# 基于机器学习的基础算法研究综述

重庆三峡学院 电子与信息工程学院：雷国平 肖科 罗秀英 杨森 姚佳佳

【摘要】机器学习就是通过对机器输入数据，然后让电脑对数据分析得到一定的信息，并从这些信息中总结出规律，使得这些规律能够在类似的事件中发挥作用，而在这个过程中，机器学习算法起到了关键性作用，本文将对机器学习几种经典的算法做出介绍和相关探讨，这将对未来关于机器学习算法的研究和改进具有极大的价值。

【关键词】机器学习；数据分析；算法

机器学习是人工智能中解决现实问题的主要方法，经过百来年的发展，诞生出了大量经典的方法。阐述了基于机器学习的几种基础算法，着重介绍算法所应用的背景以及解决了什么问题，不同算法的优缺点，在应用中达到了什么效果等，对这些机器学习算法在日常应用中所起到的作用做出深度的思考。

## 1. 何为机器学习?

简而言之，机器学习是从获取的数据中提取，获得，解析，理解知识资讯的一连串过程然后为了对新的，没有见过的数据进行判断以及预测。机器学习过程中通过数据准备和特征工程这个两个过程，找到已知数据集中的模式，然后使用这些模式对新数据进行判断或者预测，而这个过程的核心就是机器学习算法。

## 2. 几种常见的算法

### 2.1 线性回归（Linear Regression）

线性回归分为一元线性回归以及多元线性回归。在回归分析中，若只有一个自变量以及一个因变量，并且两个变量之间能够用直线表示，这就称为一元线性回归分析，其数学模型为y=w1x+b。同理，因变量受多个自变量影响，这时一元线性回归就不能满足条件了，需要建立多元回归模型，其数学模型为y=w1x1+w2x2+w3x3+b，线性回归是使用权重参数来拟合出线性模型的一种方法。从实现的角度来看，这只是包装了为预测对象的普通最小二乘。

### 2.2 梯度下降法（Gradient Descent）

梯度下降是一种优化算法，它是是大多数机器学习算法的核心和灵魂，其作用是在梯度的负值所定义的最陡下降方向上不断的移动（迭代）从而达到最小化某些函数的作用。在机器学习过程中，可以使用梯度下降法来学习线性回归或者神经网络中每个神经元的权重，达到更新参数的目的。

通过梯度下降法，在训练模型的时候能够降低模型的损失，它在机器学习中是一种广泛用于减少损失的方法，像在上坡走路一样找到一种最快最省力的方法。

### 2.3 逻辑回归（Logistic Regression）

统计学中，逻辑模型用于对特定类别或事件（例如通过/失败）的概率进行建模。逻辑回归是使用逻辑函数来估计概率，用来测量因变量与一个或几个自变量之间的关系。因此逻辑回归一般用来分析二元分类以及多元分类。逻辑回归与线性回归的区别在于，线性回归更多的是得到一种较为准确的结果，可能某个数字；而逻辑回归得到的是一种概率，基于0到1之间，是一种非常有效计算概率的机制。逻辑回归可以用于寻找导致某一事件的因素，例如找到某一疾病发生的因素。

### 2.4 决策树演算法（Decision Tree）

决策树是一个树结构（二叉树或非二叉树）。决策树每个决策点表示的是一个特征属性上的测试，而其每个分支代表的是这个特征属性在它的某个值域上的输出结果，决策树的每个叶节点都可以存放一个类别。使用决策树来进行决策的过程是从根节点开始的，然后测试待分类项中相应的特征属性，并按照其值来选择所要输出的分支，直到到达最终的叶子节点，并将叶子节点存放的类别来作为决策结果输出。

决策树演算法背后的基本思想是使用属性选择度量（ASM）选择最佳属性以拆分记录，使得该属性成为决策节点，并将数据集分为更小的子集。通过递归，可以为每个子项不断重复这个过程，从而来启动树的构建，直到其中一个条件匹配为止。属性选择度量用于将数据划分为最佳可能方式的拆分标准的启发式算法，也可以称为拆分规则，因为有助于确定给定节点上元组的断点。

### 2.5 支持向量机（Support Vector Machines）

支持向量机（SVM）可能是目前最流行、被讨论地最多的机器学习算法之一。一般来说，支持向量机被当作是一种分类的方法，其实它可以应用于分类和回归。SVM的核心技巧是处理非线性输入空间，关于它的应用有面部检

测等等。

SVM的主要目标是以最佳方式来对所给定的数据集进行隔离，这需要从给定数据集中找出支持向量机之间具有最大可能边界的超平面。在用来处理非线性以及不可分离平面问题时，使用线性超平面无法解决，但可以通过SVM使用内核技巧将输入空间转换为更高维度的空间来解决，换句话说，可以通过向其添加更多维度将不可分离的问题转化维可以分离的问题。

