SKLearn与机器学习

刘军

2017-07-07

南京财经大学

目录

[SKLearn与机器学习 1](#_Toc487181072)

[第一章 什么是机器学习 4](#_Toc487181073)

[第二章 python基础知识 5](#_Toc487181074)

[一、 安装 5](#_Toc487181075)

[二、 常用语法 5](#_Toc487181076)

[1） 数据类型 5](#_Toc487181077)

[2） 切片 5](#_Toc487181078)

[3） 循环 5](#_Toc487181079)

[4） 推导式 5](#_Toc487181080)

[三、 几个程序包 5](#_Toc487181081)

[1） Numpy 5](#_Toc487181082)

[2） Scipy 5](#_Toc487181083)

[3） Matplotlib 5](#_Toc487181084)

[第三章 数学基础知识 6](#_Toc487181085)

[一、 导数与优化 6](#_Toc487181086)

[二、 矩阵运算 6](#_Toc487181087)

[三、 优化问题 6](#_Toc487181088)

[第四章 特征提取与选择 7](#_Toc487181089)

[第五章 降维 8](#_Toc487181090)

[第六章 模型选择 9](#_Toc487181091)

[一、 模型的评价指标 9](#_Toc487181092)

[1） 损失和顺势函数 9](#_Toc487181093)

[2） ROC曲线 9](#_Toc487181094)

[二、 模型选择 9](#_Toc487181095)

[第七章 回归问题 10](#_Toc487181096)

[一、 线性回归 10](#_Toc487181097)

[1） 最小二乘 10](#_Toc487181098)

[2） 岭回归 10](#_Toc487181099)

[3） Lesso回归 10](#_Toc487181100)

[4） 实际案例 10](#_Toc487181101)

[二、 非线性回归 10](#_Toc487181102)

[第八章 分类问题 11](#_Toc487181103)

[一、 SVM 11](#_Toc487181104)

[二、 神经网络 11](#_Toc487181105)

[三、 决策树 11](#_Toc487181106)

[四、 随机森林 11](#_Toc487181107)

[五、 Adaboost 11](#_Toc487181108)

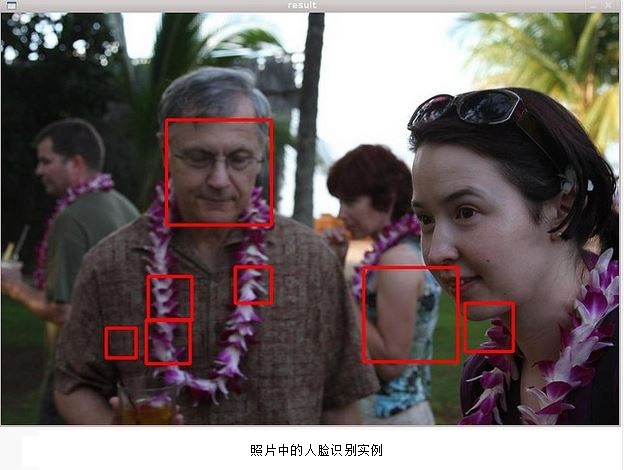
[第九章 聚类问题 12](#_Toc487181109)

[一、 Knn 12](#_Toc487181110)

# 什么是机器学习

<http://blog.csdn.net/lxf_44944/article/details/68490377>

在介绍[**机器学习**](http://lib.csdn.net/base/2)之前，我想先列几个关于机器学习的例子：

* 垃圾邮件检测：根据邮箱中的邮件，识别哪些是垃圾邮件，哪些不是。这样的模型，可以程序帮助归类垃圾邮件和非垃圾邮件。这个例子，我们应该都不陌生。
* 信用卡欺诈检测：根据用户一个月内的信用卡交易，识别哪些交易是该用户操作的，哪些不是。这样的决策模型，可以帮助程序退还那些欺诈交易。
* 数字识别：根据信封上手写的邮编，识别出每一个手写字符所代表的数字。这样的模型，可以帮助程序阅读和理解手写邮编，并根据地利位置分类信件。
* 语音识别：从一个用户的话语，确定用户提出的具体要求。这样的模型，可以帮助程序能够并尝试自动填充用户需求。带有Siri系统的iPhone就有这种功能。
* 人脸识别：根据相册中的众多数码照片，识别出那些包含某一个人的照片。这样的决策模型，可以帮助程序根据人脸管理照片。某些相机或软件，如iPhoto，就有这种功能。  
  
