伯努利分布。它的概率分布函数为：



1. 分布是n重伯努利试验成功次数的离散概率分布。如果试验E是一个n重伯努利试验，每次伯努利试验成功概率为p,X代表成功的次数，则X的概率分布是二项分布，记为，其概率质量函数为：



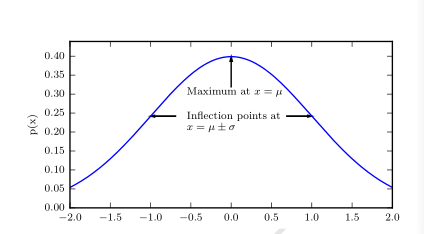


可以发现，伯努利分布其实就是二项分布在n=1时的特例。

现在，我们要来介绍一下正态分布，正态分布是对于实数上的分布最常用的分布，也称为高斯分布。



正态分布被两个参数控制，。参数给出了中心峰值的坐标，这也是分布的均值：E(X)=。分布的标准差用表示，方差用表示。下面是标准正态分布的概率密度函数，其中。



采用正态分布在很多应用中都是一个明智的选择。当我们缺乏对于某个实数分布的先验认识而不知道该选择怎样的形式时，正态分布是默认的比较好的选择，这里有两个原因。

第一，我们想要建模的很多分布真实情况都是比较接近正态分布的。中心极限定理告诉我们，很多独立随机变量的和近似服从正态分布。这意味着，在实际应用中，很多复杂系统都可以被成功的建模成正态分布的噪声，即使系统可以被分解成具有更多结构化行为的各个部分。

第二，在具有相同方差的所有可能的概率分布中，正态分布在实数上具有最大的不确定性。我们因此可以认为正态分布是对模型加入的先验知识量最少的分布。

参考资料：

《深度学习》

《程序员的数学2——概率统计》

数理统计中常用函数、概率分布函数总结 - CSDN博客 http://blog.csdn.net/DreamHome\_S/article/details/78275268

## 机器学习基本知识

### 何为机器学习

* 什么是学习

孔子云：“学而时习之，不亦说乎。”本意为：学了之后及时、经常地进行温习和实习不是一件很愉快的事吗？而自古以来“学”意为自学或有人教你学，偏重于知识理论的学习；“习”意为巩固知识、技能的行为，偏重于复习行动的实践；“学习”则意为自我概念的变化、价值与潜能的实现。既是概念的变化、又是潜能的实践，所以“学而时习之，不亦说乎”的新解就是：“学习”不仅要学习理论知识，还要经常动手实践，将理论和实践结合在一起不是一件很有成就感的事吗？温菲尔德也说过：“学习是由练习或经验引起的行为或知识的较持久的变化。”可见，“学习”的本质其实就是一个从陌生到熟悉再到掌握的过程、学习理论和实践巩固相结合的过程。

* 什么是机器学习

定义： 对于某类任务 T 和性能度量 P，如果一个计算机程序在 T 上以 P 衡量的性能随着经验 E 而自我完善，那么我们称这个计算机程序在从经验 E。

### 机器学习的主要任务

* 分类
* 回归
* 聚类

### 机器学习的基本步骤

* 数据的采集
* 数据的前处理
* 机器学习算法的处理
* 算法的对比
* 结果评价

# scikit-learn基本知识

## sklearn结构

起源

主要模块

## 加载范例数据库

# 预处理

## 标准化、归一化与正则化

数据的标准化（normalization）是将数据按照一定规则缩放，使之落入一个小的特定区间。这样去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。

前数据标准化方法有多种，归结起来可以分为直线型方法(如极值法、标准差法)、折线型方法(如三折线法)、曲线型方法(如半正态性分布)。不同的标准化方法，对系统的评价结果会产生不同的影响，然而不幸的是，在数据标准化方法的选择上，还没有通用的法则可以遵循。

* + 归一化的目标主要有两个：

1、把数变为（0，1）之间的小数

主要是为了数据处理方便提出来的，把数据映射到0～1范围之内处理，更加便捷快速，应该归到数字信号处理范畴之内。

2、把有量纲表达式变为无量纲表达式

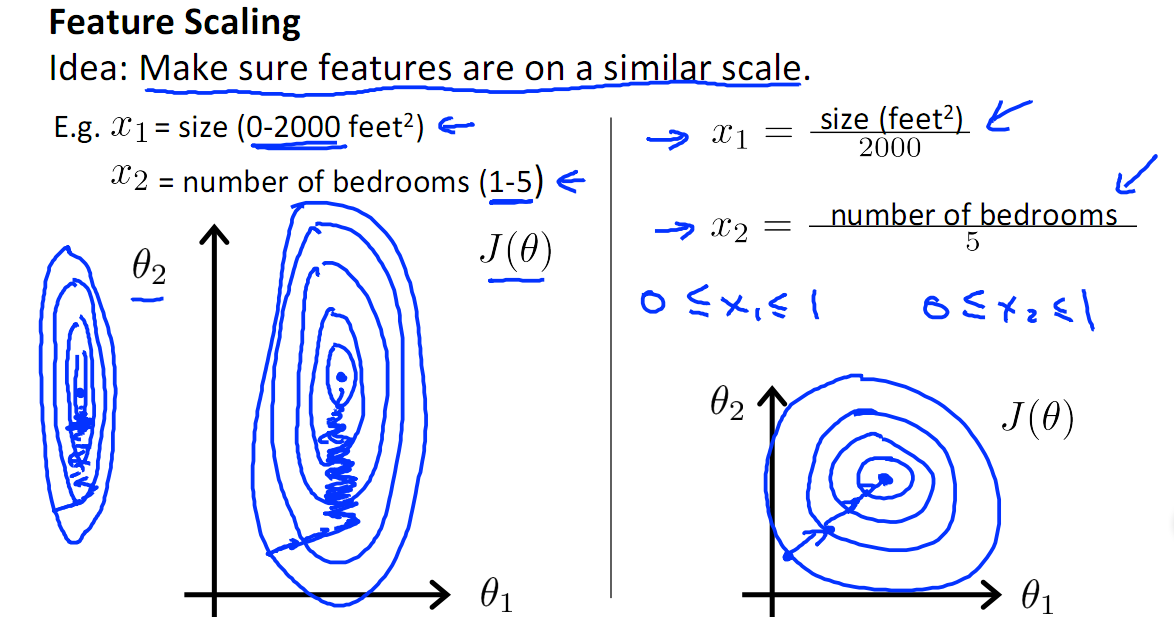
化为无量纲的表达式，成为纯量。 比如，复数阻抗可以归一化书写：Z = R + jωL = R(1 + jωL/R) ，复数部分变成了纯数量了，没有量纲。

另外，微波之中也就是电路分析、信号系统、电磁波传输等，有很多运算都可以如此处理，既保证了运算的便捷，又能凸现出物理量的本质含义。

* + 归一化后有两个好处

1. 提升模型的收敛速度

如下图，x1的取值为0-2000，而x2的取值为1-5，假如只有这两个特征，对其进行优化时，会得到一个窄长的椭圆形，导致在梯度下降时，梯度的方向为垂直等高线的方向而走之字形路线，这样会使迭代很慢，相比之下，右图的迭代就会很快（理解：也就是步长走多走少方向总是对的，不会走偏）

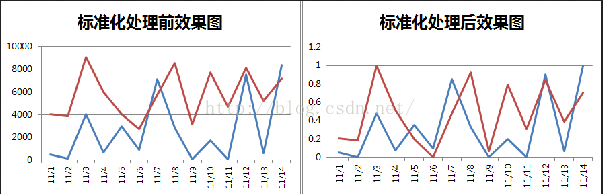


2.提升模型的精度

**P1137 25.2 Importance of Feature Scaling**

归一化的另一好处是提高精度，这在涉及到一些距离计算的算法时效果显著，比如算法要计算欧氏距离，上图中x2的取值范围比较小，涉及到距离计算时其对结果的影响远比x1带来的小，所以这就会造成精度的损失。所以归一化很有必要，他可以让各个特征对结果做出的贡献相同。

在多指标评价体系中，由于各评价指标的性质不同，通常具有不同的量纲和数量级。当各指标间的水平相差很大时，如果直接用原始指标值进行分析，就会突出数值较高的指标在综合分析中的作用，相对削弱数值水平较低指标的作用。因此，为了保证结果的可靠性，需要对原始指标数据进行标准化处理。



