**SparkML(机器学习)介绍(监督学习、半监督学习、无监督学习)**

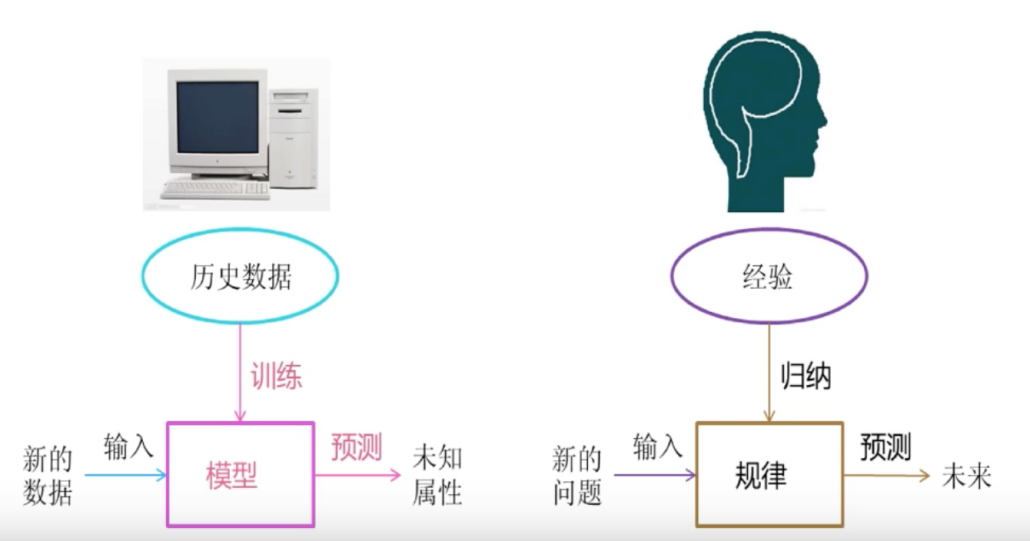
1. 机器学习定义

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。

机器学习还有模式识别、计算机视觉、语音识别、统计学习以及自然语言处理等，而在目前机器学习中，深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。



从广义上来说，机器学习是一种能够赋予机器学习的能力以引让它完成直接编程无法完成的功能的方法。但从实践的意义上来说，机器学习是一种通过利用数据，训练出模型，然后使用模型预测的一种方法。



人类在成长、生活过程中积累了很多的历史与经验。人类定期地对这些经验“归纳”，获得了生活的“规律”。当人类遇到未知的问题或需要对未来进行“推测”的时候，人类使用这些“规律”，对未知问题与未来进行“推测”，从而指导自己的生活和工作。

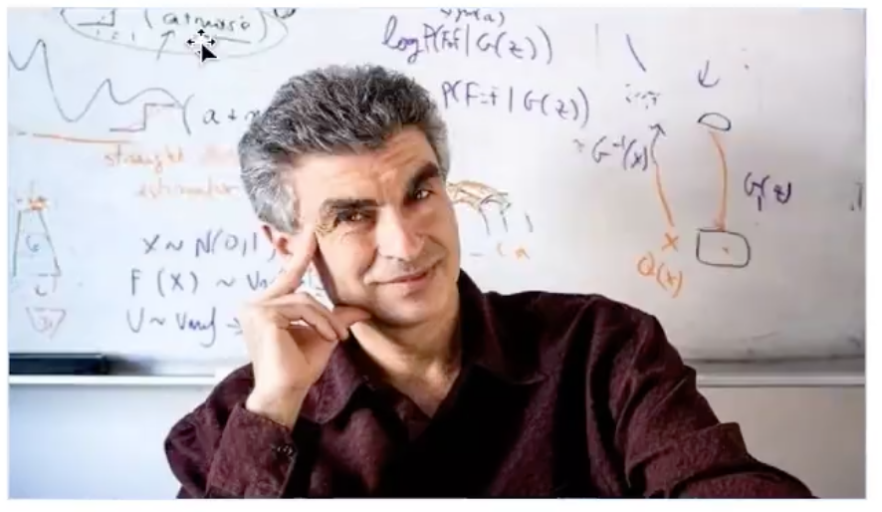
机器学习中的“训练”与“预测”过程可以对应到人类的“归纳” 和 “推测”过程。通过这样的对应，我们可以发现，机器学习的思想并不复杂，仅仅是对人类在生活中学习成长的一个模拟。由于机器学习不是基于编程形成的结果，因此它的处理过程不是因果的逻辑，而是通过归纳思想得出的相关性结论。