* 产品推荐：根据一个用户的购物记录和冗长的收藏清单，识别出这其中哪些是该用户真正感兴趣，并且愿意购买的产品。这样的决策模型，可以帮助程序为客户提供建议并鼓励产品消费。登录Facebook或GooglePlus，它们就会推荐可能有关联的用户给你。
* 医学分析：根据病人的症状和一个匿名的病人资料数据库，预测该病人可能患了什么病。这样的决策模型，可以程序为专业医疗人士提供支持。
* 股票交易：根据一支股票现有的和以往的价格波动，判断这支股票是该建仓、持仓还是减仓。这样的决策模型，可以帮助程序为金融分析提供支持。
* 客户细分：根据用户在试用期的的行为模式和所有用户过去的行为，识别出哪些用户会转变成该产品的付款用户，哪些不会。这样的决策模型，可以帮助程序进行用户干预，以说服用户早些付款使用或更好的参与产品试用。
* 形状鉴定：根据用户在触摸屏幕上的手绘和一个已知的形状资料库，判断用户想描绘的形状。这样的决策模型，可以帮助程序显示该形状的理想版本，以绘制清晰的图像。iPhone应用Instaviz就能做到这样。[1]

**机器学习(Machine Learning, ML)**是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、[**算法**](http://lib.csdn.net/base/31)复杂度理论等多门学科。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的性能。它是[**人工智能**](http://lib.csdn.net/base/2)的核心，主要使用**归纳、综合而不是演绎**[2]

关于机器学习的定义有很多，一种经常被引用的英文定义是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.[3]

上面的这个定义如何理解呢？想象一下有一个邮箱过滤程序接受到一封邮件，该程序如何判断这封邮件是不是垃圾邮件呢？首先这封邮件先观察你之前标记过是不是垃圾邮件的邮件，基于这个学习到的经验，该过滤程序为你过滤掉垃圾邮件了。在这个任务中：

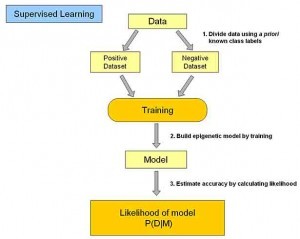
* 判断邮件是不是垃圾邮件,这个**任务T**
* 观察到你之前标记过是不是垃圾邮件的邮件，这个是**经验E**
* 正确分类垃圾邮件与非垃圾邮件的数量，这个是**性能P**



# 机器学习分类

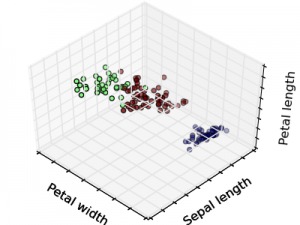
从最广义的讲，任何涉及机器学习所用的方法，只要它利用了训练样本的信息，都可以认为运用了学习。实践中和有意义的机器学习都是如此困难，以至于根本无法实现猜测出一个最佳的分类判决。因此大部分的时间都运用与研究学习问题。实现一个机器学习算法设计：给定一般的模型或分类器的形式，利用训练样本去学习或评估模型的未知参数。这里的学习指用某种算法来降低训练样本的分类误差。一大类基于梯度下降的算法，能够调节分类器的参数，使它朝着能够降低误差的方向前进，目前已成为机器学习领域的主流学习算法。[4]  
在一个机器学习的应用中，根据数据类型的不同，对一个问题的建模有不同的方式。在机器学习或者人工智能领域，人们首先会考虑算法的学习方式。在机器学习领域，有几种主要的学习方式。将算法按照学习方式分类是一个不错的想法，这样可以让人们在建模和算法选择的时候考虑能根据输入数据来选择最合适的算法来获得最好的结果。[5]