在数据分析之前，我们通常需要先将数据标准化，利用标准化后的数据进行数据分析。数据标准化也就是统计数据的指数化。数据标准化处理主要包括数据同趋化处理和无量纲化处理两个方面。数据同趋化处理主要解决不同性质数据问题，对不同性质指标直接加总不能正确反映不同作用力的综合结果，须先考虑改变逆指标数据性质，使所有指标对测评方案的作用力同趋化，再加总才能得出正确结果。数据无量纲化处理主要解决数据的可比性。经过上述标准化处理，原始数据均转换为无量纲化指标测评值，即各指标值都处于同一个数量级别上，可以进行综合测评分析。

从经验上说，归一化是让不同维度之间的特征在数值上有一定比较性，可以大大提高分类器的准确性。

**归一化是把数据映射到（0，1）区间内，一般方法是(x-min(x))/(max(x)-min(x)),而标准化是一种统计的处理，基于正态分布的假设，但是即使数据不服从正态分布，也可以用此法，区别是标准化后的数据可正可负，但是一般绝对值不会太大。具体的办法是(x-mean(x))/std(x)。标准化后的数据的均值＝0，标准差＝1**

**归一化（Normalization）**

        1.把数据变为（0，1）之间的小数。主要是为了方便数据处理，因为将数据映射到0～1范围之内，可以使处理过程更加便捷、快速。

        2.把有量纲表达式变换为无量纲表达式，成为纯量。经过归一化处理的数据，处于同一数量级，可以消除指标之间的量纲和量纲单位的影响，提高不同数据指标之间的可比性。

        主要算法：

        1.线性转换，即min-max归一化（常用方法）

        y=(x-min)/(max-min)

        2. 对数函数转换

        y=log10(x)

        3.反余切函数转换

        y=atan(x)\*2/PI

**标准化（Standardization）**

        数据的标准化是将数据按比例缩放，使之落入一个小的特定区间。

        主要方法：

        1.z-score标准化，即零-均值标准化（常用方法）

        y=(x-μ)/σ

        是一种统计的处理，基于正态分布的假设，将数据变换为均值为0、标准差为1的标准正态分布。但即使数据不服从正态分布，也可以用此法。特别适用于数据的最大值和最小值未知，或存在孤立点。

        2.小数定标标准化

        y=x/10^j  （j确保max(|y|)<1）

        通过移动x的小数位置进行标准化

        3.对数Logistic模式

        y=1/(1+e^(-x))

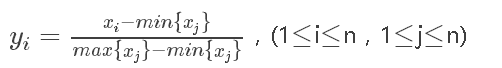
**正则化（Regularization）**

        用一组与原不适定问题相“邻近”的适定问题的解，去逼近原问题的解，这种方法称为正则化方法。如何建立有效的正则化方法是反问题领域中不适定问题研究的重要内容。通常的正则化方法有基于变分原理的Tikhonov 正则化、各种迭代方法以及其它的一些改进方法。

**总的来说**，归一化是为了消除不同数据之间的量纲，方便数据比较和共同处理，比如在神经网络中，归一化可以加快训练网络的收敛性；标准化是为了方便数据的下一步处理，而进行的数据缩放等变换，并不是为了方便与其他数据一同处理或比较，比如数据经过零-均值标准化后，更利于使用标准正态分布的性质，进行处理；正则化而是利用先验知识，在处理过程中引入正则化因子(regulator)，增加引导约束的作用，比如在逻辑回归中使用正则化，可有效降低过拟合的现象。

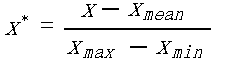
### 离差标准化

通常意义上的归一化，也称**Min-Max标准化，**离差标准化是将某变量中的观察值减去该变量的最小值，然后除以该变量的极差。即



其中max{xj}为样本数据的最大值，min{xj}为样本数据的最小值。

如果想要将数据映射到[-1,1]，则将公式换成：



这种方法有**一个缺陷就是当有新数据加入时，可能导致max和min的变化，需要重新定义。**

当然，在构造类对象的时候也可以直接指定最大最小值的范围：feature\_range=(min, max)，此时应用的公式变为：

X\_std=(X-X.min(axis=0))/(X.max(axis=0)-X.min(axis=0))

X\_scaled=X\_std/(max-min)+min

经过离差标准化后，各种变量的观察值的数值范围都将在〔0，1〕之间，并且经标准化的数据都是没有单位的纯数量。离差标准化是消除量纲（单位）影响和变异大小因素的影响的最简单的方法。有一些关系系数（例如绝对值指数尺度）在定义时就已经要求对数据进行离差标准化，但有些关系系数的计算公式却没有这样要求，当选用这类关系系数前，不妨先对数据进行标准化，看看分析的结果是否为有意义的变化。

将属性缩放到一个指定的最大和最小值（通常是1-0）之间，这可以通过preprocessing.MinMaxScaler类实现。

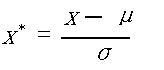
使用这种方法的目的包括：

1、对于方差非常小的属性可以增强其稳定性。

2、维持稀疏矩阵中为0的条目。

### 标准差标准化

也叫Z-score 标准化，经过处理的数据符合标准正态分布，即均值为0，标准差为1，其转化函数为：



其中μ为所有样本数据的均值，σ为所有样本数据的标准差。

经过 Z-score 标准化后，各变量将有约一半观察值的数值小于0，另一半观察值的数值大于0，变量的平均数为0，标准差为1。经标准化的数据都是没有单位的纯数量。它是当前用得最多的数据标准化方法。如果特征非常稀疏，并且有大量的0（现实应用中很多特征都具有这个特点），Z-score 标准化的过程几乎就是一个除0的过程，结果不可预料。

大于0说明高于平均水平，小于0说明低于平均水平。

z-score标准化方法适用于属性A的最大值和最小值未知的情况，或有超出取值范围的离群数据的情况。

示例：

Sklearn的StandardScaler提供了对数据进行**标准差标准化处理**的功能，下面我们用PCA算法来演示标准化处理的影响，数据我们选用的是UCI数据集中的Wine数据集，该数据集的特征值分别代表了葡萄酒的酒精度、苹果酸等的含量，因此取值范围差异较大，例如酒精度的取值范围在（，）之间，苹果酸的取值范围为（，）之间。代码如下：

from \_\_future\_\_ import print\_function

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

from sklearn.naive\_bayes import GaussianNB

from sklearn import metrics

import matplotlib.pyplot as plt

from sklearn.datasets import load\_wine

from sklearn.pipeline import make\_pipeline

print(\_\_doc\_\_)

RANDOM\_STATE = 42

FIG\_SIZE = (10, 7)

features, target = load\_wine(return\_X\_y=True)

# Make a train/test split using 30% test size

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, target,

test\_size=0.30,

random\_state=RANDOM\_STATE)

# Fit to data and predict using pipelined GNB and PCA.

unscaled\_clf = make\_pipeline(PCA(n\_components=2), GaussianNB())

unscaled\_clf.fit(X\_train, y\_train)

pred\_test = unscaled\_clf.predict(X\_test)

# Fit to data and predict using pipelined scaling, GNB and PCA.

std\_clf = make\_pipeline(StandardScaler(), PCA(n\_components=2), GaussianNB())

std\_clf.fit(X\_train, y\_train)

pred\_test\_std = std\_clf.predict(X\_test)

# Show prediction accuracies in scaled and unscaled data.

**print**('**\n**Prediction accuracy for the normal test dataset with PCA')

**print**('{:.2%}**\n**'.format(metrics.accuracy\_score(y\_test, pred\_test)))

**print**('**\n**Prediction accuracy for the standardized test dataset with PCA')

**print**('{:.2%}**\n**'.format(metrics.accuracy\_score(y\_test, pred\_test\_std)))

# Extract PCA from pipeline

pca = unscaled\_clf.named\_steps['pca']

pca\_std = std\_clf.named\_steps['pca']

# Show first principal componenets

**print**('**\n**PC 1 without scaling:**\n**', pca.components\_[0])

**print**('**\n**PC 1 with scaling:**\n**', pca\_std.components\_[0])

# Scale and use PCA on X\_train data for visualization.

scaler = std\_clf.named\_steps['standardscaler']

