# 第一章：赋予计算机从数据学习的能力

在我看来，*机器学习，*与数据相关的算法及其应用，是所有计算机科学中最激动人心的一个领域！我们生活在一个数据充沛的时代，通过使用机器学习领域的自主学习算法，我们可以把这些数据转化为知识。得益于近些年来开发的许多强大的开源库，可能再没有比现在更适合于投入到机器学习领域，来学习如何利用强大的算法去发现数据的规律并且对未来的事件进行预测。

在本章中，我们会学习机器学习的主要概念以及它的不同类型。在对其基本术语进行介绍的同时，我们还会为使用机器学习进行实际问题的解决做好铺垫。

本章将会覆盖如下知识点：

1、机器学习的基本概念

2、机器学习的三种类型及其基本术语

3、成功设计机器学习系统的基石

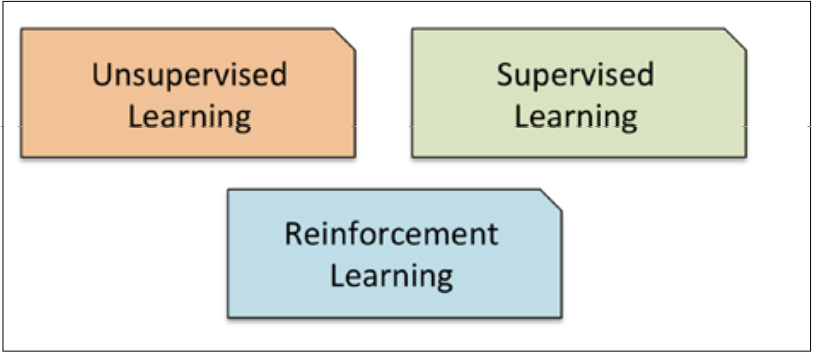
4、安装以及搭建Python环境用于数据分析以及机器学习

## 创建智能机器来把数据转化为知识

在当今这个科技时代，有一种资源我们异常丰富：大量的有结构的和无结构的数据。在20世纪下半叶，机器学习作为*人工智能*的子领域被提出，它涉及到自主学习算法的发展来从数据获取信息以便进行预测。不同于需要人工从大量数据中导出规则然后进行建模，机器学习提供了一种更为高效的方法来挖掘数据中的知识来逐渐改善预测模型的性能，并且做出基于数据的决策。机器学习不仅在计算机科学研究中扮演着越来越重要的角色，在我们的日常生活中也随处可见其身影。得益于机器学习，我们享受到了强大的垃圾邮件过滤、方便的文字以及语音识别软件、可靠的web搜索引擎、具有挑战的棋类AI，可能在不久的将来，还可以体验到安全且高效的自动驾驶汽车。

## 机器学习的三种不同类型

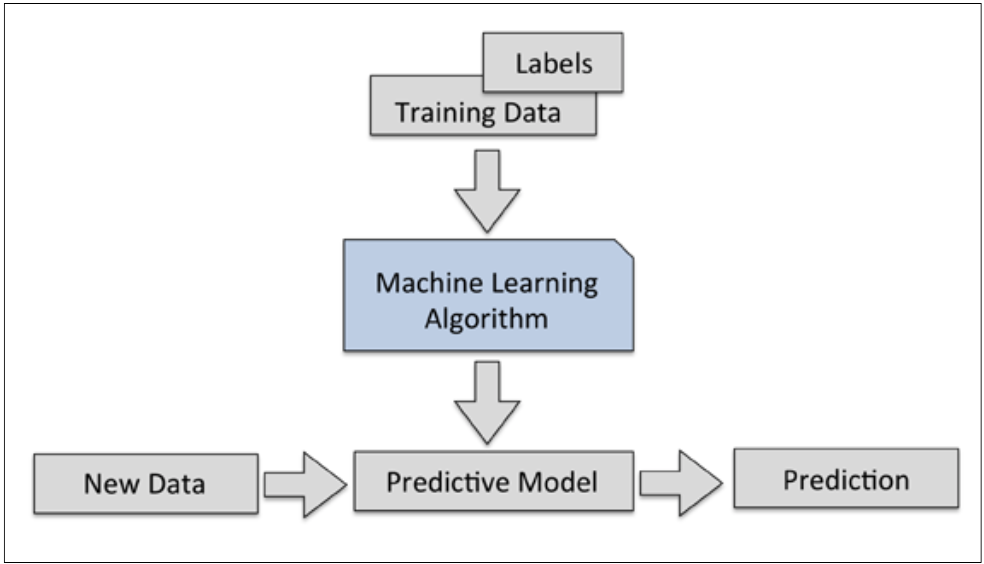
在本节中，我们来看一下机器学习的三种类型：*监督学习*、*无监督学习*以及*增强学习*。我们会学习这三者之间基本的区别，并且通过概念性的例子，我们会形成它们能够用在哪些实际问题解决领域的直觉：