### 监督式学习



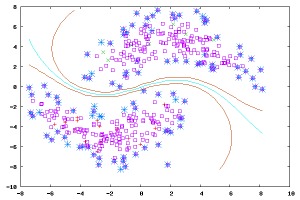
在监督式学习下，输入数据被称为“训练数据”，每组训练数据有一个明确的标识或结果，如对防垃圾邮件系统中“垃圾邮件”“非垃圾邮件”，对手写数字识别中的“1“，”2“，”3“，”4“等。在建立预测模型的时候，监督式学习建立一个学习过程，将预测结果与“训练数据”的实际结果进行比较，不断的调整预测模型，直到模型的预测结果达到一个预期的准确率。监督式学习的常见应用场景如分类问题和回归问题。常见的算法有：逻辑回归，感知机，SVM支持向量机。

### 非监督学习



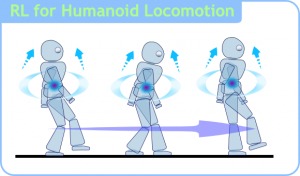
在非监督式学习中，数据并不被特别标识，学习模型是为了推断出数据的一些内在结构。常见的应用场景包括关联规则的学习以及聚类等。常见算法包括：Apriori算法以及k-Means算法。

### 半监督学习



在此学习方式下，输入数据部分被标识，部分没有被标识，这种学习模型可以用来进行预测，但是模型首先需要学习数据的内在结构以便合理的组织数据来进行预测。应用场景包括分类和回归，算法包括一些对常用监督式学习算法的延伸，这些算法首先试图对未标识数据进行建模，在此基础上再对标识的数据进行预测。如图论推理算法（Graph Inference）或者拉普拉斯支持向量机（Laplacian SVM.）等。

### 强化学习



在这种学习模式下，输入数据作为对模型的反馈，不像监督模型那样，输入数据仅仅是作为一个检查模型对错的方式，在强化学习下，输入数据直接反馈到模型，模型必须对此立刻作出调整。常见的应用场景包括动态系统以及机器人控制等。常见算法包括Q-Learning以及时间差学习（Temporal difference learning）

### 机器学习框架

在机器学习领域有很多的框架、库以及软件。比较出名的有：C++的opencv, shark等，java中weka、hadoop的mahout、spark的mllib，matlab中的LibSVM、caffe，python的scikit-learn、NLTK、theano、brain、caffe等。

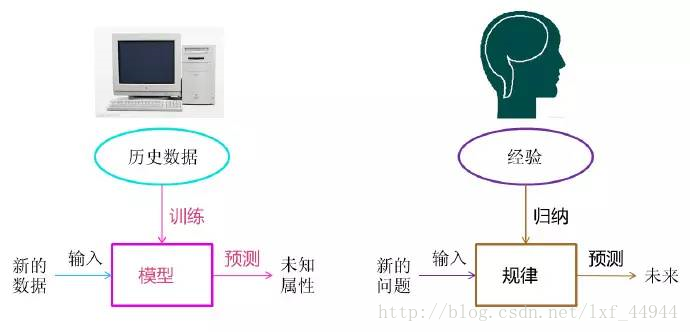
今天我就根据个人的理解带着大家一起来看看机器学习是什么，而机器又是怎样学习的。

我从小就对神奇的事物感兴趣，有一次看到一位魔术师准确地猜出了路人心中的数字时，便着迷了很长一段时间。在我学习魔术不久之后，便发现很大一部分预测类魔术都会利用到大众心理学以及统计学的相关知识，而往往这类魔术的表演成功率也并非百分之百。



我们来看一个简化版的读心术，有兴趣的读者可以跟着选一下：“现在请你从1到4中选一个数字出来，然后我来猜，选好了么？我猜你选的应该是3！”很多人也许会大吃一惊，觉得我很懂你的心，其实真相并非如此。有心理学家专门做过这个实验：随机挑选出一部分人，让他们参与测验，发现针对这一问题选3的人最多，如果我猜你选3，猜对的概率最大，这便是上面这个读心术的诀窍（其实这个魔术也同时利用了人的心理，在问题中提及1和4就会大大降低选择这两个数字的概率）。

