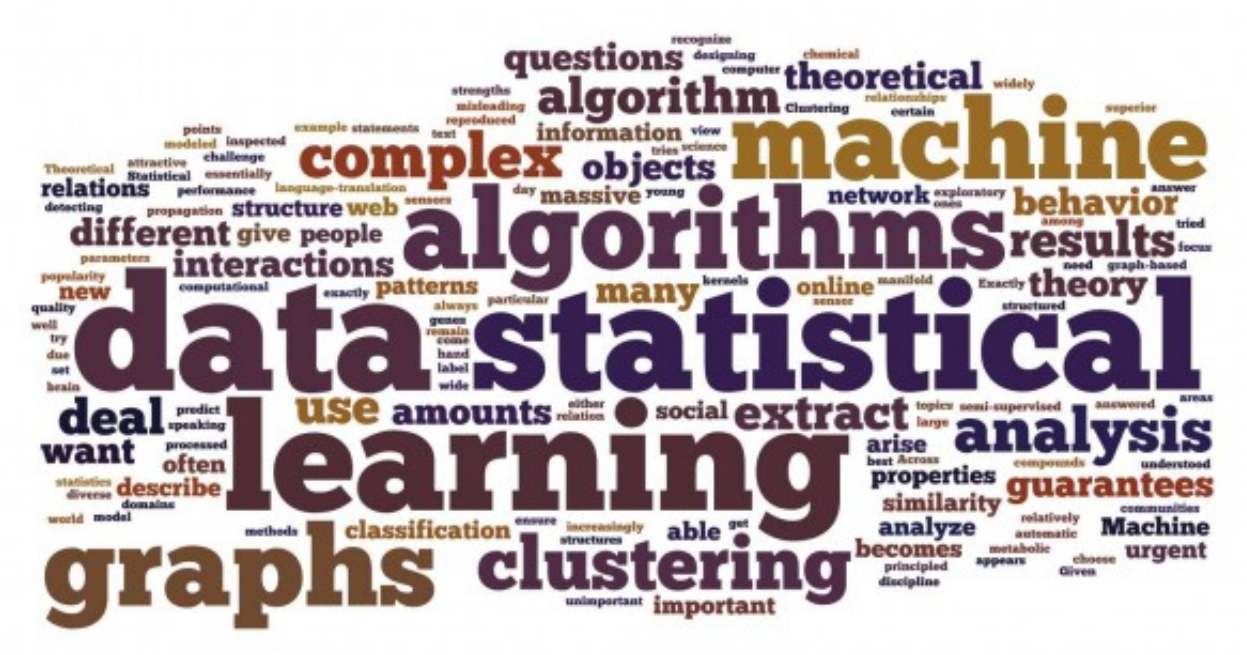
机器学习常见算法分类汇总

作者： 王萌　　星期三, 六月 25, 2014　　Big Data, 大数据, 应用, 热点, 计算　　11 条评论



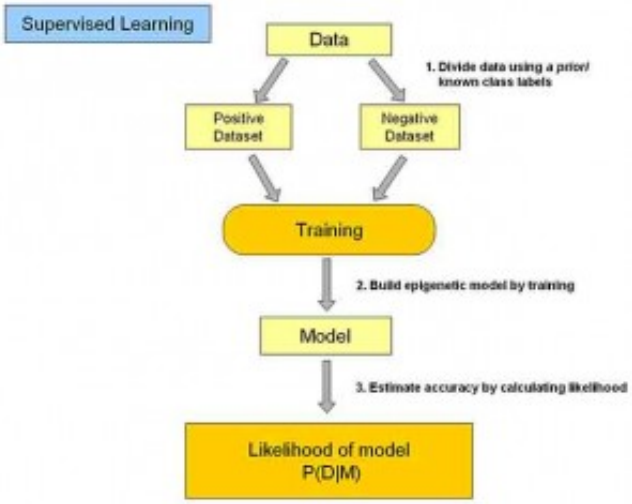
机器学习无疑是当前数据分析领域的一个热点内容。很多人在平时的工作中都或多或少会用到机器学习的算法。这里IT经理网为您总结一下常见的机器学习算法，以供您在工作和学习中参考。

机器学习的算法很多。很多时候困惑人们都是，很多算法是一类算法，而有些算法又是从其他算法中延伸出来的。这里，我们从两个方面来给大家介绍，第一个方面是学习的方式，第二个方面是算法的类似性。

