## 2.1机器学习

**2.1.1 机器学习的定义**

机器学习（Machine Learning）是一门人工智能的学科，它所讨论的是在“从有限观察概括特定问题世界模型的机器学习”也“从有限观察发现观测数据中暗含的各种关系的数据分析”的方法，其领域的主要研究对象是人工智能，研究热点是如何在经验学习中改善具体算法的性能。机器学习主要应用于数据挖掘、语音和手写识别、计算机视觉、生物特征识别、自然语言处理、医学诊断、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、DNA序列测序、证券市场分析、战略游戏和机器运用等领域。

**2.1.2 机器学习的发展史**

20世纪90年代初期，当时的美国副总统提出了一个重要的计划：国家信息基本设施计划（National Information Infrastructure）。这个计划的技术含义主要包含了四个方面内容：

1．不分时间与地域，可以方便地获得信息。

2．不分时间与地域，可以有效地利用信息。

3．不分时间与地域，可以有效地利用软硬件资源。

4．保证信息安全。

“信息有效利用”问题的本质是：如何根据用户特定的需求从海量的数据中建立适当的模型或者发现有用的知识。对计算机科学来言，这就是机器学习。机器学习作为人工智能领域较为年轻的分支，它的发展历程大致上可分为四个时期[40] [43]。

第一阶段是从20世纪50年代中叶到60年代中叶，属于热烈时期。

第二阶段是从20世纪60年代中叶至70年代中叶，被称为机器学习的冷静时期。

第三阶段是从20世纪70年代中叶至80年代中叶，称为复兴时期。

机器学习的最新阶段始于1986年。机器学习开启进入新阶段的主要表现在下列几个方面：

1．机器学习已成为新兴的热门学科，并在高校形成一门课程。它综合了[应用心理学](http://baike.baidu.com/view/46434.htm)、数学、自动化、生物学、神经生理学和[计算机科学](http://baike.baidu.com/view/92404.htm)形成机器学习理论基础。

2．结合多种学习方法，以取长补短等多种形式的形成的集成学习[系统研究](http://baike.baidu.com/view/1446932.htm)正在兴起。特别是连接学习和[符号学习](http://baike.baidu.com/view/6217848.htm)的耦合能够更好地解决连续性[信号处理](http://baike.baidu.com/view/642820.htm)中知识与技能的获取与求精问题而受到重视。

3．人工智能和机器学习的各种基础问题的统一性的观点正在形成。比如问题求解和学习相结合进行、知识表达使得学习的观点发生了通用智能系统的模块学习。类比学习和问题求解结合的基于示例方法已成为[经验](http://baike.baidu.com/view/21717.htm)学习的一个重要方向。

4．随着各种学习方法的应用范围不断拓宽，一部分方法已经形成了商品。归纳学习的知识获取工具已在诊断分类型专家系统中广泛的使用。其中连接学习在声图文的识别应用中占优势；而分析学习已经用于设计综合型的专家系统；另外，强化学习和[遗传算法](http://baike.baidu.com/view/45853.htm)在工程控制应用中有很好的应用前景；最后与[符号系统](http://baike.baidu.com/view/6552866.htm)耦合的[神经网络](http://baike.baidu.com/view/5348.htm)方法已经在智能机器人运动规划和企业的智能管理中发挥着重要作用。

5．机器学习的学术会议空前繁多。国际上除每年一次的机器学习研讨会外，还有遗传算法会议以及计算机学习理论会议等。

**2.1.3 机器学习分类**

机器学习有很多算法，分类方式也有多种。主要的分类方式有基于学习策略的分类、按应用领域分类、基于所获取知识的表示形式分类、学习形式分类和综合分类等。其中最为普遍接受的是按学习形式分类。

机器学习按学习形式分类有监督学习（Supervised Learning）、非监督学习（Unsupervised Learning）和强化学习等几种类型。实践中应用较多的也就是监督学习和非监督学习。

监督学习就是数据样本会告诉计算机在该情形下的正确输出结果，希望计算机能够在面对没有见过的输入样本时也给出正确的输出结果，从而达到预测未知的目的。

根据输出结果是离散还是连续，监督学习可以分为分类问题和回归问题两大类。他们在文字、语音、图像识别、垃圾邮件分类与拦截、网页检索、股票预测等方面有着广泛应用。

非监督学习是指数据样本中没有给出正确的输出结果信息。无监督学习希望从数据中挖掘的正是这一类信息，常见的例子有聚类、关联规则挖掘、离群点检测等等。

强化学习是智能系统从环境到行为映射的学习，以使奖励信号（强化信号）函数值最大，强化学习不同于连接主义学习中的监督学习，主要表现在强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价（通常为标量信号），而不是告诉强化学习系统RLS（Reinforcement Learning System）如何去产生正确的动作。由于外部环境提供的信息很少，RLS必须靠自身的经历进行学习。通过这种方式，RLS在行动评价的环境中获得知识，改进行动方案以适应环境。