X\_train\_std = pca\_std.transform(scaler.transform(X\_train))

# visualize standardized vs. untouched dataset with PCA performed

fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=FIG\_SIZE)

**for** l, c, m **in** zip(range(0, 3), ('blue', 'red', 'green'), ('^', 's', 'o')):

ax1.scatter(X\_train[y\_train == l, 0], X\_train[y\_train == l, 1],

color=c,

label='class %s' % l,

alpha=0.5,

marker=m

)

**for** l, c, m **in** zip(range(0, 3), ('blue', 'red', 'green'), ('^', 's', 'o')):

ax2.scatter(X\_train\_std[y\_train == l, 0], X\_train\_std[y\_train == l, 1],

color=c,

label='class %s' % l,

alpha=0.5,

marker=m

)

ax1.set\_title('Training dataset after PCA')

ax2.set\_title('Standardized training dataset after PCA')

**for** ax **in** (ax1, ax2):

ax.set\_xlabel('1st principal component')

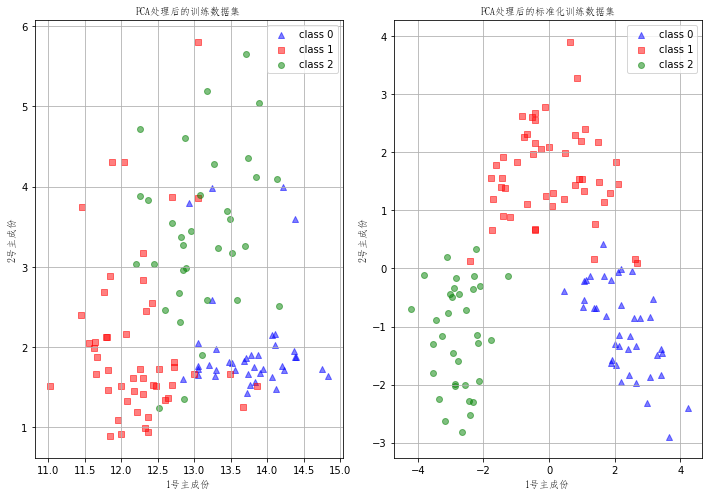
ax.set\_ylabel('2nd principal component')

ax.legend(loc='upper right')

ax.grid()

plt.tight\_layout()

plt.show()



结果是显著的，我们可以看出第一个主成份在非标准化数据集

### 规范化

规范化是将不同变化范围的值映射到相同的固定范围，常见的是[0,1]，此时也称为归一化。《机器学习》周志华

正则化的过程是将每个样本缩放到单位范数（每个样本的范数为1），如果后面要使用如二次型（点积）或者其它核方法计算两个样本之间的相似性这个方法会很有用。

将每个样本变换成unit norm。Normalization主要思想是对每个样本计算其p-范数，然后对该样本中每个元素除以该范数，这样处理的结果是使得每个处理后样本的p-范数（l1-norm,l2-norm）等于1。

|  |  |
| --- | --- |
| 1  2 | X = [[ 1, -1, 2],[ 2, 0, 0], [ 0, 1, -1]]  sklearn.preprocessing.normalize(X, norm='l2') |

得到：

|  |  |
| --- | --- |
| 1 | array([[ 0.40, -0.40, 0.81], [ 1, 0, 0], [ 0, 0.70, -0.70]]) |

可以发现对于每一个样本都有，0.4^2+0.4^2+0.81^2=1,这就是L2 norm，变换后每个样本的各维特征的平方和为1。类似地，L1 norm则是变换后每个样本的各维特征的绝对值和为1。还有max norm，则是将每个样本的各维特征除以该样本各维特征的最大值。

在度量样本之间相似性时，如果使用的是二次型kernel，需要做Normalization。

Scaling inputs to unit norms is a common operation for text classification or clustering for instance. For instance the dot product of two l2-normalized TF-IDF vectors is the cosine similarity of the vectors and is the base similarity metric for the Vector Space Model commonly used by the Information Retrieval community。P1999

**sklearn.preprocessing.Normalizer**

## 离散特征二值化

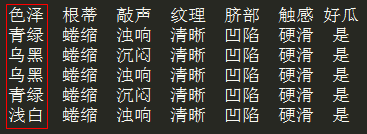
scikit-learn中离散特征二值化

【http://blog.csdn.net/u012328159/article/details/71617381?utm\_source=itdadao&utm\_medium=referral】

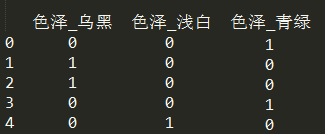
最近在看西瓜书用scikit-learn中的CART去跑西瓜数据集，结果遇到麻烦了，西瓜数据集特征不光离散的，而且还是中文的。。（PS：其实我们的数据集中特征值常常是离散的类别，这个很正常），但在scikit-learn中不支持这种离散的类别特征作为输入，这点不得不说weka的人性化，直接输入原始数据集就可以了。。为了解决这个问题，就要用到独热编码（One-Hot Encoding），下面来说下这个One-Hot Encoding：

* **One-Hot Encoding**

关于one-hot编码的定义，简单的说就是用n位bit（0和1）表示n个状态，并且任意时候只有一位有效数字。直接看例子，我觉得举例子比较好理解。用西瓜数据集（只选了部分样本）举例子吧，西瓜数据集如下：



就拿“色泽”这个属性来看看one-hot encoding，显然色泽有“青绿，乌黑，浅白”三个不同的离散值（类别），那么就需要三位二进制数来表示这个特征，则边玩码后应该是这样的:



来讲下上图中数据是怎样产生的，即one-hot encoding是怎样编码的。显然“色泽”这个属性的三个离散值按照[乌黑，浅白，青绿]排列的，来看第一个样本（图中index为0的）色泽=青绿，因此和[乌黑，浅白，青绿]一比对，对应的编码就是001，依次类推2,3,4,5个样本色泽这个属性的编码分为为100,100,001,010。其实说白了就是把一个属性变成了“三个”，因为模型才不关心你几个特征呢，它只管你给他他能计算的，能挖掘出信息的。这就是one-hot编码的大概内容。下面回到scikit-learn中的实现：自然而言想到sklearn的 的Preprocessing模块中的OneHotEncoder函数，这个函数的具体细节大家去看文档[链接](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html" \l "sklearn.preprocessing.OneHotEncoder" \t "_blank)。但这个函数最坑的地方在于没法对字符串类型的变量进行编码。。这个是最让人无语的地方，不知道后续的更新版本中会不会支持字符串类型的变量。下面只能造个例子来展示下OneHotEncoder怎么用：

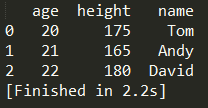
* 造的数据集

**[python]** [view plain](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "view plain) [copy](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "copy)

[print](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "print)[?](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "?)

1. **import** pandas as pd
2. **from** sklearn.preprocessing **import** OneHotEncoder
3. **from** sklearn.preprocessing **import** LabelBinarizer
4. **from** sklearn.feature\_extraction **import** DictVectorizer
6. data = pd.DataFrame({'name':['Tom','Andy','David'],'age':[20,21,22],'height':[175,165,180]})
7. **print**(data)

运行结果如下：



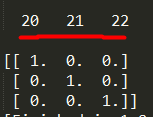
下面使用OneHotEncoder类来对特征进行onehot编码：

**[python]** [view plain](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "view plain) [copy](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "copy)

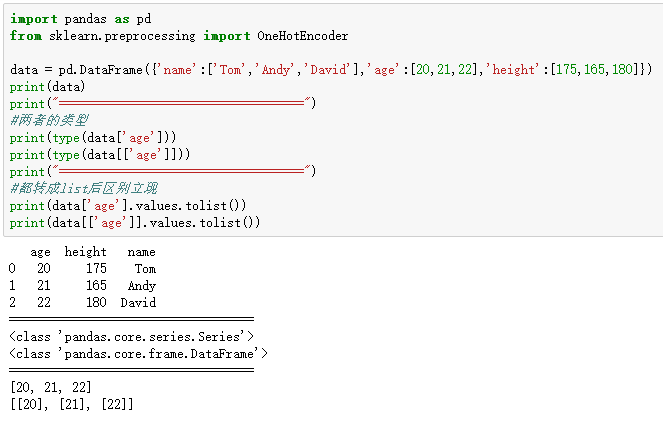
[print](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "print)[?](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "?)