2.6 K最近邻算法（K-Nearest Neighbor）

K最近邻算法（KNN）是比较成熟的算法，也是最简单的机器学习算法之一，该方法的思路是：在特征空间中，如果一个样本附近的K个最近(即特征空间中最邻近)样本的大多数属于某一个类别，则该样本也属于这个类别。KNN在解决实际问题中也得到了广泛的运用，例如信用评级等。它用于分类和回归问题，基于特征相似的分类演算法。

在KNN中，K是指最近邻居的数量，这是核心决定因素。这里如何选择最佳邻居数是关键，可以将K视为预测模型的控制变量。研究表明，没有最优数量的邻居数适合所有类型的数据集，每个数据集都有自己的要求，在少数邻居的情况下，噪声将对结果具有更高的影响，在大量邻居情况下，计算成本将较高。研究还表明少数邻居具有低偏差，高方差的特点，大量邻居就有更平滑的决策边界，即更低的方差，更高的偏差，因此可以通过在不同的K值上生成模型来检查其性能。

2.7 随机森林分类法（Random forests Classifiers）

随机森林演算法主要用于分类和回归。森林是由树木（特征）组成，拥有的树越多，森林（特征）越坚固。随机森林分类法就是从随机选择的数据样本上创建决策树，从每棵树上获得预测并提供投票，来选择最佳的解决方案。随机森林应用也非常多，例如推荐引擎等。

在分类问题过程中，每棵树进行投票后，最受欢迎的类会被当作最终结果；在回归问题过程中，所有树输出的平均值会被当作最终结果。随机森林演算法的过程是从给定数据集中选择随机样本，然后为每个样本构建决策树，并从每棵决策树上获得预测的结果，再对每个预测的结果来进行投票，最终选择所投票最多的预测结果作为最终的预测结果。

随机森林由于参与该过程的决策树数量被认为是高度准确和稳健的方法，不会受到过度拟合问题的影响，因为它取消了所有预测的平均值，从而抵消了偏差。与其他算法相比，随机森林生成预测的速度很慢，因为它有多个决策树，每当在进行预测的过程时，森林中所有的树都会对相同的给定输入做出预测，然后再对其进行投票，过程较耗时。

2.8 聚类演算法（Clustering Algorithms）

在机器学习中，无监督学习方法一直是我们追的方向而其中的聚类算法更是发现隐藏数据结构与知识的有效手段。聚类是将样本进行分类，然后将样本收集到相类似的组中，根据某些预定义的相似性或距离（不相似性），来进行采样测量。它的应用包括使用案例包括细分客户等。

Kmeans聚类法是聚类算法中最为原始的一类，它可以随机将每个观察样本分配到 k 类中的某一类，然后来计算每个类的平均值，再然后它重新将每个观察的样本分配到与其最为接近的均值所在的类别，最后再重新计算其均值。这一步需要不停的重复，直到不再需要出现新的分配为止。其是在未标记的多维数据集中搜索预定数量的聚类，聚集中心是属于聚类所有点的算术平均值，每个点都靠近自己的集群中心，而不是靠近其他集群的中心。

聚类算法对大数据集处理具有相对可伸缩和高效率的特性，它会尝试找出到使平方误差函数值最小的k个划分。聚类算法是使用最简单的聚类算法之一，其他类似的聚类算法（例如分层聚类等）都是在其基础上深化得来的。

## 3. 结束语

机器学习近些年来发展得越来越迅速，在深度学习，计算机视觉以及语音识别方面得到了广泛得使用并且衍生出大量的实用产品，展现出它向前推进的无限动力，但是要做好机器学习算法方面的研究并不是一件容易的事，因为没有那一个算法能解决所有的实际问题，需要我们后续不断地更新和拓展，将不同场景和不同的算法结合，达到最佳效果的目的，用实际问题来匹配算法，灵活运用。

参考文献：

[1]周志华，机器学习，清华大学出版社，2016.

[2]Nick T G , Campbell K M . Logistic Regression[J]. Methods in Molecular Biology, 2007, 404(404):273.

[3]刘颖超，张纪元. 梯度下降法[J]. 南京理工大学学报(自然科学版), 1993.

[4]None. Applied Logistic Regression Analysis[J]. Technometrics, 1996, 38(2):192-192.

[5]唐华松，姚耀文. 数据挖掘中决策树算法的探讨[J]. 计算机应用研究, 2001(08):21-22+25.

[6]关晓蔷. 基于决策树的分类算法研究[D]. 山西大学.

[7]范昕炜. 支持向量机算法的研究及其应用[D]. 浙江大学, 2003.

[8]Zhang ML, Zhou Z H . ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7):2038-2048.

[9]Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.

[10]Rasmussen E. Clustering algorithms[M]// Information retrieval. Prentice-Hall, Inc. 1992.