## 使用监督学习对未来进行预测

监督学习的主要目标是从标记训练数据中学到一个模型，使得我们可以对未知以及未来的数据进行预测。在这里，*监督*一词指的是一系列**已经知道预期结果**的样本。

我们来看一下邮件过滤的例子，我们可以在一个标记了的e-mail集上使用一个监督机器学习算法来训练一个模型，e-mail被正确的打上垃圾或者非垃圾的标签，来预测一个新的e-mail是属于这两类中的哪个。具有离散*类标签*的监督学习任务，比如上面提到的垃圾邮件过滤的例子，也被称为*分类（classification）*任务。监督学习的另一个子领域是*回归（regression）*，它的输出信号是一个连续的值：

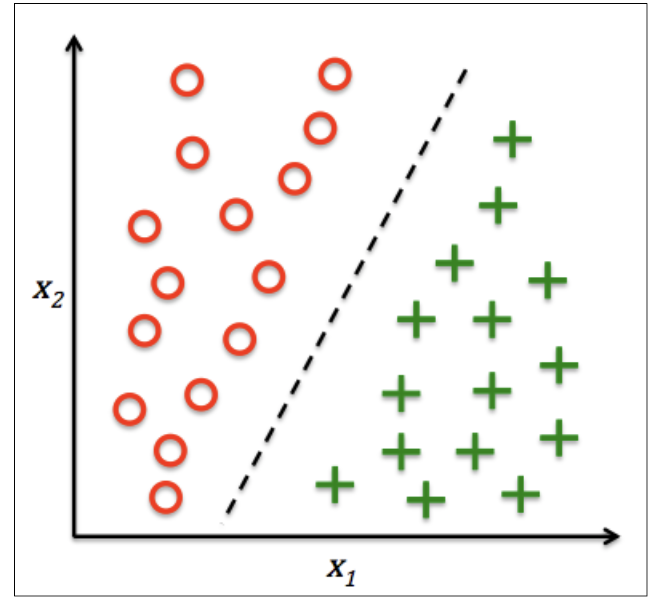


## 预测类标签的分类

分类（classification）是监督学习的子领域，其目标是通过过去的观测来预测新实例的类型标签。那些类型标签是离散的、无序的值，可以把它们理解为这些实例的*组成员关系*。前面提到的垃圾邮件筛选例子可以表示一个典型的*二元分类*任务，机器学习算法学习一系列规则，来将实例归为这两类：垃圾或者非垃圾邮件。

但是，类型标签并不一定就是二元的。通过监督学习算法学到的预测模型可以将任何训练数据集中表示的类型标签指定给一个新的、未打上标签的实例。一个*多类分类*任务的典型例子是手写字符识别。在这个例子中，我们可以收集一个训练集，它包含字母表上每个字母的多个手写例子。接着，如果用户通过输入设备提供一个新的手写字符，我们的预测模型就能够相当精确的预测该手写字符对应字母表上哪个字符。不过，我们的机器学习系统并不能准确的识别数字0到9，如果它们不存在于我们的训练集中。

下图说明了一个二元分类任务的概念，给定一个含有30个样本的训练集：15个训练样本被标记为*负类*（圆圈），另外15个训练样本被标记为*正类*（加号）。在这个场景中，我们的数据集是二维的，表示每个样本关联着两个值：x1和x2 。现在，我们可以使用一个监督机器学习算法来学习一个规则—决策边界被表示为一个黑色虚线—通过它可以将两种类型进行区分，并且一旦给定其x1和x2的值，可以把一个新的数据归类到这两种类型中的一种。