学习方式

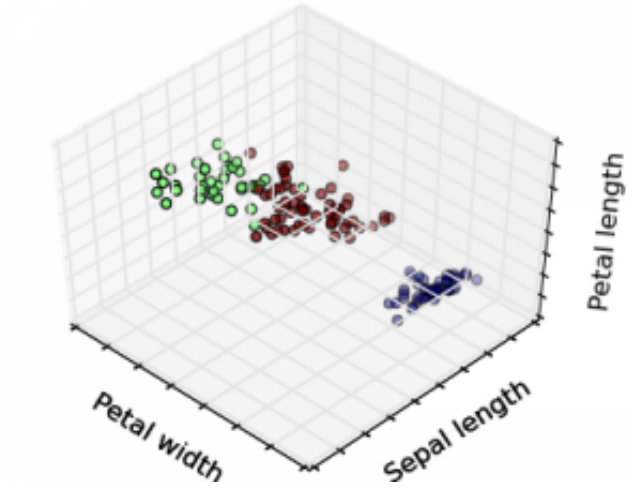
根据数据类型的不同，对一个问题的建模有不同的方式。在机器学习或者人工智能领域，人们首先会考虑算法的学习方式。在机器学习领域，有几种主要的学习方式。将算法按照学习方式分类是一个不错的想法，这样可以让人们在建模和算法选择的时候考虑能根据输入数据来选择最合适的算法来获得最好的结果。

监督式学习：



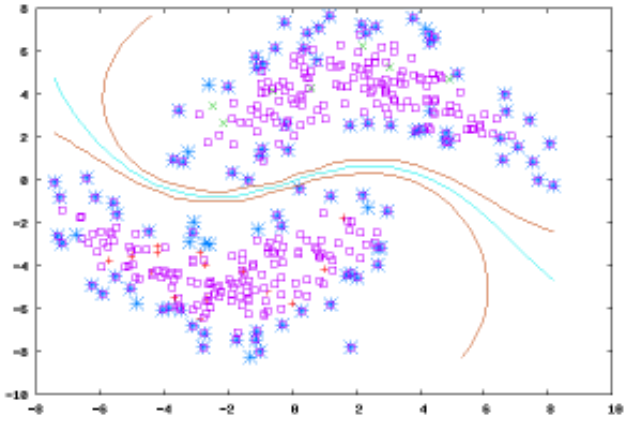
在监督式学习下，输入数据被称为“训练数据”，每组训练数据有一个明确的标识或结果，如对防垃圾邮件系统中“垃圾邮件”“非垃圾邮件”，对手写数字识别中
的“1“，”2“，”3“，”4“等。在建立预测模型的时候，监督式学习建立一个学习过程，将预测结果与“训练数据”的实际结果进行比较，不断的调整预测模型，直到模型的预测结果达到一个预期的准确率。监督式学习的常见应用场景如分类问题和回归问题。常见算法有逻辑回归（Logistic Regression）和反向传递神经网络（Back Propagation Neural Network）

非监督式学习：



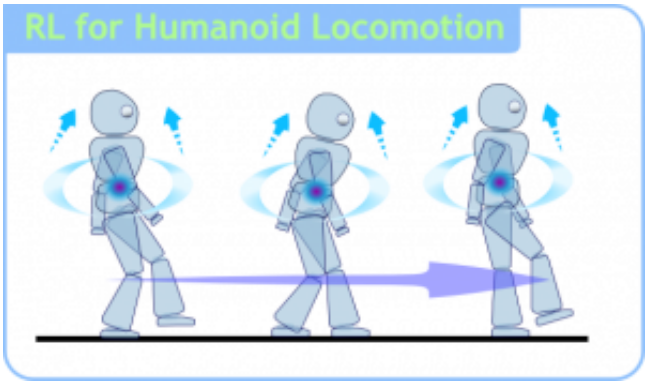
在非监督式学习中，数据并不被特别标识，学习模型是为了推断出数据的一些内在结构。常见的应用场景包括关联规则的学习以及聚类等。常见算法包括Apriori算法以及k-Means算法。