**机器学习的原理也是如此，我们通常会利用已有的测试答案总结出规律，然后通过这个规律再去预测一个陌生人的答案，这里的测试答案便是我们的“样本”，而总结规律的过程就是“训练模型”，总结出的规律我们当做“训练好的模型”继续预测更多人的答案。**

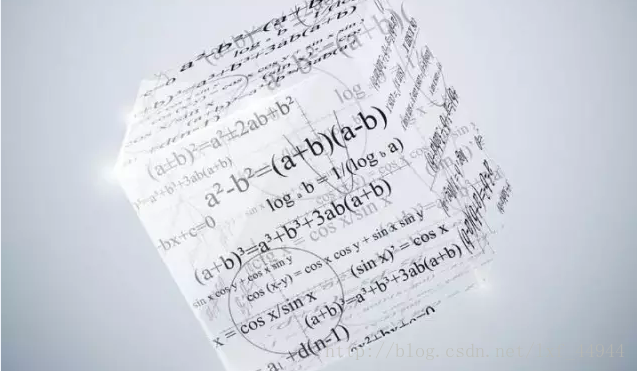


通过以上描述我们可以看出，机器学习和我们传统的计算机工作流程有很大区别。过去，我们输入指令，让**计算机按照我们的指令执行出结果，更强调因果**，越是资深的编程人员，对这一点越是深信不疑，时刻会强调逻辑关系。而机器学习是**利用数据得到结果，从数据中“学”到规律**，而不是依靠指令，于是机器学习更多考虑的是相关性，而非因果。其实这和我们的生活智慧很相似，生活中处理很多事情上没有绝对的对错，只有相对的合理与否，如果我们可以理解透彻这一点，不管是处理机器学习还是生活中的问题时，都会有很大帮助。

我们的祖先早已对机器学习的原理有所领悟，唐太宗李世民曾说：“以史为镜，可以知兴替”，掌握了历史兴替的规律，便可以指导未来的发展方向，避免重蹈覆辙。再推前至上古时代，伏羲倚仗天地变化之术，创建了先天八卦，这不正是最强大的模型么，它归纳了万物规律，通过八卦模型便可预测未知结果，让人不得不惊叹它的神奇（有机会我们可以专门探讨一下关于易经与计算机科学千丝万缕的联系，非常有趣）。



这时，有的同学可能会发现模型被提到了很多次，但大多是抽象的描述，那么究竟什么是模型？回答这个问题前我们不妨先来看一道小学数学题：“一艘船离开码头1公里后开始匀速前行，1小时后船离码头10公里，请问2小时后船离码头几公里？”将这个问题转化机器学习的描述就变成：我们希望通过两个样本(0,1)和(1,10)训练出一个模型，然后输入2，预测出离码头的距离。通过题目中的业务场景，我们可以分析出样本是程线性关系的，于是我们可以考虑选用一个简单的线性模型“Y=aX+b”来表示这种情况。接下来我们将样本(0,1)和(1,10)代入模型训练，求得参数a=9，b=1，于是训练好的模型为Y=9X+1（如果还有额外的样本点，我们可以代入检查一下这个模型的准确性），此时我们将想要预测的X=2代入求得Y=19，这样我们便成功预测出2小时后船离码头19公里，今后不管是希望获得几小时的距离，我们都可以通过这个训练好的模型进行预测了。



总结一下整个过程：**我们首先会通过样本的规律寻找合适的模型，再用样本数据训练模型（训练时通常会将样本分为两部分，一部分用来训练，另一部分用来检验训练后模型的正确率，以评估模型的好坏），之后我们就可以通过训练好的模型进行新数据的预测了。**

虽然真实的业务并不会这么简单，样本也不会如此规整，但我希望可以通过这样一个理想状态下的小例子让大家更直观地了解什么是“模型”、“训练”以及“预测”，我们后面将要学习的很多[**算法**](http://lib.csdn.net/base/31)模型的大多都是同样的训练过程。