1. #新版本不允许传入一维数组，因此传入data[['age']],sparse=False为不使用稀疏矩阵表示
2. arr = OneHotEncoder(sparse = False).fit\_transform(data[['age']])
3. **print**(arr)

输出结果为（OneHotEncoder会按属性排序，形成有序值，因此下面三列是20,21,22）：



讲到这个地方，要插一个插曲，也是遇到的坑，记录一下，就是当时OneHotEncoder(sparse = False).fit\_transform(data['age'])参数传的是data['age']，然后有提示就是新版不再允许传入一维数组，一直没搞懂data['age']和data[['age']]的区别，请教石神后，方知data['age']返回的是series，data[['age']]因为传入的是列表['age']所以返回的是dataframe，知道这点区别，为了更加弄明白本质区别是什么，做了实验，相信看完下面的对比，还有和我一样不明白的同学能够有一个清晰的认识：



一目了然，转成列表后，一个是一维的一个是二维的。

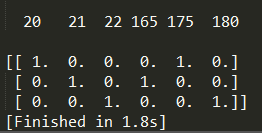
回到正题，还可以传入多个特征一次性处理完：

**[python]** [view plain](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "view plain) [copy](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "copy)

[print](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "print)[?](http://blog.csdn.net/u012328159/article/details/71617381?utm_source=itdadao&utm_medium=referral" \t "_blank" \o "?)

1. arr = OneHotEncoder(sparse = False).fit\_transform(data[['age','height']])
2. **print**(arr)

结果为（注意：列名20 21.....是我为了方便大家看加的，并不是实际输出结果）：



* **LabelBinarizer()**

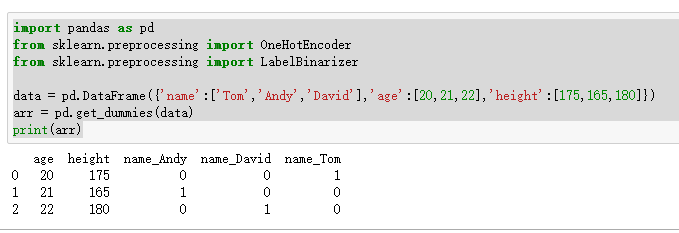
OneHotEncoder()千般好万般好，唯一的不好就是这货没法处理字符串类型的特征值。。比如我们处理‘name’属性就会报错。。提示不能处理字符串类型。。没办法生活还得继续啊，只能曲线救国了，查看文档后发现了LabelBinarizer()类，大家看名字也能看出是干嘛的，对，就是用于类别（label）编码的，上代码：



发现完全可以解决问题，但 新问题是LabelBinarizer()只能传入一维数组处理，不接受二维的，也就是说如果你想一次对多个特征处理，不能全部传入。。。博主试图用pandas的apply（）解决，但无奈水平不到家，没解决，有解决的可以在下方评论贴出，只求别太麻烦。

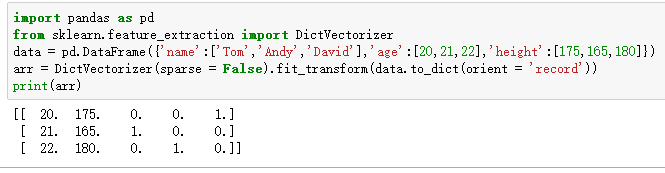
* 新大陆1(get\_dummies())

经过探索发现pandas的get\_dummies()函数可以完美解决问题，这个函数的功能用官方文档的话就是Convert categorical variable into dummy/indicator variables，把类别型的变量转换成dummy variables（用0和1表示类别的出现）。上代码看效果吧：



* **新大陆2(**DictVectorizer**)**

翻了参考资料,发现sklearn的feature\_extraction模块中提供了DictVectorizer类，即把字典向量化，采用0/1二值向量化。上代码：



### 特征二值化（Binarization）

给定阈值，将特征转换为0/1

|  |  |
| --- | --- |
| 1  2 | binarizer = sklearn.preprocessing.Binarizer(threshold=1.1)  binarizer.transform(X) |

### 标签二值化（Label binarization）

lb = sklearn.preprocessing.LabelBinarizer()

### 类别特征编码

有时候特征是类别型的，而一些算法的输入必须是数值型，此时需要对其编码。

|  |  |
| --- | --- |
| 1  2  3 | enc = preprocessing.OneHotEncoder()  enc.fit([[0, 0, 3], [1, 1, 0], [0, 2, 1], [1, 0, 2]])  enc.transform([[0, 1, 3]]).toarray() #array([[ 1., 0., 0., 1., 0., 0., 0., 0., 1.]]) |

上面这个例子，第一维特征有两种值0和1，用两位去编码。第二维用三位，第三维用四位。

### .标签编码（Label encoding）

le = sklearn.preprocessing.LabelEncoder()

le.fit([1, 2, 2, 6])

le.transform([1, 1, 2, 6]) #array([0, 0, 1, 2])

#非数值型转化为数值型

le.fit(["paris", "paris", "tokyo", "amsterdam"])

le.transform(["tokyo", "tokyo", "paris"]) #array([2, 2, 1])

## 特征中含异常值时

|  |  |
| --- | --- |
| 1 | sklearn.preprocessing.robust\_scale |

此Scaler删除中位数并根据分位数范围缩放数据（默认为**四分位距**：Interquartile Range）。 IQR是第一个四分之一（第25个分位数）和第三个四分位数（第75个分位数）之间的范围。

数据集的标准化是许多机器学习估计器的常见要求。 通常，这是通过去除平均值和缩放到单位方差来完成的。 然而，异常值常常会影响样本平均值/方差为负。 在这种情况下，中位数和四分位数范围往往会给出更好的结果。

## 生成多项式特征

这个其实涉及到特征工程了，多项式特征/交叉特征。

|  |  |
| --- | --- |
| 1  2 | poly = sklearn.preprocessing.PolynomialFeatures(2)  poly.fit\_transform(X) |

原始特征：[http://img.blog.csdn.net/20160203140705738](http://img.blog.csdn.net/20160203140705738)

转化后：[http://img.blog.csdn.net/20160203140725613](http://img.blog.csdn.net/20160203140725613)

## 插补处理

preprocessing.Imputer

## 独热编码

**sklearn.preprocessing.OneHotEncoder**

## 特征选择

## 降维

# 一般线性模型

## 线性模型概述

首先，给定一个示例，它具有d个属性，我们将它表示为，其中表示在第i个属性上的取值。线性模型试图学得一个与属性有关的线性组合并据此来进行预测的函数，即

，（1.1）

一般用向量形式表示，即

，（1.2）

其中。和b的值确定之后，模型就能得到确定了。

给定数据集，其中。线性回归即假设特征满足线性关系，根据给定的训练集D训练一个线性模型，并用此模型进行预测。

我们先考虑一种最简单的情形：输入属性的数目只有一个。那么预测函数为。接下来就是和b的确定。显然，关键在于使与实际值的差别尽可能的小。因为均方误差是回归任务中最常用的性能度量，所以我们将均方误差作为衡量差别的工具，目的是让均方误差最小化，即

，（1.3）

基于均方误差最小化来进行模型求解的方法称为“最小二乘法”。在线性回归中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小。均方误差的几何意义对应的就是欧几里得距离，简称欧式距离。

求解w和b使最小化的过程，称为线性回归模型的最小二乘“参数估计”。我们可将分别对w和b求导，得到

，（1.4）

，（1.5）

然后令偏导数为零，即可得到w和b最优解的闭式解

，（1.6）

，（1.7）

其中为的均值。

讨论完只有一个属性的简单形式，我们来更一般的情形，即样本具有d个属性，此时我们试图学得的函数为。这称为“多元线性回归”。

类似的，可利用最小二乘法来对和b进行估计。为了便于讨论，我们把和b用向量形式来表示，把数据集D表示为一个大小的矩阵，其中每行对应一个示例，该行前个元素对应于示例的d个属性值，最后一个元素都为1，即