半监督式学习：



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：



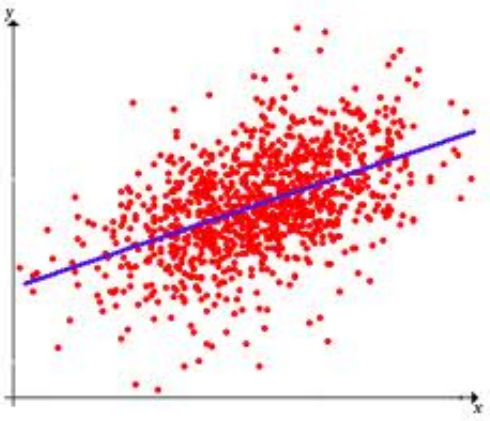
在这种学习模式下，输入数据作为对模型的反馈，不像监督模型那样，输入数据仅仅是作为一个检查模型对错的方式，在强化学习下，输入数据直接反馈到模型，模型必须对此立刻作出调整。常见的应用场景包括动态系统以及机器人控制等。常见算法包括Q-Learning以及时间差学习（Temporal difference learning）

在企业数据应用的场景下， 人们最常用的可能就是监督式学习和非监督式学习的模型。 在图像识别等领域，由于存在大量的非标识的数据和少量的可标识数据， 目前半监督式学习是一个很热的话题。 而强化学习更多的应用在机器人控制及其他需要进行系统控制的领域。

算法类似性

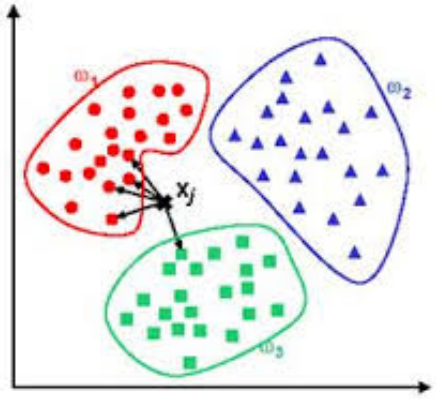
根据算法的功能和形式的类似性，我们可以把算法分类，比如说基于树的算法，基于神经网络的算法等等。当然，机器学习的范围非常庞大，有些算法很难明确归类到某一类。而对于有些分类来说，同一分类的算法可以针对不同类型的问题。这里，我们尽量把常用的算法按照最容易理解的方式进行分类。

回归算法：



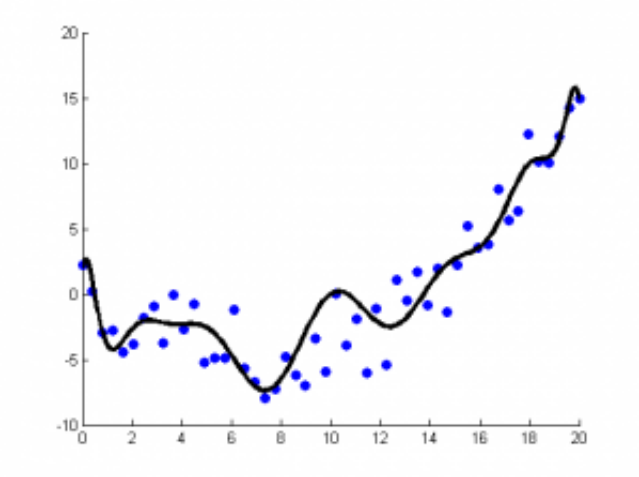
回归算法是试图采用对误差的衡量来探索变量之间的关系的一类算法。回归算法是统计机器学习的利器。在机器学习领域，人们说起回归，有时候是指一类问题，有时候是指一类算法，这一点常常会使初学者有所困惑。常见的回归算法包括：最小二乘法（Ordinary Least Square），逻辑回归（Logistic Regression），逐步式回归（Stepwise Regression），多元自适应回归样条（Multivariate Adaptive Regression Splines）以及本地散点平滑估计（Locally Estimated Scatterplot Smoothing）

