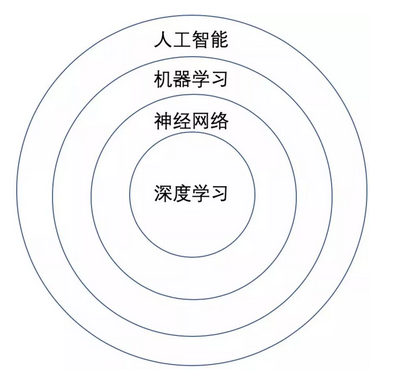
# 概述

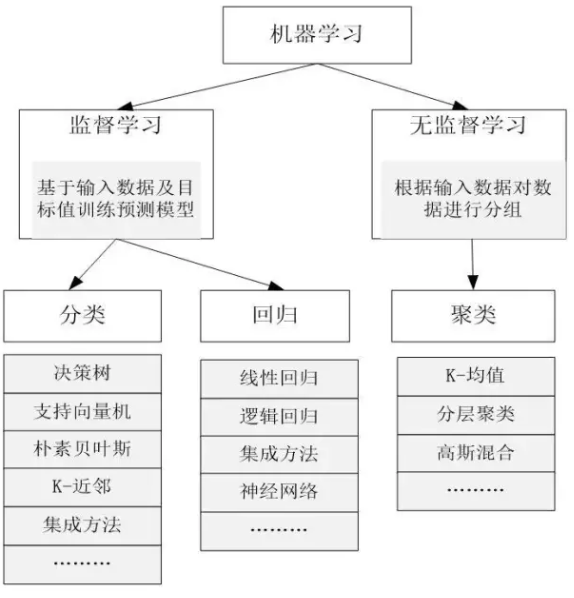
## 人工智能

 人工智能，机器学习和深度学习的关系：

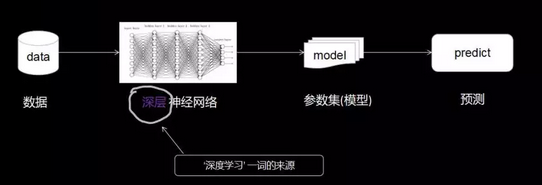
人工智能：能够独立推理思考、有自我意识、有自己的世界观的机器（程序）。

## 机器学习

机器学习：对于某类任务T和性能度量P，如果计算机程序在T上以P衡量的性能随着经验E而自我完善，那么就成这个计算机程序从经验E学习（机器学习）。

 机器学习按照方法主要可以分为两大类：监督学习和无监督学习（还有强化学习）。其中监督学习主要由分类和回归等问题组成，无监督学习主要由聚类和关联分析等问题组成。**深度学习则属于监督学习当中的一种**。

## 深度学习

 深度学习：对于某类任务T和性能度量P，如果计算机程序在T上以P度量的性能随着经验E而自我完善，那么就称这个计算机程序从经验E学习。

## 模式识别

　　模式识别=机器学习。两者的主要区别在于前者是从工业界发展起来的概念，后者则主要源自计算机学科。在著名的《Pattern Recognition And Machine Learning》这本书中，Christopher M. Bishop在开头是这样说的“模式识别源自工业界，而机器学习来自于计算机学科。不过，它们中的活动可以被视为同一个领域的两个方面，同时在过去的10年间，它们都有了长足的发展”。

## 数据挖掘

　　数据挖掘=机器学习+数据库。这几年数据挖掘的概念实在是太耳熟能详。几乎等同于炒作。但凡说数据挖掘都会吹嘘数据挖掘如何如何，例如从数据中挖出金子，以及将废弃的数据转化为价值等等。但是，我尽管可能会挖出金子，但我也可能挖的是“石头”啊。这个说法的意思是，数据挖掘仅仅是一种思考方式，告诉我们应该尝试从数据中挖掘出知识，但不是每个数据都能挖掘出金子的，所以不要神话它。一个系统绝对不会因为上了一个数据挖掘模块就变得无所不能(这是IBM最喜欢吹嘘的)，恰恰相反，一个拥有数据挖掘思维的人员才是关键，而且他还必须对数据有深刻的认识，这样才可能从数据中导出模式指引业务的改善。大部分数据挖掘中的算法是机器学习的算法在数据库中的优化。