**现在，相信大家对机器学习已经有了更清晰的认识，其实机器学习学到的东西就是我们所谓的经验或者规律，机器学习会把这些学到的东西固化到模型中，以便解决新的问题。**

经过无数研究者的努力，我们已经拥有非常丰富的模型库，而当前多样的技术也将这些模型的训练与使用变得非常容易。即便如此，很多人还是会担心做机器学习的数学门槛会不会很高，实际上不是所有的模型都用到了很高深的数学知识，很多时候我们知道这个模型算法的大概思路以及它的应用场景就可以满足我们的需求了，而走这样的应用路线同样会有很广阔的前景和舞台（当然如果想深入研究这个领域，数学越好会让你走得越远）。

# python基础知识

## 安装

## 常用语法

### 数据类型

### 切片

### 循环

### 推导式

## 几个程序包

### Numpy

### Scipy

### Matplotlib

## 集成开发环境

## 如何获取帮助

# 数学基础知识

## 导数与优化

## 矩阵运算

矩阵初步

1. 矩阵的基本概念：

（1）实数、向量、矩阵、张量

（2）矩阵和向量的关系

（3）矩阵和方程组的关系

（4）特殊矩阵：单位矩阵，数量矩阵，对角矩阵，三角矩阵

2. 矩阵的基本运算：

（1）矩阵的加、减、乘（数乘和矩阵乘法）、除，转置（对称矩阵），求行列式

（2）方程组的矩阵表示

（3）矩阵三个初等变换

（4）矩阵的逆、秩、迹

矩阵变换和线性空间

1. 线性空间

（1）线性的概念、线性相关、线性表示（线性组合）和方程组的关系

（2）线性空间、线性子空间（生成子空间）、向量组的值域和核

（3）线性空间的范数、范数（p=1、2、）、矩阵的Frobenius范数

2. 矩阵分解

（1）特征分解、正交分解、正定（半正定）矩阵、正定矩阵和特征分解之间的关系、SVD分解（Moore-Penrose伪逆）

（2）应用举例：PCA

## 优化问题

凸优化

数值优化是机器学习的理论基础与核心内容，在各种机器学习算法中普通存在。优化方法通过数学公式将目标函数定量表示，便于机器学习灵活地嵌入计算机视觉、自然语言处理等领域知识。本课程介绍机器学习及优化相关的数学基础与思维方式，将数值优化技术定制到特定的机器学习问题中，探讨凸优化（如支持向量机）与非凸优化（如深度学习）算法中的性能与精度等问题。

# 特征提取与选择

# 降维

PCA

HPCA

# 模型选择

## 模型的评价指标

### 损失和顺势函数

### ROC曲线

<http://blog.csdn.net/jzwong/article/details/62217526>

http://blog.csdn.net/pzy20062141/article/details/48711355

https://www.douban.com/note/284051363/

ROC曲线首先是由二战中的电子工程师和雷达工程师发明的，用来侦测战场上的敌军载具（飞机、船舰），也就是信号检测理论。之后很快就被引入了[心理学](https://zh.wikipedia.org/wiki/%E5%BF%83%E7%90%86%E5%AD%A6" \t "_blank" \o "心理学)来进行信号的[知觉](https://zh.wikipedia.org/wiki/%E7%9F%A5%E8%A7%89)检测。数十年来，ROC分析被用于[医学](https://zh.wikipedia.org/wiki/%E5%8C%BB%E5%AD%A6" \t "_blank" \o "医学)、[无线电](https://zh.wikipedia.org/wiki/%E6%97%A0%E7%BA%BF%E7%94%B5)、[生物学](https://zh.wikipedia.org/wiki/%E7%94%9F%E7%89%A9%E5%AD%B8)、[犯罪心理学](https://zh.wikipedia.org/wiki/%E7%8A%AF%E7%BD%AA%E5%BF%83%E7%90%86%E5%AD%B8)领域中，而且最近在[机器学习](https://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0)（machine learning）和[数据挖掘](https://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \o "数据挖掘" \t "_blank)（data mining）领域也得到了很好的发展。

## 模型选择

# 回归问题

## 线性回归

### 最小二乘

### 岭回归

### Lesso回归

### 实际案例

## 非线性回归

# 分类问题

## SVM

## 神经网络

## 决策树

## 随机森林

## Adaboost

# 聚类问题

## Knn