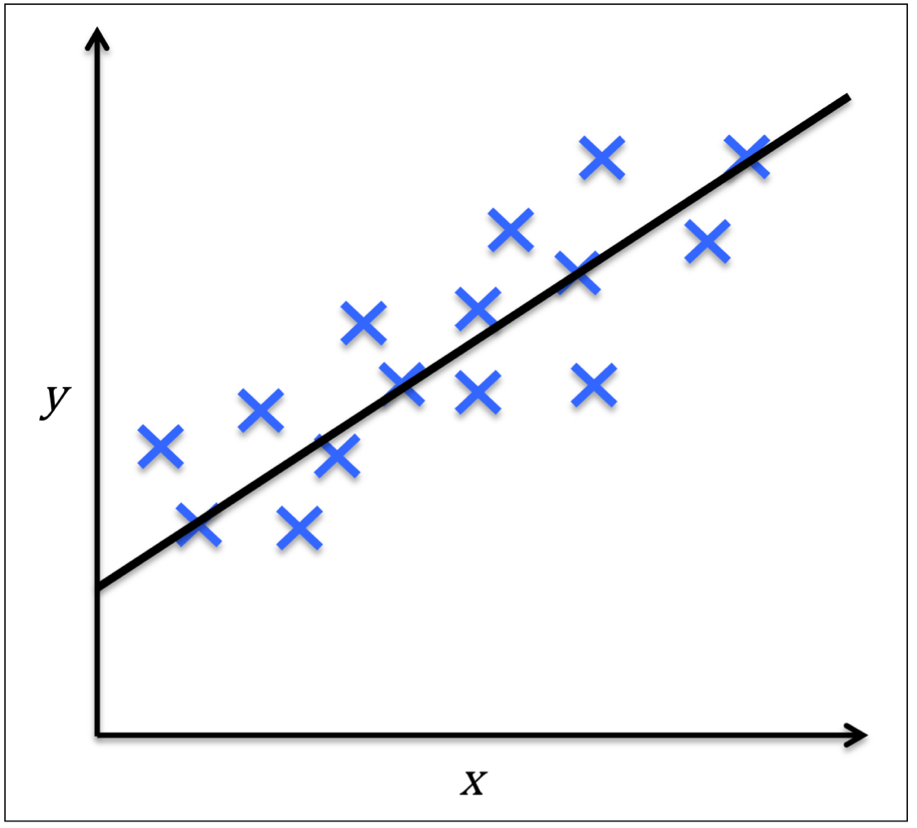
## 预测连续结果的回归（regression）

通过上节的学习，我们知道了分类的任务是将用于分类的、无序的标签指定到实例上。监督学习的第二种类型是对连续结果的预测，也被称为回归分析。在*回归分析*中，我们被给定一些预测变量以及连续的结果变量，我们需要找到这些变量之间的联系以对结果进行预测。

比如说，假设我们对于预测同学们的SAT成绩感兴趣。假设在最终考试成绩和你花在考试上的学习时间之间有某种联系，那我们就可以使用这些数据作为训练集来学习到一个模型，用来预测未来参加这个测试的学生的考试成绩。

*回归*一词由Francis Galton在他1886年发表的论文*在身高遗传上回归趋向平庸（Regression Towards Mediocrity in Hereditary Stature）*中被提出。Galton描述了这样的一个生物现象，物种的身高并不会随着时间不断增加。他观察到，父母的身高并不会遗传给孩子，但是孩子的身高总是回归趋向于物种的平均身高。

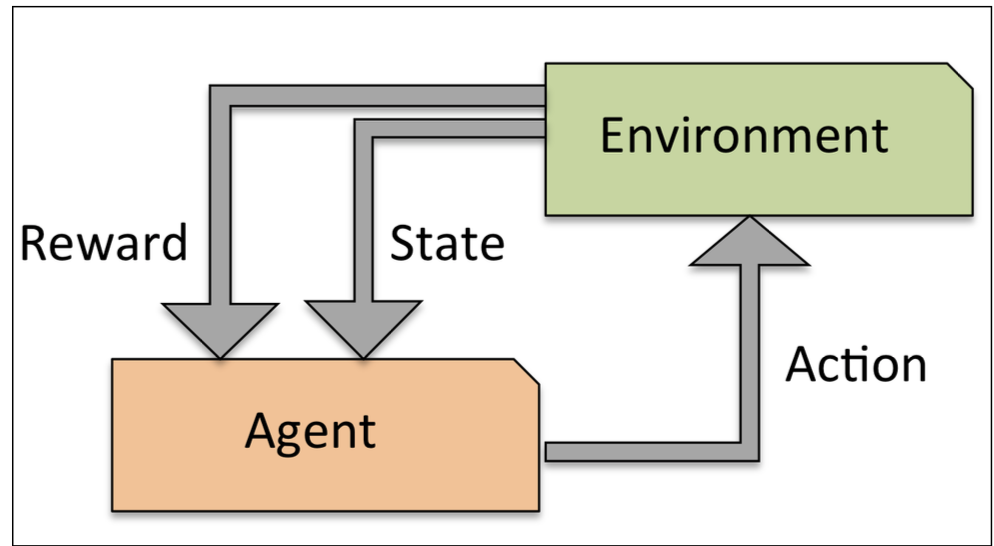
下图说明了*线性回归*的概念。给定一个预测变量x，以及一个结果变量y，我们找出一条直线，使得样本数据与直线之间的距离（最常见的平均平方距离）最短。我们现在就可以用从这个数据集上学习得到的截距和斜率来预测新数据的结果变量。