## 2.2最小二乘法回归

### 2.2.1线性回归模型

考虑到本课题中数据的多维性，这里用多元回归方程表示：

式中：θi称为假设模型即映射输入特征x与结果y的线性函数h的参数parameters，为了简化表示，我们在输入特征中加入x0 = 1，于是得到：

参数θ和输入特征x都为矢量，n是输入的特征x的个数（不包含x0）。

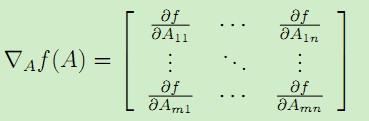
使得预测值h(x)尽可能接近y，为了达到这个目的，我们对于每一个参数θ，定义一个**代价函数cost function**用来描述h(x(i))'与对应的y(i)'的接近程度：

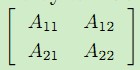
于是我们的目标就变为了调整θ使得代价函数J(θ)取得最小值，方法有梯度下降法，最小二乘法等。

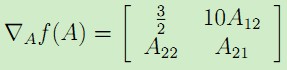
### 2.2.2 最小二乘法

梯度下降算法给出了一种计算θ的方法，但是需要迭代的过程，比较费时而且不太直观。下面介绍的最小二乘法是一种直观的直接利用矩阵运算可以得到θ值的算法。为了理解最小二乘法，首先回顾一下矩阵的有关运算：

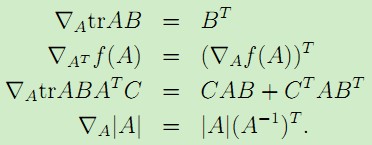
　　假设函数f是将m\*n维矩阵映射为一个实数的运算，即https://images0.cnblogs.com/blog/392228/201410/302115238007040.jpg，并且定义对于矩阵A，映射f(A)对A的梯度为：



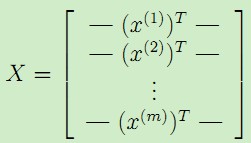
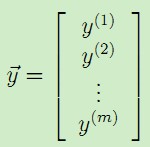
因此该梯度为m\*n的矩阵。例如对于矩阵A=，而且映射函数f(A)定义为：F(A) = 1.5A11 + 5A122 + A21A22，于是梯度为：

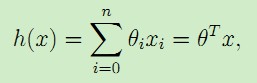
。

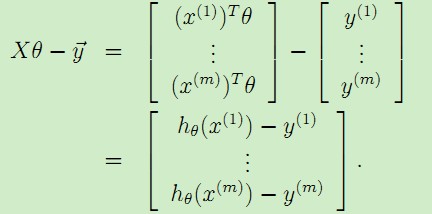
　　另外，对于矩阵的迹的梯度运算，有如下规则：

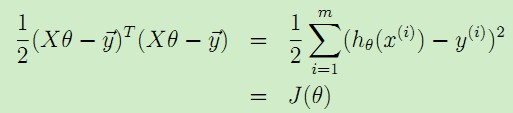
。

　　下面，我们将测试集中的输入特征x和对应的结果y表示成矩阵或者向量的形式，有：

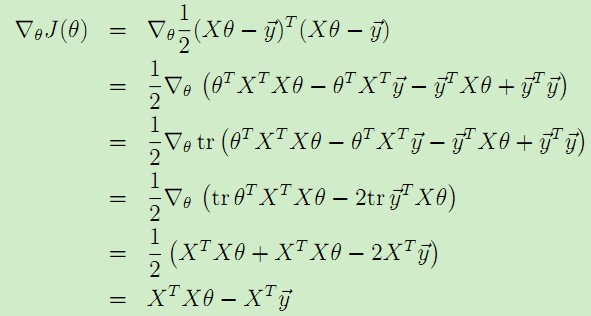
，，

对于预测模型有，即https://images0.cnblogs.com/blog/392228/201410/302126335346998.jpg，于是可以很容易得到：

，

所以可以得到。

　　于是，我们就将代价函数J(θ)表示为了矩阵的形式，就可以用上述提到的矩阵运算来得到梯度：

，

令上述梯度为0，得到等式：https://images0.cnblogs.com/blog/392228/201410/302131411597784.jpg，于是得到θ的值：

https://images0.cnblogs.com/blog/392228/201410/302132163156945.jpg。这就是最小二乘法得到的假设模型中参数的值。