，

再把标记也写成向量形式，则类似于式（1.3），有

.（1.8）

令，对求导得到

. （1.9）

同样的，令上式为零就可以得到最优解的闭式解。但是由于涉及矩阵的逆运算，我们需要做一个简单的讨论。

当为满秩矩阵或者正定矩阵时，令式（1.9）为零可得

.（1.10）

令，则最终学得的多元线性回归模型为

.（1.11）

然而，现实任务中往往不是满秩矩阵。例如在有的任务中，我们会遇到示例的属性个数超过样例数，也就是的列数要多于行数，显然不满秩。这时候解出的不唯一，它们都能够使均方误差最小化。这时候决定选择哪一个解作为输出，就要由学习算法的归纳偏好决定了。常见的做法就是引入正则化项。

参考资料：《机器学习》周志华

## 线性判别分析 P60

# 非线性回归

<http://blog.csdn.net/loveliuzz/article/details/78024355>

<http://blog.csdn.net/AriesSurfer/article/details/41310525>

高正明，赵娟，贺升平 多项式回归分析及回归方程的显著性检验[J] 中国核科学技术进展报告（第二卷）2011.10

## Logistic Regression

1、概率

（1）定义：概率（Probability）：对一件事情发生的可能性的衡量。

（2）取值范围：

（3）计算方法：根据个人置信、根据历史数据、根据模拟数据

（4）条件概率：在事件B已经发生的情况下，事件A发生的概率等于事件A、B同时发生的概率除以B事件发生的概率。



1. Logistic Regression (逻辑回归)

Logistic回归为概率型非线性回归模型，是研究二分类观察结果与一些影响因素之间关系的一种多变量分析方法。通常的问题是，研究某些因素条件下某个结果是否发生，比如医学中根据病人的一些症状来判断它是否患有某种病。

Regression问题的常规步骤为：

寻找h函数（即hypothesis）；

构造J函数（损失函数）；

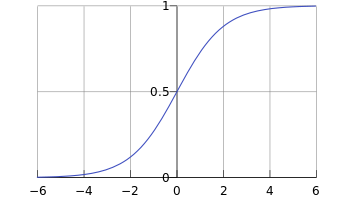
想办法使得J函数最小并求得回归参数（θ）。

（1）构造预测函数h

Logistic回归虽然名字里带“回归”，但是它实际上是一种分类方法，主要用于两分类问题（即输出只有两种，分别代表两个类别），所以利用了Logistic函数（或称为Sigmoid函数），函数形式为：



Sigmoid 函数在有个很漂亮的“S”形，如下图所示（引自维基百科）：



➀对于线性边界的情况，边界形式如下：



构造预测函数为：



函数的值有特殊的含义，它表示结果取1的概率，因此对于输入x分类结果为类别1和类别0的概率分别为：

 （1）

构造损失函数J

Cost函数和J函数如下，它们是基于最大似然估计推导得到的：

****



下面详细说明推导的过程：

1. 式综合起来可以写成：



取似然函数为：



对数似然函数为：



最大似然估计就是求使取最大值时的θ，其实这里可以使用梯度上升法求解，求得的θ就是要求的最佳参数。但是，由于，此式乘了一个负的系数-1/m，所以取最小值时的θ为要求的最佳参数，所以梯度下降法求的最小值。

更新过程：







更新过程可以写成：



下面介绍向量化的过程：

约定训练数据的矩阵形式如下，x的每一行为一条训练样本，而每一列为不同特征取值：

，，





g(A)的参数A为一列向量，所以实现g函数时要支持列向量作为参数，并返回列向量。

更新过程可以改为：



综上所述，向量化后更新的步骤如下：

1. 求；
2. ；
3. 。

## 多项式回归

4.2.1多项式拟合的普适性

根据高等数学理论，若连续函数在的某一领域内具有直到（n+1）阶的导数，则函数在处可展开为n阶Taylor级数形式：

 其中是与之间的某个值，这说明任意连续的在领域内具有阶导数的函数均可以简化为多项式函数表达式：



因此，对于任意母体的n组统计独立的样本数据（）（），在母体规律存在的前提下，必然存在一组系数，使式表征母体分布规律，因此多项式拟合对统计数据的普适的拟合方法，适用于任意母体规律存在的样本数据回归分析过程。

4.2.2 多项式回归参数的最小二乘估计

对于任意实际问题的n组观测数据（）（），有



设





代入上式可得到其矩阵表达形式：



根据最小二乘估计原则，回归参数的估计值应使估计值与观测值之间的残差达到最小，即：



其中即有：



即有当存在时：



此外，零均值正态分布的随机误差的方差估计量为：



其中样本残差：



# k-邻近算法

参考资料：基于K近邻的分类算法研究\_桑应宾.caj

《机器学习》周志华

《机器学习实战》

## k-邻近算法概述

k-近邻（k-Nearest Neighbor，简称kNN）是一种常用的监督学习方法。它具有精度高、对异常值不敏感、无数据输入假定的特点。不足之处在于计算复杂度高、空间复杂度高。适用于数值型和标称型数据。

### 基本思想

它的工作原理是：给定一个训练样本集，基于某种距离度量找出训练集中与新数据最靠近的k个训练样本，然后将根据这k个“邻居”的特征来预测新数据的分类标签。它并没有显示的训练过程，而仅仅在训练阶段将样本保存起来，待收到测试样本后再进行处理，它的训练时间开销为零。

### 算法概述

近邻算法是基于实例学习的分类算法中比较常用的一种方法。令，其中每一个样本所属的类别均已知。对于测试样本点，在集合中距离它最近的点记为，那么，最近邻规则的分类方法就是把点分为所属的类别。最近邻点的标记是一个随机变量。的概率为后验概率。

当样本个数非常大的时候，有理由认为距离足够近，使得。

因为这恰好是状态位于的概率，因此最近邻规则自然是真实概率的一个有效的近似。

K近邻规则是最近邻规则的一个推广。从这个规则名词可知，该规则将是一

个测试数据点分类为与它最接近的K个近邻中出现最多的那个类别。K近邻算

法从测试样本点开始生长，不断的扩大区域，直到包含进K 个训练样本点为止，并且把测试样本点归为这最近的K个训练样本点中出现频率最大的类别。

如果K值固定，并且允许训练样本个数趋向于无穷大，那么，所有的这K 个

近邻都将收敛于x 。如同最近邻规则一样，K个近邻的标记都是随机变量，概率

都是相互独立的。而最近邻规则则以概率选取类别。而根据K近邻规则，只有当K个最近邻中的大多数的标记记为,才判定为类别。做出这样判定的概率为



通常K值越大，选择类别的概率也越大。

K近邻规则可以被看作是另一种从样本中估计后验概率的方法。为了得到可靠的估计必须使得 K值越大越好。另一方面，又希望x的K个近邻距离x越近越好，因为这样才能保证尽可能逼近。在选取K值的时候，就不得不做出某种折衷。只有当n趋近于无穷大时，才能保证K近邻规则几乎是最优的分类规则。

算法步骤：

（1）构建训练样本集合X。

（2）设定K的初值。K值的确定没有一个统一的方法（根据具体问题选取的K值可能有较大的区别）。一般方法是先确定一个初始值，然后根据实验结果不断调试，最终达到最优。

（3）在训练样本集中选出与待测样本最近的K个样本。假定样本点x属于n维空间，样本之间的“近邻”一般由欧式距离来度量。设第i个样本，其中表示第i个样本第个特征属性值。那么两个样本和之间的欧氏距离定义为：



1. 给定一个待分类的样本，表示与距离最近的K个样本，设离散的目标函数（分类问题）为，表示第i个类别的标签，标签集合定义为，

，

表示对的估计，当时，；否则，。

1. 就是待测样本的类别。

## 算法范例：k-邻近算法应用范例

## 小结

# 支持向量机

参考资料：1、机器学习中的算法(2)-支持向量机(SVM)基础 - LeftNotEasy - 博客园 <https://www.cnblogs.com/LeftNotEasy/archive/2011/05/02/basic-of-svm.html>