机器学习代表人物：

1. Andrew Ng (吴思达) 不久前刚从百度的首席数字科学家的职位离开，他还是斯坦福大学计算机科学系副教授，在线教育平台Coursera 的联合创始人之一。



1. Yoshua Bengio 是蒙特利尔大学计算机科学教授，他因在人工神经网络和深度学习方面的研究而著名。Bengio曾说，他的研究背后最大的野心是理解“获取智慧的学习的原则”。除了AI和ML的其他原则之外，他发表的大部分工作设计用于编码或格式化非结攀比化数据的自动编码器，通过无监督机器学习令计算机可以理解。



1. Yann LeCun作为自2013年以来Fackbook的AI研究主管，LeCun在计算机视觉领域的开创性工作获得广泛认可，计算机视觉是教机器以与人类识别物体相同的方式去“看见”物体，并通过对它他们进行分类的方式学习它他。LeCun也被认为是卷积神经网络的发明人之一，这类模型的目标是创建与生物有机体（例如眼睛或大脑）相同的方式去获取及解释信息的算法。LeCun是纽约大学数据科学中心的创始人。



1. Demis Hassabis是DeepMind联合创始人，DeepMind是谷歌在2014年收购的一家英国AI创业公司。Hassabis的工作重点是结合机器学习和神经科学方法促进人工神经网络的发展。迄今为止，DeepMind最广人知的成功是AlphaGo，这是去年成为第一个打败专业围棋棋士的计算机程序。在这个突破之前，即使是最好的AI也只是跟人类业余围棋玩家交手，而且在实践中经常被人类打败。



1. Hinton是在Google Engineering Fellow，多伦多大学计算机科学系教授。早在1992年，他就发表了关于使用人工神经网络让机器模拟人类处理信息的论文。Hinton在取得AI博士学位之前，最初学过心理学，并将他对人类认知过程的理解应用于计算机。



机器学习例子：

1. 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，可以让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。
2. 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。
3. 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，并根据地理位置分类信件。
4. 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。
5. 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。
6. 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清费。登录FaceBook或GooglePlus，它们就会推荐可能有关联的用户给你。

**监督学习**

通过已有的一部分输入数据与输出数据之间的对应关系，生成一个函数，将输入映射到合适的输出，例如：分类。

监督是从给定的训练数据集中学习一个函数（模型），当新的数据到来时，可以根据这个函数（模型）预测结果。监督学习的训练集要求包括输入和输出，也可以说是特征和目录。训练集中的目标是由人标注（标量）的。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“1”、“2”、“3”等。在建立预测模型，直到模型的预测结果达到一个预期的准确率。常见的监督学习算法包括回归分析和统计分类：

1. 二元分类是机器学习要解决的基本问题，将测试数据分成两个类，如垃圾邮件的判别、房贷是否逾期等问题的判断。
2. 多元分类是二元分类的逻辑延伸。例如，在因特网的流分类的情况下，根据问题的分类，网页可以被归类为体育、新闻、技术等，依此类推。

**无监督学习**

与监督学习相比，无监督学习的训练集没有人为标注的结果。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算法和k-Means算法。这类学习类型的目标不是让效用函数最大化，而是找到训练数据中的近似点。

**半监督学习**

半监督学习（Semi-supervised Learning）是介于监督学习与无监督学习之间一种机器学习方式，是模式识别和机器学习领域研究的重点问题。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。

**Spark 机器学习ML库**



分类：贝叶斯算法、二元SVM分类、逻辑回归分类

聚类：K-means聚类、LAD主题模型

推荐：ALS推荐

回归：线性回归等

决策树：决策树分类

Spark 机器学习算法库主要有两大类

从spark 2.0开始，基于rdd api 实现的算法训，进入到维护模式，spark官方建议你在使用spark机器学习框架的时候，建议你使用基于DataFrame API实现的算法库Spark-ML

1. 基于DataFrame API实现的算法库

Spark-ML

官方说明文档：<http://spark.apache.org/docs/latest/ml-guide.html>

所在的包：org.apache.spark.ml

1. 基于RDD API实现的算法库

Spark-MLlib

官方说明文档：<http://spark.apache.org/docs/latest/mllib-guide.html>

所在的包：org.apache.spark.mllib

在spark ml或mllib中，主要提供了四大类我们常用的算法：

1. 分类算法（Classification）

分类算法主要用于，预测两个类别中的一个或多个类别中的一个

1. 回归算法（Regression）
2. 聚类算法（Clustering）
3. 推荐算法（Collaborative filtering协同过虑）