## 2.3 算法实现

**2.3.1 Python语言简介**

Python语言诞生于20世纪90年代初，它已被逐渐广泛应用于[Web](http://baike.baidu.com/subview/3912/15992867.htm)编程和系统管理任务的处理等领域[22]。Python语言是一个高层次的结合了编译性、解释性、互动性和面向对象的脚本编程语言。Python语言具有诸多特点：易于学习、数据库GUI编程、易于阅读、广泛的标准库、易于维护、可扩展性、交互式、可移植、可嵌入等优势。Python语言凭借着自身的诸多优势，广泛应用于系统编程、数据挖掘、图形处理、数学处理、黑客编程、文本处理、网络编程、数据库编程、Web编程和多媒体应用等领域或方向。

Python语言区别于其他的编程语言还有一个特点：具有丰富和强大的库，并能够把用其他语言制作的各种模块联结起来。最常见的一种应用情形：将Python快速生成原型程序，然后将其中有特别配置的部分，用更合适的语言编写。比如[3D游戏](http://baike.baidu.com/view/96860.htm)中的图形渲染的模块对于性能要求特别高，这时候就可以用C++重写，然后再将其封装为Python语言可以调用的扩展类库。

除了自己编写程序模块之外，众多开源的科学计算软件包也提供了Python的调用[接口](http://baike.baidu.com/view/159864.htm)。比如著名的计算机视觉库[OpenCV](http://baike.baidu.com/view/1343775.htm)、医学图像处理库ITK和三维可视化库VTK。而Python专用的科学计算扩展库也有很多。比如经典的科学计算扩展库：SciPy、NumPy和Matplotlib，它们分别为Python提供了数值运算、快速数组处理以及绘图功能。本论文使用这三个标准库，其中更主要的使用了面向机器学习的Scikit-learn标准模块库。

**2.3.2 Scikit-learn**

Scikit-learn是一个开源的基于Python语言的科学计算工具包。它基于SciPy，针对不同的应用领域已经发展了很多的分支版本，它们统一称为Scikits，就是SciPy工具包的意思。在这些分支版本中，最著名的也是专门面向机器学习的一个就是Scikit-learn。Scikit-learn最早由数据科学家David Cournapeau在2007年提出，它依赖于NumPy和SciPy等程序包的支持，是Python语言中面向机器学习应用开发的一款开源框架[24]。

Scikit-learn的基本功能分为六大部分：分类、回归、聚类、数据降维、模型选择和数据预处理。

（1）分类是辨别对象所属的类别，是监督学习的一种。最常见的应用包括垃圾邮件识别和图像识别等应用。目前Scikit-learn所包含的算法有：支持向量机（SVM）、最近邻、随机森林、逻辑回归、决策树以及[神经网络](http://www.leiphone.com/news/201505/t3T1XQy2g3spCUdd.html)等等。另外由于Scikit-learn本身不支持[深度学习](http://www.leiphone.com/news/201701/LqwiP7VUJO9DgBPi.html)，也不支持GPU加速。

（2）回归是指预测给定对象相关联的连续值的问题，其中最常见的应用场景包括预测药物反应和预测股票价格等问题。目前Scikit-learn已经包含的算法有：支持向量回归（S[VR](http://www.leiphone.com/category/arvr)）、弹性网络（Elastic Net）、脊回归（Ridge）、Lasso回归、最小角回归（LARS）和贝叶斯等回归算法。

（3）聚类则是指自动识别具有相似属性的对象的问题。聚类属于无监督学习的典型，最常见的应用场景主要包括顾客类型细分和试验结果分组等问题。目前Scikit-learn已经包含的算法有：K-均值聚类、均值偏移、谱聚类、分层聚类和DBSCAN聚类等算法。

（4）数据降维指的是用主成分分析（PCA）、特征选择和非负矩阵分解（NMF）等降维算法来减少需要考虑的随机变量个数，其主要应用场景包括效率提升和可视化处理等问题。

（5）模型选择的主要内容是对于给定的模型和参数的比较、验证和选择。模型选择的主要目的是通过参数的调整来提升精度。目前Scikit-learn所包含的模块有：交叉验证、格点搜索和各种预测误差评估的度量函数。

（6）数据预处理的主要内容是数据的归一化和特征提取，这是所有的机器学习过程中的首要也是最重要的一个环节。归一化是将输入数据转换为具有零均值和单位权方差的变量。虽然大多数时候都做不到精确等于零，但是往往会设置一个可接受的范围，一般都要求落在0～1之间。特征提取所做的是将图像数据和文本数据转换为可用于机器学习的数值变量。

总结以上六点来说，Scikit-learn实现了完整用于数据降维、模型选择、特征提取和归一化的完整算法模块，同时还提供了每个算法和模块以及其丰富的参考样例和详细的说明文档。

本文中使用了Scikit-learn的回归、模型选择和数据预处理等多个模块。