1. SVM支持向量机算法 - CSDN博客 <http://blog.csdn.net/androidlushangderen/article/details/42780439>
2. 《机器学习》周志华

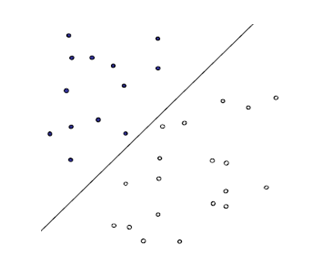
## 概述

支持向量机(Support Vector Machine，简称SVM)是一种高效的的监督学习算法。它的主要思想可以概括为2点：（1）针对线性可分情况进行分析。（2）对于线性不可分的情况，通过使用核函数，将低维线性不可分空间转化为高维线性可分的情况，然后再进行分析。

## 支持向量机算法

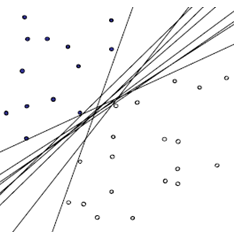
### 线性分类器

首先给出一个非常非常简单的分类问题（线性可分），我们要用一条直线，将下图中黑色的点和白色的点分开，很显然，图上的这条直线就是我们要求的直线之一（可以有无数条这样的直线）。



假如说，我们将黑色的点用-1表示， 白色的点用+1表示，直线，这儿的、是向量，其实写成这种形式也是等价的，当向量的维度=2的时候，表示二维空间中的一条直线，当的维度=3的时候，表示3维空间中的一个平面，当x的维度=n > 3的时候，表示n维空间中的n-1维超平面。

但是，我们怎样才能取得一个最优的划分直线呢？下图的直线表示几条可能的。



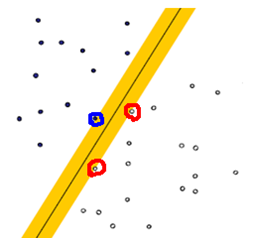
从直观上来说，就是分割的间隙越大越好，把两个类别的点分得越开越好。在SVM中，称为最大间隔，是SVM的一个理论基础之一。选择使得间隙最大的函数作为分割平面是由很多道理的，比如说从概率的角度上来说，就是使得置信度最小的点置信度最大，从实践的角度来说，这样的效果非常好，等等。

假设超平面能将训练样本正确分类，即对于，则有

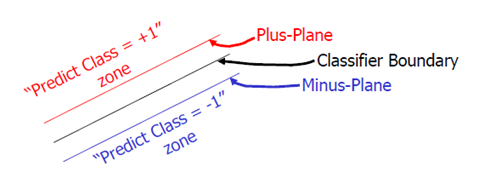
 （3.1）

如下图，被红色和蓝色的线圈出来的点，即距离超平面最近的几个训练样本点就是所谓的支持向量(support vector)。两个异类支持向量到超平面的距离之和被称为“间隔”，

， （3.2）



下图就是一个对之前说的类别中的间隙的一个描述。表示的是超平面，红色和蓝色的线就是支持向量所在的面，红色、蓝色线之间的间隙就是我们要最大化的分类间的间隙。



想要找到具有最大间隔的划分超平面，也就是要找到能满足式（3.1）中约束的参数和b，使得最大，即

（3.3）

显然，为了最大化间隔，仅需最大化，这等价于最小化。于是，式（3.3）可重写为

（3.4）

这就是支持向量机（SVM）的基本型。

### 对偶问题

式（3.4）优化问题可以用拉格朗日乘子法得到其“对偶问题”并求解，使用了KKT条件的理论。具体地说就是对（3.4）的每条约束添加拉格朗日乘子，则该问题的拉格朗日函数可写为

 （3.5）

其中。令式（3.5）对和b的偏导为零可得

，（3.6）

。（3.7）

将两式带回（3.5）得到对偶问题的表达式

（3.8）



解出后，求出和b即可得到模型

 （3.9）

（3.8）就是我们需要最终优化的式子。至此，得到了线性可分问题的优化式子。求解这个式子，有很多的方法，比如SMO等等。

### 软间隔

接下来谈谈线性不可分的情况，因为线性可分这种假设实在是太有局限性了。线性不可分的分类图就是我们没有办法用一条直线去将其分成两个区域，每个区域只包含一种颜色的点。

要想在这种情况下的分类器，有两种方式，一种是用曲线去将其完全分开，曲线就是一种非线性的情况，跟之后将谈到的核函数有一定的关系；另外一种还是用直线，不过不用去保证可分性，就是包容那些分错的情况，不过我们得加入惩罚函数，使得点分错的情况越合理越好。其实在很多时候，不是在训练的时候分类函数越完美越好，因为训练函数中有些数据本来就是噪声，可能就是在人工加上分类标签的时候加错了，如果我们在训练（学习）的时候把这些错误的点学习到了，那么模型在下次碰到这些错误情况的时候就难免出错了。这种学习的时候学到了“噪声”的过程就是一个过拟合（over-fitting），这在机器学习中是一个大忌，我们宁愿少学一些内容，也坚决杜绝多学一些错误的知识。还是回到主题，用直线怎么去分割线性不可分的点：

我们可以为分错的点加上一点惩罚，对一个分错的点的惩罚函数就是这个点到其正确位置的距离。我们在原函数上面加上一个惩罚函数，并且带上其限制条件为：

，（3.10）



这就是常用的“软间隔支持向量机”。C是一个由用户去指定的系数，表示对分错的点加入多少的惩罚，当C很大的时候，分错的点就会更少，但是过拟合的情况可能会比较严重，当C很小的时候，分错的点可能会很多，不过可能由此得到的模型也会不太正确，所以如何选择C是有很多学问的，不过在大部分情况下就是通过经验尝试得到的。

接下来就是同样的，求解一个拉格朗日对偶问题，得到一个原问题的对偶问题的表达式：

 （3.12）



是与线性可分的对偶问题表达式的不同之处。在线性不可分情况下得到的对偶问题，不同的地方就是的范围从[0, +∞)，变为了[0, C]，增加的惩罚没有为对偶问题增加什么复杂度。

### 核函数

线性不可分的情况下，如果使用某些非线性的方法，可以得到将两个分类完美划分的曲线，比如接下来将要说的核函数。一个低维的样本集映射到高维则可以变成线性可分，那样才能使用SVM工作。

设映射函数为，则映射后的空间分类函数变成



但是，如果拿到低维数据直接映射到高维的话，维度的数目会呈现爆炸性增长。所以这里需要引入核函数（kernal function）。核函数的思想是寻找一个函数，这个函数使得在低维空间中进行计算的结果和映射到高维空间中计算内积的结果相同。

这样就避开直接在高维空间中进行计算，而最后的结果却是等价的。现在，分类函数就变成了这样：



其中就是核函数。由于对任意数据集找到它合适的映射是困难的且没有必要，所以通常会从常用核函数中选择。

## 算法范例：支持向量机分类应用范例

## 小结

# 决策树

<https://www.cnblogs.com/pinard/p/6053344.html>

《机器学习》清华大学出版社 周志华著

## 简介

决策树是一类常见的机器学习方法，它是基于数结构来进行决策的，这恰是人类在面临决策问题时一种很自然的处理机制。一般的，一颗决策树包括一个根结点、若干个内部结点和若干个叶结点；叶结点对应于决策结果，其他每个结点则对应于一个属性测试；每个结点包含的样本集合根据属性测试的结果被划分到子结点中；根结点包含样本全集。从根结点到每个叶结点的路径对应了一个判定测试序列。决策树学习的目的是为了产生一颗泛化能力强，即处理未见示例能力强的决策树。

### 决策树表示法

## 决策树算法

8.2.1信息增益

“信息熵”是度量样本集合纯度最常用的一种指标。假定当前样本集合D中第k类样本所占的比例为（k=1,2,...,），则D的信息熵定义为



的值越小，则D的纯度越高。

假定离散属性a有V个可能的取值，若使用a 来对样本集D进行划分，则会产生V个分支结点，其中第v个分支结点包含了D中所有在属性a上取值为的样本，记为。可根据定义计算出的信息熵，再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重，可计算出用属性a对样本集D进行划分所获得的“信息增益”