## 统计学习

　　统计学习近似等于机器学习。统计学习是个与机器学习高度重叠的学科。因为机器学习中的大多数方法来自统计学，甚至可以认为，统计学的发展促进机器学习的繁荣昌盛。例如著名的支持向量机算法，就是源自统计学科。但是在某种程度上两者是有分别的，这个分别在于：统计学习者重点关注的是统计模型的发展与优化，偏数学，而机器学习者更关注的是能够解决问题，偏实践，因此机器学习研究者会重点研究学习算法在计算机上执行的效率与准确性的提升。

## 计算机视觉

　　计算机视觉=图像处理+机器学习。图像处理技术用于将图像处理为适合进入机器学习模型中的输入，机器学习则负责从图像中识别出相关的模式。计算机视觉相关的应用非常的多，例如百度识图、手写字符识别、车牌识别等等应用。这个领域是应用前景非常火热的，同时也是研究的热门方向。随着机器学习的新领域深度学习的发展，大大促进了计算机图像识别的效果，因此未来计算机视觉界的发展前景不可估量。

## 语音识别

　　语音识别=语音处理+机器学习。语音识别就是音频处理技术与机器学习的结合。语音识别技术一般不会单独使用，一般会结合自然语言处理的相关技术。目前的相关应用有苹果的语音助手siri等。

## 自然语言处理

　　自然语言处理=文本处理+机器学习。自然语言处理技术主要是让机器理解人类的语言的一门领域。在自然语言处理技术中，大量使用了编译原理相关的技术，例如词法分析，语法分析等等，除此之外，在理解这个层面，则使用了语义理解，机器学习等技术。作为唯一由人类自身创造的符号，自然语言处理一直是机器学习界不断研究的方向。按照百度机器学习专家余凯的说法“听与看，说白了就是阿猫和阿狗都会的，而只有语言才是人类独有的”。如何利用机器学习技术进行自然语言的的深度理解，一直是工业和学术界关注的焦点。

　　可以看出机器学习在众多领域的外延和应用。机器学习技术的发展促使了很多智能领域的进步，改善着我们的生活。

# 分类

依据机器学习算法如何学习数据可分为3类：

有监督学习：从有标签的数据学习，得到模型参数，对测试数据正确分类；

无监督学习：没有标签，计算机自己寻找输入数据可能的模型；

强化学习（reinforcement learning）：计算机与动态环境交互，学习错误反馈达到更优的目的。

依据机器学习期望结果来分类：

分类：输入被分为N个类别的一种；

回归：输出是连续值；如依据房子的大小，时间，位置来预测房子的价格；

聚类：使用无监督学习将输入聚为N类；

密度估计（density estimation）：找到输入可能的分布；

机器学习根据任务类型，可以划分为：

**监督学习任务**：从已标记的训练数据来训练模型。 主要分为：分类任务、回归任务、序列标注任务。

**无监督学习任务**：从未标记的训练数据来训练模型。主要分为：聚类任务、降维任务。

**半监督学习任务**：用大量的未标记训练数据和少量的已标记数据来训练模型。

**强化学习任务**：从系统与环境的大量交互知识中训练模型。

机器学习根据算法类型，可以划分为：

**传统统计学习**：基于数学模型的机器学习方法。包括SVM、逻辑回归、决策树等。

这一类算法基于严格的数学推理，具有可解释性强、运行速度快、可应用于小规模数据集的特点。

**深度学习**：基于神经网络的机器学习方法。包括前馈神经网络、卷积神经网络、递归神经网络等。

这一类算法基于神经网络，可解释性较差，强烈依赖于数据集规模。但是这类算法在语音、视觉、自然语言等领域非常成功。

## 监督学习

从已标记的训练数据来训练模型。 主要分为：分类任务、回归任务、序列标注任务。

监督学习的两种主要类型是**分类**和**回归**。

在分类中，训练的机器将把一组数据分成特定的类。比如邮箱的垃圾邮件过滤器，过滤器分析之前标记为垃圾邮件的邮件，并将其与新邮件进行比较。如果达到某个百分比，则这些新邮件会被标记为垃圾邮件，并发送到相应的文件夹；不像垃圾邮件的将被归类为正常并发送到收件箱。