## 使用强化学习解决交互问题

机器学习的另一种类型是强化学习。在强化学习中，我们的目标是开发一个系统（代理），它能够通过与*环境*的交互来改善自己的性能。由于当前环境状态的信息通常包含一个*奖励*信号，我们可以把强化学习看作与监督学习相关的一个领域。但是，在强化学习中，这种反馈并不是正确的标签或者值，而是通过一种奖励函数对其表现进行评判的度量值。通过和环境进行交互，代理接着可以使用强化学习通过一种探索性试错法或者慎重的规划来学到一系列动作使得获得的奖励最大化。

强化学习最流行的例子就是棋类引擎。在这里，代理基于当前棋盘的状态（环境）来决定下一步的动作，在游戏结束时，奖励可以被定义为*成功*或者*失败*：



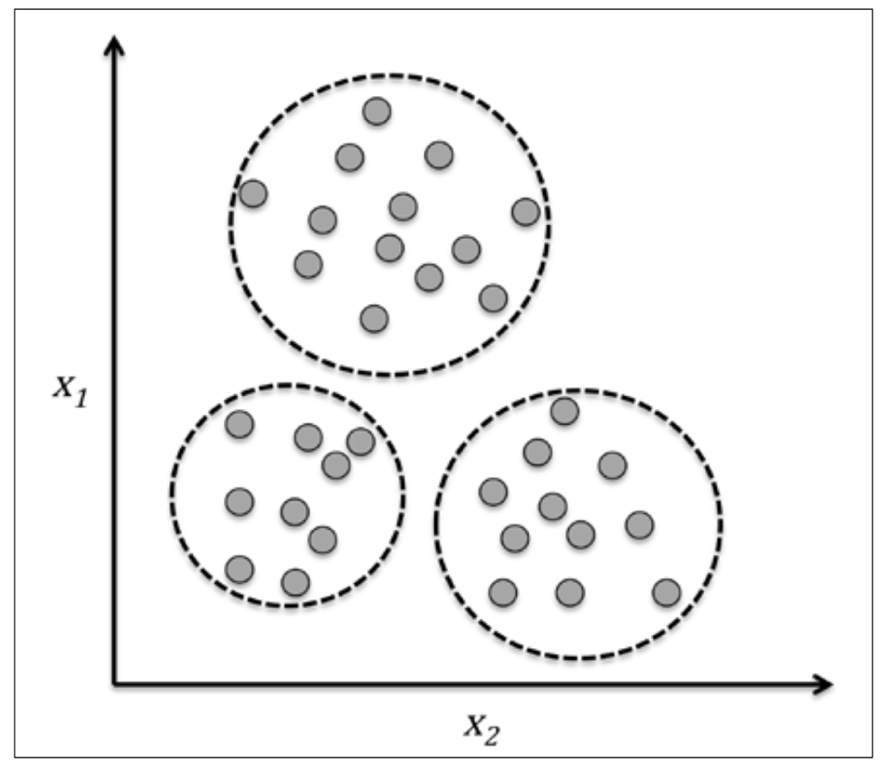
## 使用非监督学习发掘隐含结构

在监督学习中，我们在训练我们的模型之前就知道了*正确答案*，在强化学习中，我们给代理的某些动作定义了一个奖励度量。但在非监督学习中，我们处理的是未标记的数据或者有着*未知结构*的数据。使用非监督学习技术，我们可以探索我们数据的结构来挖掘有意义的信息，而不需要一个已知输出变量或者奖励函数的指导。

## 通过聚集找到子群

*聚集（clustering）*是一种探索性数据分析技术，它允许我们将一堆信息组织成为有意义的子群（集群），而不用事先知道子群的组织成员信息。在分析的过程中，一部分对象就会渐渐展现出有别于其他对象的特征，这些有相似特征的对象便会组成一个个集群，这也是为什么聚集有时也被称为“无监督分类”。聚集很适用于使信息结构化和使无意义的数据形成有意义的关联上，比如说，它使得营销人员可以给予顾客对商品感兴趣程度的不同，来对顾客进行分类，从而有针对性开发不同的营销计划。

下图展示了聚集是如何根据数据的x1和x2特征来将未标记的数据分成三个不同的组：



## 降维以压缩数据

无监督学习的另一个子领域是“降维”，通常我们处理的数据都有很高的维度—每次观察都需要大量测量—这给有限的存储空间以及机器学习算法的计算性能带来了挑战。无监督降维常用于特征预处理，来为数据降噪