ID3算法就是用信息增益大小来判断当前节点应该用什么特征来构建决策树，用计算出的信息增益最大的特征来建立决策树的当前节点。

著名的C4.5决策树算法不直接使用信息增益，而是使用“增益率”来选择最优划分属性，其定义为：



其中



8.2.2 CART算法的原理

分类回归树(CART,Classification And Regression Tree)算法采用一种二分递归分割的技术，将当前的样本集分为两个子样本集，使得生成的的每个非叶子节点都有两个分支。CART分类树算法使用基尼系数来代替信息增益比，基尼系数代表了模型的不纯度，基尼系数越小，则不纯度越低，特征越好。这和信息增益(比)是相反的。

具体的，在分类问题中，假设有K个类别，第k个类别的概率为, 则数据集D基尼系数的表达式为:



属性a的基尼指数定义为：



CART具体的算法流程：

算法输入是训练集D，基尼系数的阈值，样本个数阈值，输出是决策树T。我们的算法从根节点开始，用训练集递归的建立CART树。

（1）对于当前节点的数据集为D，如果样本个数小于阈值或者没有特征，则返回决策子树，当前节点停止递归。

（2）计算样本集D的基尼系数，如果基尼系数小于阈值，则返回决策树子树，当前节点停止递归。

（3）计算当前节点现有的各个特征的各个特征值对数据集D的基尼系数，。

（4）在计算出来的各个特征的各个特征值对数据集D的基尼系数中，选择基尼系数最小的特征A和对应的特征值a。根据这个最优特征和最优特征值，把数据集划分成两部分D1和D2，同时建立当前节点的左右节点，做节点的数据集D为D1，右节点的数据集D为D2.

（5）对左右的子节点递归的调用1-4步，生成决策树。

对于生成的决策树做预测的时候，假如测试集里的样本A落到了某个叶子节点，而节点里有多个训练样本。则对于A的类别预测采用的是这个叶子节点里概率最大的类别。

由于决策时算法很容易对训练集过拟合，而导致泛化能力差，为了解决这个问题，我们需要对CART树进行剪枝，即类似于线性回归的正则化，来增加决策树的返回能力。CART采用的办法是后剪枝法，即先生成决策树，然后产生所有可能的剪枝后的CART树，然后使用交叉验证来检验各种剪枝的效果，选择泛化能力最好的剪枝策略。

## 决策树学习的适用问题

## 算法范例：决策树回归范例

## 小结

# 人工神经网络

<http://blog.csdn.net/fengbingchun/article/details/50274471>

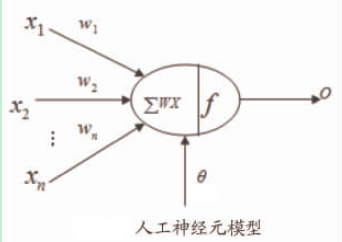
http://blog.csdn.net/zhongkejingwang/article/details/44514073

## 简介

人工神经网络（Artificial Neural Network，ANN）简称神经网络(NN)，是基于生物学中神经网络的基本原理，在理解和抽象了人脑结构和外界刺激响应机制后，以网络拓扑知识为理论基础，模拟人脑的神经系统对复杂信息的处理机制的一种数学模型。该模型以并行分布的处理能力、高容错性、智能化和自学习等能力为特征，将信息的加工和存储结合在一起，以其独特的知识表示方式和智能化的自适应学习能力，引起各学科领域的关注。它实际上是一个有大量简单元件相互连接而成的复杂网络，具有高度的非线性，能够进行复杂的逻辑操作和非线性关系实现的系统。

## 人工神经元结构

人工神经元的研究源于脑神经元学说，19世纪末，在生物、生理学领域，Waldeger等人创建了神经元学说。人工神经网络是由大量处理单元经广泛互连而组成的人工网络，用来模拟脑神经系统的结构和功能。而这些处理单元我们把它称作人工神经元。人工神经网络可看成是以人工神经元为节点，用有向加权弧连接起来的有向图。在此有向图中，人工神经元就是对生物神经元的模拟，而有向弧则是轴突----突触----树突对的模拟。有向弧的权值表示相互连接的两个人工神经元间相互作用的强弱。人工神经元结构如下图所示：

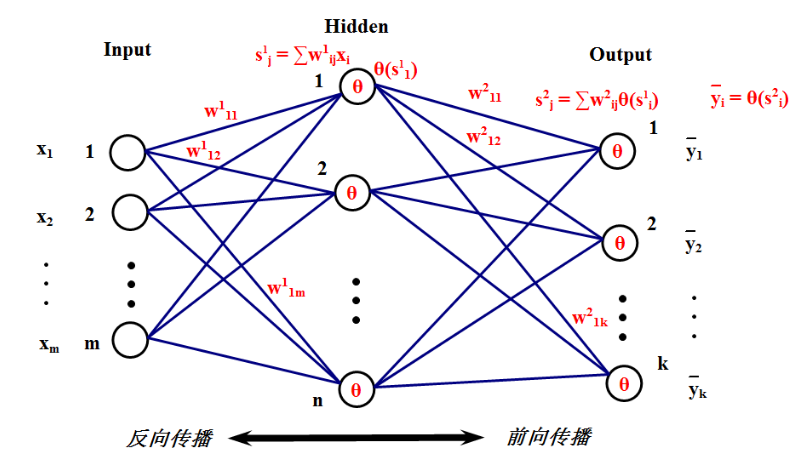


## 人工神经网络算法（BP算法）

9.3.1 BP算法简介

BP神经网络，BP即Back Propagation的缩写，也就是反向传播的意思。它可以对组成前向多层网络的各人工神经元之间的连接权值进行不断的修改，从而使该前向多层网络能够将输入它的信息变换成所期望的输出信息。同时也可以在修改各人工神经元的连接权值时，所依据的是该网络的实际输出与其期望的输出之差，将这一差值反向一层一层的向回传播，来决定连接权值的修改。

9.3.2 BP算法的数学原理



为了简单起见，这里只介绍只有一个隐层的BP网络，多个隐层的也是一样的原理。这个网络的工作原理应该很清楚了，首先，一组输入来到输入层，然后通过与隐层的连接权重产生一组数据作为隐层的输入，然后通过隐层节点的激活函数后变为其中表示隐层的第j个节点产生的输出，这些输出将通过隐层与输出层的连接权重产生输出层的输入，这里输出层的处理过程和隐层是一样的，最后会在输出层产生输出，这里j是指输出层第j个节点的输出。这只是前向传播的过程。在这里，先解释一下隐层的含义，可以看到，隐层连接着输入和输出层，它到底是什么？它就是特征空间，隐层节点的个数就是特征空间的维数，或者说这组数据有多少个特征。而输入层到隐层的连接权重则将输入的原始数据投影到特征空间，比如就表示这组数据在特征空间中第j个特征方向的投影大小，或者说这组数据有多少份量的j特征。而隐层到输出层的连接权重表示这些特征是如何影响输出结果的，比如某一特征对某个输出影响比较大，那么连接它们的权重就会比较大。关于隐层的含义就解释这么多，至于多个隐层的，可以理解为特征的特征。

既然在输出层产生输出了，现在就来分析一下什么东西在影响输出。显然，输入的数据是已知的，变量只有那些个连接权重了，那这些连接权重如何影响输出呢？现在假设输入层第i个节点到隐层第j个节点的连接权重发生了一个很小的变化Δ，那么这个Δ将会对产生影响，导致也出现一个变化Δ，然后产生Δ，然后传到各个输出层，最后在所有输出层都产生一个误差Δe。所以说，权重的调整将会使得输出结果产生变化，那么如何使这些输出结果往正确方向变化呢？这就是接下来的任务：如何调整权重。对于给定的训练样本，其正确的结果已经知道，那么由输入经过网络的输出和正确的结果比较将会有一个误差，如果能把这个误差降到最小，那么就是输出结果靠近了正确结果，就可以说网络可以对样本进行正确分类了。怎样使得误差最小呢？首先，把误差表达式写出来，为了使函数连续可导，这里最小化均方根差，定义损失函数如下：



在这里我们用随机梯度下降来最小化L。随机梯度下降也就是对每个训练样本都使权重往其负梯度方向变化。现在的任务就是求L对连接权重w的梯度。

用表示输入层第i个节点到隐层第j个节点的连接权重，表示隐层第i个节点到输出层第j个节点的连接权重，表示隐层第j个节点的输入，表示输出层第j个几点的输入，区别在右上角标，1表示第一层连接权重，2表示第二层连接权重。那么有：