第二种是回归。在回归中，机器使用先前标注的数据来预测未来。比如天气应用。利用天气的相关历史数据（即平均温度，湿度和降水量），手机的天气应用可以查看当前天气，并对一定时间范围内的天气进行预测。

## 无监督学习

从未标记的训练数据来训练模型。主要分为：聚类任务、降维任务。

在无监督学习中，数据是未标注的。由于现实中，大多数的数据都是未标注的，因此这些算法特别有用。

无监督学习分为**聚类**和**降维**。

聚类用于根据属性和行为对象进行分组。这与分类不同，因为这些组不会提供给你。聚类将一个组划分为不同的子组（例如，根据年龄和婚姻状况），然后进行有针对性的营销。

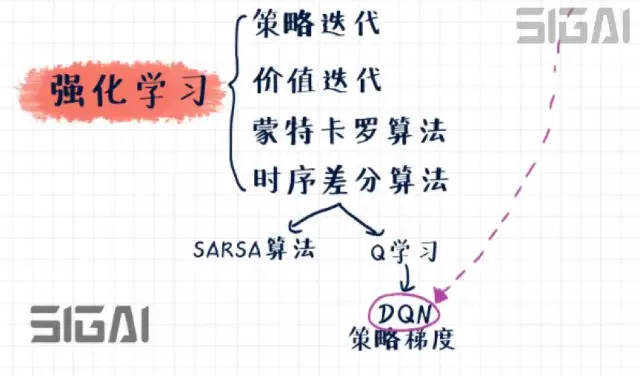
另一方面，降维涉及通过查找共性来减少数据集的变量。大多数数据可视化使用降维来识别趋势和规则。

## 半监督学习

用大量的未标记训练数据和少量的已标记数据来训练模型。

## 强化学习

强化学习使用机器的历史和经验来做出决策。强化学习的经典应用是游戏。

与监督和无监督学习相反，强化学习不注重提供“正确”的答案或输出。相反，它专注于性能，这类似人类根据积极和消极后果进行学习。如果孩子碰到了热炉，他很快就会学会不再重复这个动作。同样在国际象棋中，计算机可以学习不将王移动到对手的棋子可以到达的地方。根据这个原理，在游戏中机器能够最终击败顶级的人类玩家。