基于实例的算法



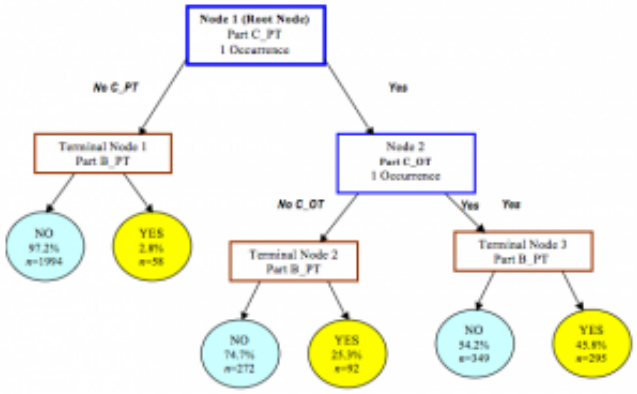
基于实例的算法常常用来对决策问题建立模型，这样的模型常常先选取一批样本数据，然后根据某些近似性把新数据与样本数据进行比较。通过这种方式来寻找最佳的匹配。因此，基于实例的算法常常也被称为“赢家通吃”学习或者“基于记忆的学习”。常见的算法包括 k-Nearest Neighbor(KNN), 学习矢量量化（Learning Vector Quantization， LVQ），以及自组织映射算法（Self-Organizing Map ， SOM）

正则化方法



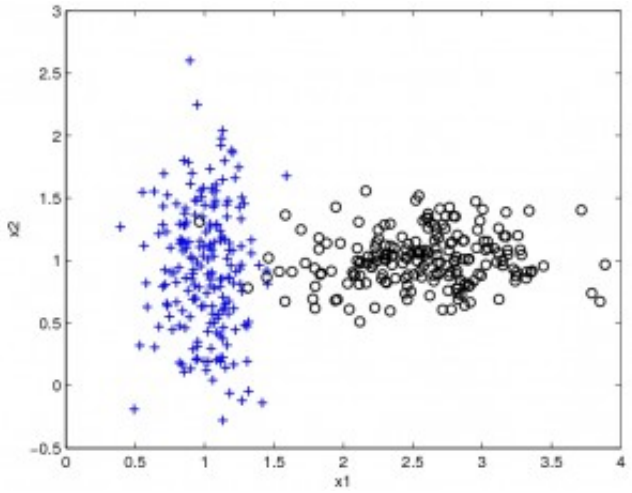
正则化方法是其他算法（通常是回归算法）的延伸，根据算法的复杂度对算法进行调整。正则化方法通常对简单模型予以奖励而对复杂算法予以惩罚。常见的算法包括：Ridge Regression， Least Absolute Shrinkage and Selection Operator（LASSO），以及弹性网络（Elastic Net）。

决策树学习



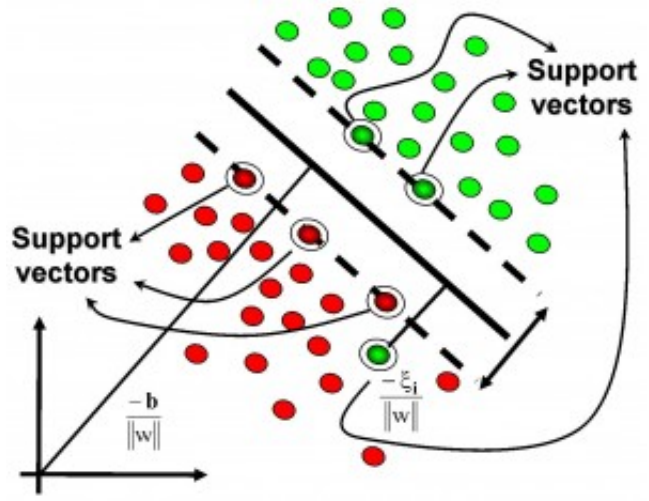
决策树算法根据数据的属性采用树状结构建立决策模型， 决策树模型常常用来解决分类和回归问题。常见的算法包括：分类及回归树（Classification And Regression Tree， CART），ID3 (Iterative Dichotomiser 3)， C4.5， Chi-squared Automatic Interaction Detection(CHAID), Decision Stump, 随机森林（Random Forest），多元自适应回归样条（MARS）以及梯度推进机（Gradient Boosting Machine， GBM）