由于



所以



代入前面式子可得



接下来只需要求出即可。

由于对所有输出层都有影响，所以



由于



所以



代入前面的式子可得：



现在记



则隐层为



输出层为









到这一步，可以看到是什么反向传播了吧？没错，就是误差e！

反向传播过程是这样的：输出层每个节点都会得到一个误差e，把e作为输出层反向输入，这时候就像是输出层当输入层一样把误差往回传播，先得到输出层δ，然后将输出层δ根据连接权重往隐层传输，即前面的式子：



现在再来看第一层权重的梯度：



第二层权重梯度：



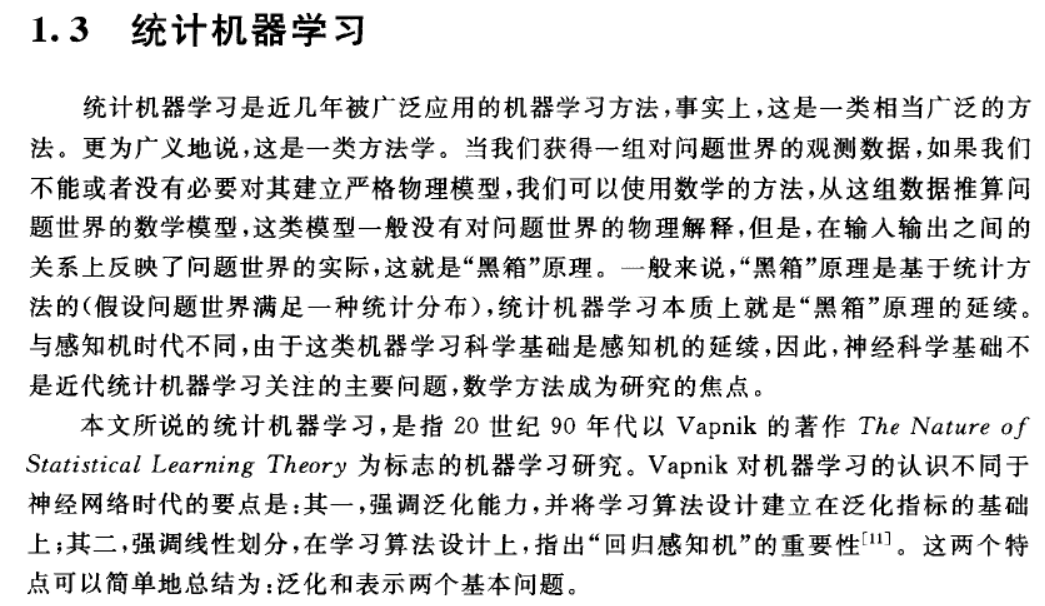
可以看到一个规律：每个权重的梯度都等于与其相连的前一层节点的输出（即和）乘以与其相连的后一层的反向传播的输出（即和）。这样反向传播得到所有的以后，就可以更新权重了。

## 人工神经网络算法适用的问题

## 算法范例：BP神经网络应用范例

## 小结

素材



# AdaBoost算法

参考资料：Adaboost 算法的原理与推导（读书笔记） <http://www.360doc.com/content/18/0112/12/51961713_721308110.shtml>

## 基本思想

AdaBoost，是英文"Adaptive Boosting"（自适应增强）的缩写，由Yoav Freund和Robert Schapire在1995年提出。它的自适应在于：前一个基本分类器分错的样本会得到加强，加权后的全体样本再次被用来训练下一个基本分类器。同时，在每一轮中加入一个新的弱分类器，直到达到某个预定的足够小的错误率或达到预先指定的最大迭代次数。

具体说来，整个AdaBoost 迭代算法就3步：

（1）初始化训练数据的权值分布。如果有N个样本，则每一个训练样本最开始时都被赋予相同的权重：1/N。

（2）训练弱分类器。具体训练过程中，如果某个样本点已经被准确地分类，那么在构造下一个训练集中，它的权重就被降低；相反，如果某个样本点没有被准确地分类，那么它的权重就得到提高。然后，权重更新过的样本集被用于训练下一个分类器，整个训练过程如此迭代地进行下去。

（3）将各个训练得到的弱分类器组合成强分类器。各个弱分类器的训练过程结束后，加大分类误差率小的弱分类器的权重，使其在最终的分类函数中起着较大的决定作用，而降低分类误差率大的弱分类器的权重，使其在最终的分类函数中起着较小的决定作用。换言之，误差率低的弱分类器在最终分类器中占的权重较大，否则较小。

## 算法流程

给定一个训练数据集，其中实例，而实例空间，属于标记集合{-1,+1}，AdaBoost的目的就是从训练数据中学习一系列弱分类器或基本分类器，然后将这些弱分类器组合成一个强分类器。

其算法流程如下：

步骤1。首先，初始化训练数据的权值分布。每一个训练样本最开始时都被赋予相同的权重：1/N。



步骤2。进行多轮迭代，用表示迭代的第多少轮。

a.使用具有权值分布的训练数据集学习，得到基本分类器：



b.计算在训练数据集上的分类误差率：



由上述式子可知，在训练数据集上的误差率就是被误分类样本的权值之和。

c. 计算的系数，表示在最终分类器中的重要程度，目的是为了得到基本分类器在最终分类器中所占的权重：



由上述式子可知，时，，且随着的减小而增大，意味着分类误差率越小的基本分类器在最终分类器中的作用越大。

d.更新训练数据集的权值分布（目的：得到样本的新的权值分布），用于下一轮迭代:





使得被基本分类器误分类样本的权值增大，而被正确分类样本的权值减小。就这样，通过这样的方式，AdaBoost方法能“聚焦于”那些较难分的样本上。其中，是规范化因子，使得成为一个概率分布：



步骤3。组合各个弱分类器：



从而得到最终分类器，如下：



# 随机森林

http://blog.csdn.net/zrjdds/article/details/50133843

## 简介

随机森林是一种比较新的机器学习模型。上世纪八十年代Breiman等人发明分类树的算法（Breiman et al. 1984），通过反复二分数据进行分类或回归，计算量大大降低。2001年Breiman把分类树组合成随机森林（Breiman 2001a），即在变量（列）的使用和数据（行）的使用上进行随机化，生成很多分类树，再汇总分类树的结果。随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元公线性不敏感，结果对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用（Breiman 2001b），被誉为当前最好的算法之一（Iverson et al. 2008）。

随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。

## 随机森林基本原理

随机森林由LeoBreiman（2001）提出，它通过自助法（bootstrap）重采样技术，从原始训练样本集N中有放回地重复随机抽取k个样本生成新的训练样本集合，然后根据自助样本集生成k个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定。其实质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。特征选择采用随机的方法去分裂每一个节点，然后比较不同情况下产生的误差。能够检测到的内在估计误差、分类能力和相关性决定选择特征的数目。单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样品可以通过每一棵树的分类结果经统计后选择最可能的分类。

## 随机森林实现过程

随机森林中的每一棵分类树为二叉树，其生成遵循自顶向下的递归分裂原则，即从根节点开始依次对训练集进行划分；在二叉树中，根节点包含全部训练数据， 按照节点纯度最小原则，分裂为左节点和右节点，它们分别包含训练数据的一个子集，按照同样的规则节点继续分裂，直到满足分支停止规则而停止生长。若节点n上的分类数据全部来自于同一类别，则此节点的纯度I(n)=0，纯度度量方法是Gini准则，即假设P(X)是节点n上属于类样本个数占训练。

具体实现过程如下：

（1）原始训练集为N，应用bootstrap法有放回地随机抽取k个新的自助样本集，并由此构建k棵分类树，每次未被抽到的样本组成了k个袋外数据；

（2）设有个变量，则在每一棵树的每个节点处随机抽取个变量( n )，然后在中选择一个最具有分类能力的变量，变量分类的阈值通过检查每一个分类点确定；

（3）每棵树最大限度地生长, 不做任何修剪；

（4）将生成的多棵分类树组成随机森林，用随机森林分类器对新的数据进行判别与分类，分类结果按树分类器的投票多少而定。

# 模型评价

## 模型评价指标

## 模型评价范例

## 模型选择范例