# 原理

线性代数概念Top 3：

1. 矩阵运算

2. 特征值/特征向量

3. 向量空间和范数

微积分概念Top 3：

1. 偏导数

2. 向量值函数

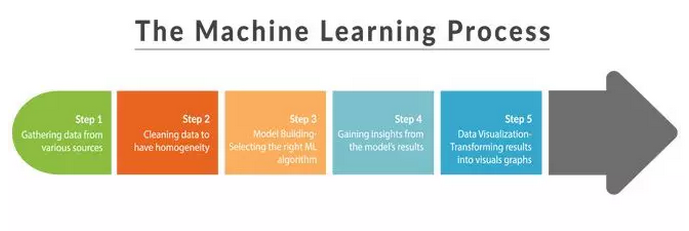
3. 方向梯度

统计概念Top 3：

1. 贝叶斯定理

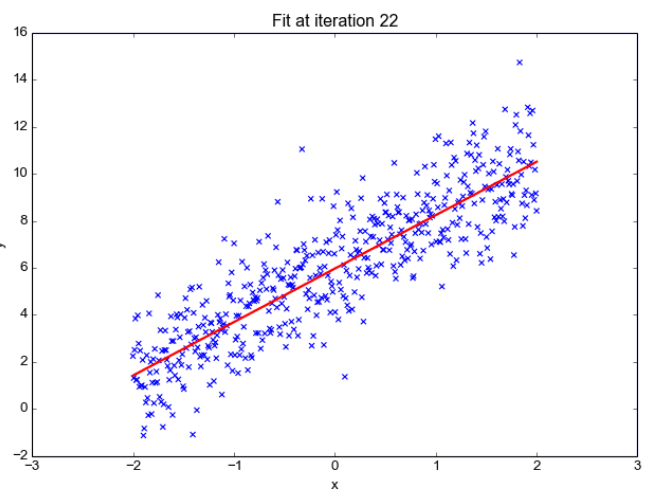
2. 组合学

3. 抽样方法

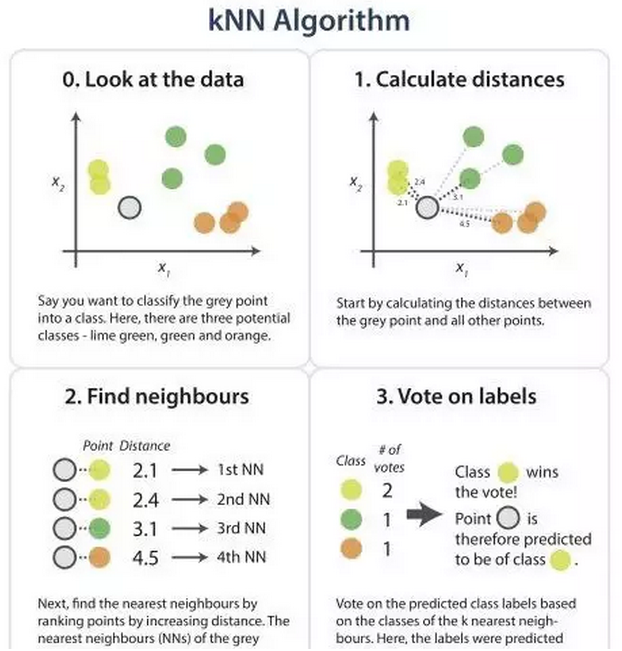
一旦掌握了基本的数学概念，就可以入门机器学习了，有5个主要步骤：

# 算法

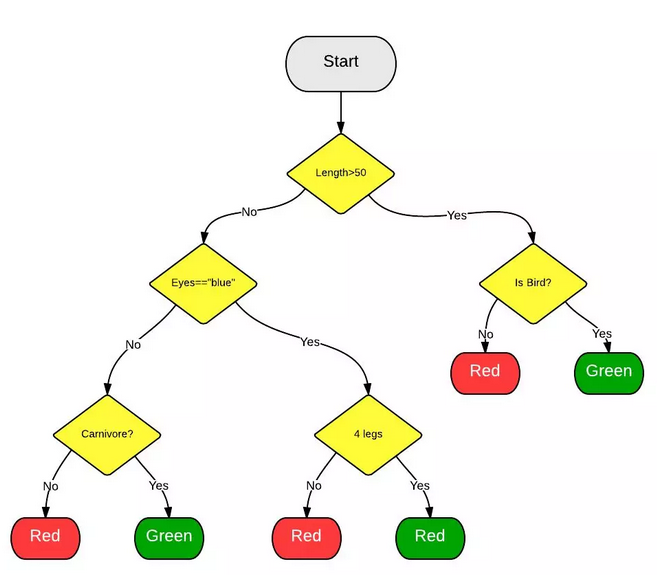
## 回归算法

这可能是最流行的机器学习算法，线性回归算法是基于连续变量预测特定结果的监督学习算法。另一方面，逻辑回归专门用于预测离散值。这些算法都以其速度而闻名，它们被认为是最快的机器学习算法之一。