贝叶斯方法



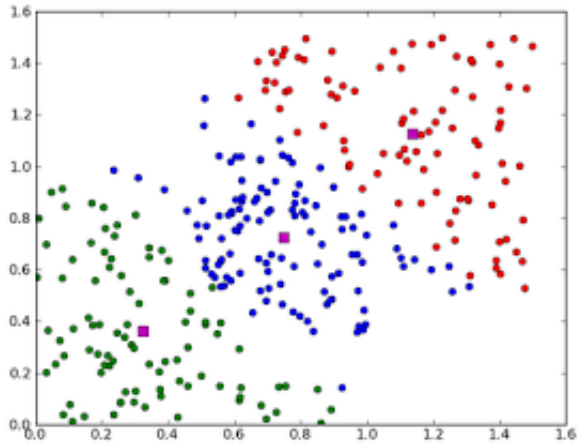
贝叶斯方法算法是基于贝叶斯定理的一类算法，主要用来解决分类和回归问题。常见算法包括：朴素贝叶斯算法，平均单依赖估计（Averaged One-Dependence Estimators， AODE），以及Bayesian Belief Network（BBN）。

基于核的算法



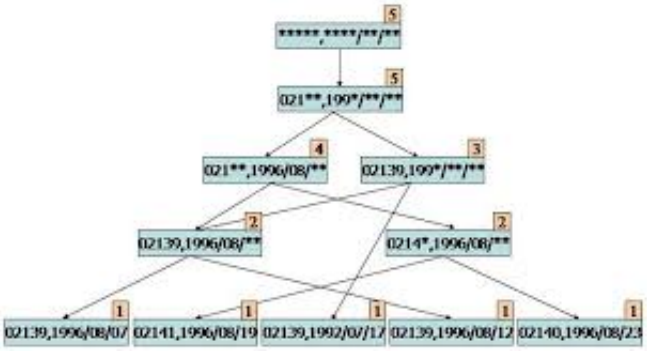
基于核的算法中最著名的莫过于支持向量机（SVM）了。 基于核的算法把输入数据映射到一个高阶的向量空间， 在这些高阶向量空间里， 有些分类或者回归问题能够更容易的解决。 常见的基于核的算法包括：支持向量机（Support Vector Machine， SVM）， 径向基函数（Radial Basis Function ，RBF)， 以及线性判别分析（Linear Discriminate Analysis ，LDA)等

聚类算法



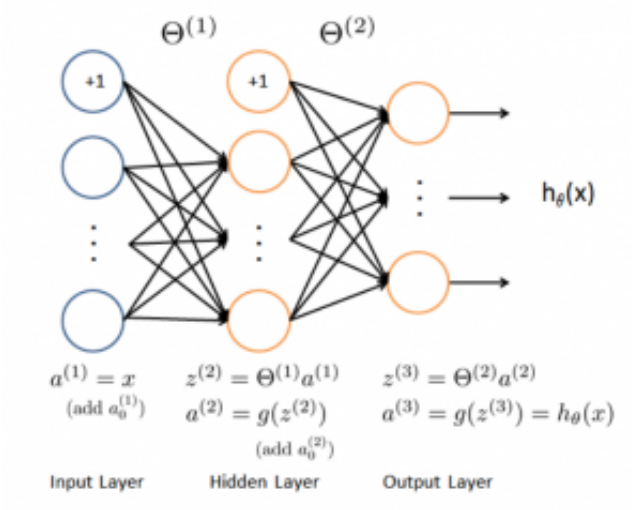
聚类，就像回归一样，有时候人们描述的是一类问题，有时候描述的是一类算法。聚类算法通常按照中心点或者分层的方式对输入数据进行归并。所以的聚类算法都试图找到数据的内在结构，以便按照最大的共同点将数据进行归类。常见的聚类算法包括 k-Means算法以及期望最大化算法（Expectation Maximization， EM）。

关联规则学习



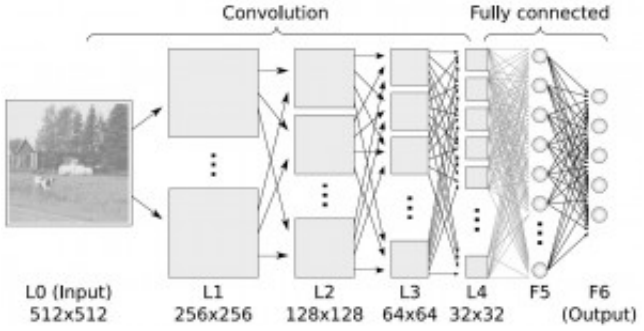
关联规则学习通过寻找最能够解释数据变量之间关系的规则，来找出大量多元数据集中有用的关联规则。常见算法包括 Apriori算法和Eclat算法等。