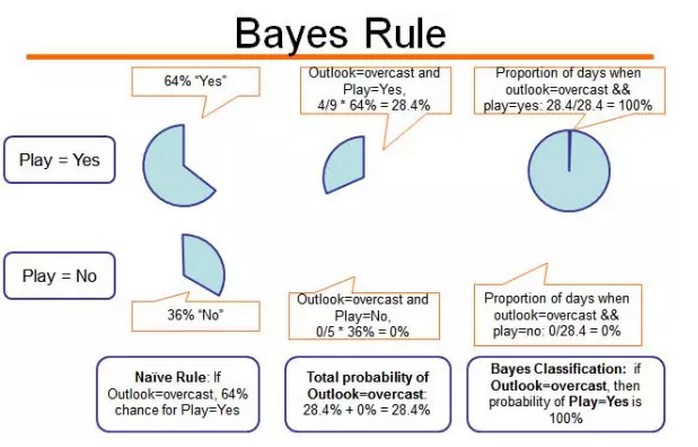
## 基于实例的算法

基于实例的分析根据提供数据的特定实例来预测结果。最著名的基于实例算法是k-Nearest Neighbor，也称为kNN。用于分类中，kNN比较数据点的距离并将每个点分配给它最接近的组。

## 决策树算法

决策树算法聚集“弱”学习元素，让它们一起形成强大的算法，这些元素以树状结构组成。其中比较流行的决策树算法是随机森林算法。在该算法中，弱学习元素是随机选择的。在下面的例子中，我们可以发现许多共同的特征（比如眼睛为蓝色或非蓝色），这都无法对动物种类进行辨别。然而，当我们将所有这些观察结果结合在一起时，我们能够形成更完整的理解并进行更准确的预测。

## 贝叶斯算法

这些算法基于贝叶斯定理的，最受欢迎的是朴素贝叶斯算法，它经常用于文本分析。例如，大多数垃圾邮件过滤器都使用贝叶斯算法。它们使用按类别标记的用户输入数据来比较新数据，并对其进行适当分类。

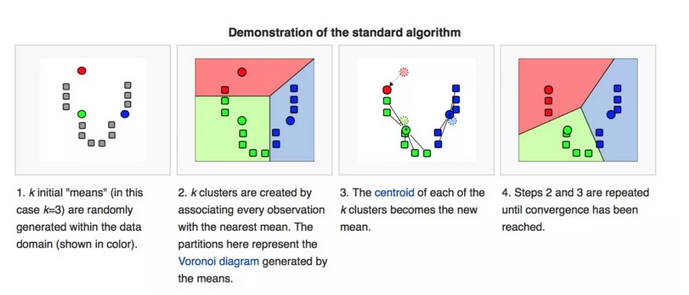
## 降维算法

降维算法也是一种无监督学习算法，其主要特征是将数据从高维降低到低维层次。在这里，维度其实表示的是数据的特征量的大小，例如，房价包含房子的长、宽、面积与房间数量四个特征，也就是维度为4维的数据。可以看出来，长与宽事实上与面积表示的信息重叠了，例如面积=长 × 宽。通过降维算法我们就可以去除冗余信息，将特征减少为面积与房间数量两个特征，即从4维的数据压缩到2维。于是我们将数据从高维降低到低维，不仅利于表示，同时在计算上也能带来加速。

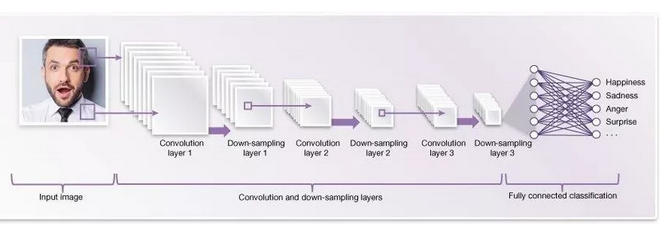
　　刚才说的降维过程中减少的维度属于肉眼可视的层次，同时压缩也不会带来信息的损失(因为信息冗余了)。如果肉眼不可视，或者没有冗余的特征，降维算法也能工作，不过这样会带来一些信息的损失。但是，降维算法可以从数学上证明，从高维压缩到的低维中最大程度地保留了数据的信息。因此，使用降维算法仍然有很多的好处。

　　降维算法的主要作用是压缩数据与提升机器学习其他算法的效率。通过降维算法，可以将具有几千个特征的数据压缩至若干个特征。另外，降维算法的另一个好处是数据的可视化，例如将5维的数据压缩至2维，然后可以用二维平面来可视。降维算法的主要代表是PCA算法(即主成分分析算法)。

## 聚类算法

聚类算法专注找到元素间的共性，并相应地对它们进行分组。常见的聚类算法是K-Means聚类。根据K-Means，分析人员选择聚类的数量（由变量K表示），算法将元素按物理距离分组到适当的聚类中。

## 深度学习和神经网络算法

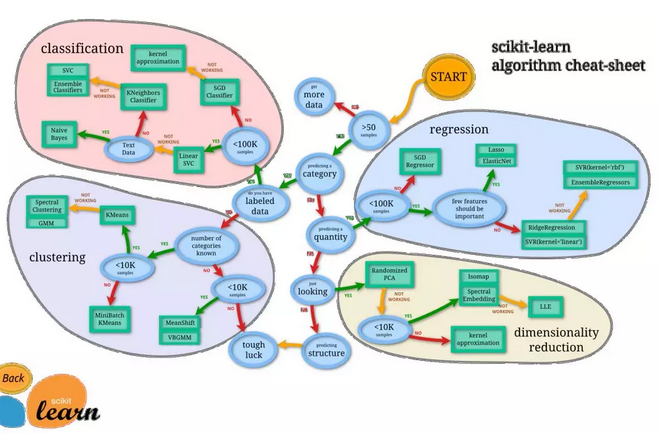
人工神经网络算法基于生物神经网络的结构。深度学习采用神经网络模型并对其进行更新。它们是大型且极其复杂的神经网络，使用少量标注数据和大量未标注数据。神经网络和深度学习具有许多输入，这些输入在产生一个或多个输出之前要通过若干隐藏层。这些连接形成一个特定的循环，模仿人脑处理信息和建立逻辑联系的方式。此外，随着算法的运行，隐藏层通常会变得更小、更细微。

## 集成算法

包括提升树boosting tree和XGBoost

## 其他算法

下面的图表注明了主要的机器学习算法，它们的类别以及之间的关系。

一旦选择并运行算法，你还需要一个非常重要的步骤：对结果进行可视化。虽然与算法编程相比，这看似很简单而没有技术含量。但出色的可视化能力对于数据科学家来说是至关重要的。即使你得出的分析见解再好，一旦没有人能理解也是毫无价值的。

# 框架

numpy数学计算框架

scipy物理计算框架

pandas数据分析框架

matplotlib绘图的框架

scikit-learn机器学习的框架

tensorflow谷歌开源出来的深度学习的框架

keras开源出来的深度学习的框架

# 应用

我们需要明确的是，机器学习有潜力能够改变世界。通过Google Brain和斯坦福机器等研究团队的努力，我们正朝着真正的人工智能迈进。但是，机器学习即将影响的领域有哪些呢？

1. 物联网

物联网或IOT，指家庭和办公室中与网络连接的物理设备。其中一个流行的物联网设备是智能灯泡，其销售额在过去几年中飙升。随着机器学习的进步，物联网设备比以往更智能，更复杂。

物联网相关的机器学习应用主要有两方面，提高设备性能和收集数据。

提高设备性能非常简单，我们可以使用机器学习来定制环境，比如用面部识别软件识别谁是房间里，并相应地调整温度和空调。

收集数据更加简单，通过连接网络的设备（如Amazon echo），亚马逊等公司将收集的用户数据提供给广告商，这些数据包括你会看哪些节目，起床和睡觉的时间，你家里有几口人等等。

2. 聊天机器人

在过去几年，聊天机器人的数量激增，复杂的语言处理算法在不断改进。公司在自己的移动应用和第三方应用上使用聊天机器人，从而提供比更快、更高效的客户服务。例如，如果要从H＆M订购衬衫，你可以告诉他们的聊天机器人你想要的款式和尺寸，轻松订购产品。

3. 自动驾驶汽车

如今，雪佛兰、优步和特斯拉等几家大公司正在开发自动驾驶汽车。这些汽车使用通过机器学习进行导航、维护和安全程序。比如交通标志传感器，它使用监督学习算法来识别交通标志，并与标注数据集进行比较。因此，汽车看到停车标志时，系统将进行确认并停车。