人工神经网络



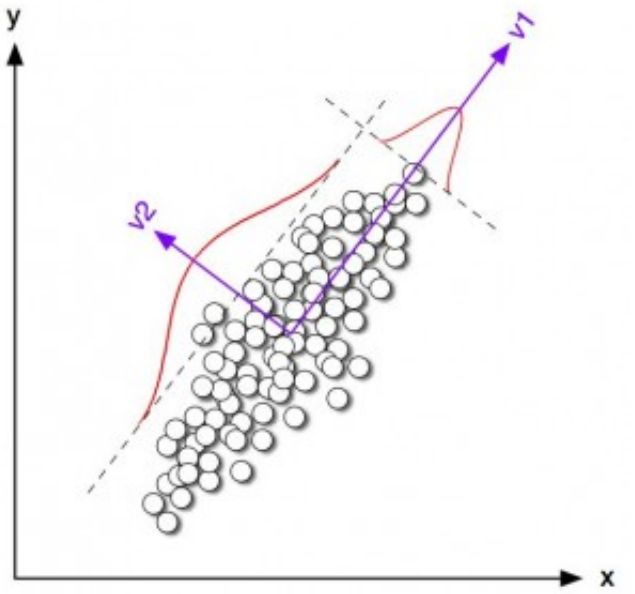
人工神经网络算法模拟生物神经网络，是一类模式匹配算法。通常用于解决分类和回归问题。人工神经网络是机器学习的一个庞大的分支，有几百种不同的算法。（其中深度学习就是其中的一类算法，我们会单独讨论），重要的人工神经网络算法包括：感知器神经网络（Perceptron Neural Network），反向传递（Back Propagation）， Hopfield网络，自组织映射（Self-Organizing Map, SOM）。学习矢量量化（Learning Vector Quantization， LVQ）

深度学习



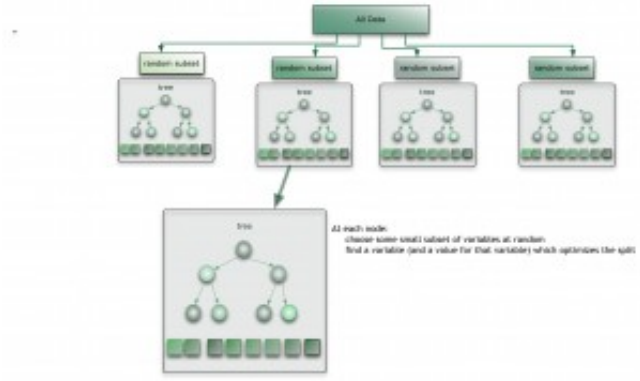
深度学习算法是对人工神经网络的发展。 在近期赢得了很多关注， 特别是百度也开始发力深度学习后， 更是在国内引起了很多关注。 在计算能力变得日益廉价的今天，深度学习试图建立大得多也复杂得多的神经网络。很多深度学习的算法是半监督式学习算法，用来处理存在少量未标识数据的大数据集。常见的深度学习算法包括：受限波尔兹曼机（Restricted Boltzmann Machine， RBN），Deep Belief Networks（DBN），卷积网络（Convolutional Network），堆栈式自动编码器（Stacked Auto-encoders）。

降低维度算法



像聚类算法一样，降低维度算法试图分析数据的内在结构，不过降低维度算法是以非监督学习的方式试图利用较少的信息来归纳或者解释数据。这类算法可以用于高维数据的可视化或者用来简化数据以便监督式学习使用。常见的算法包括：主成份分析（Principle Component Analysis， PCA），偏最小二乘回归（Partial Least Square Regression，PLS）， Sammon映射，多维尺度（Multi-Dimensional Scaling, MDS），投影追踪（Projection Pursuit）等。

集成算法：



集成算法用一些相对较弱的学习模型独立地就同样的样本进行训练，然后把结果整合起来进行整体预测。集成算法的主要难点在于究竟集成哪些独立的较弱的学习模型以及如何把学习结果整合起来。这是一类非常强大的算法，同时也非常流行。常见的算法包括：Boosting， Bootstrapped Aggregation（Bagging），AdaBoost，堆叠泛化（Stacked Generalization， Blending），梯度推进机（Gradient Boosting Machine, GBM），随机森林（Random Forest）。