机器学习的介绍

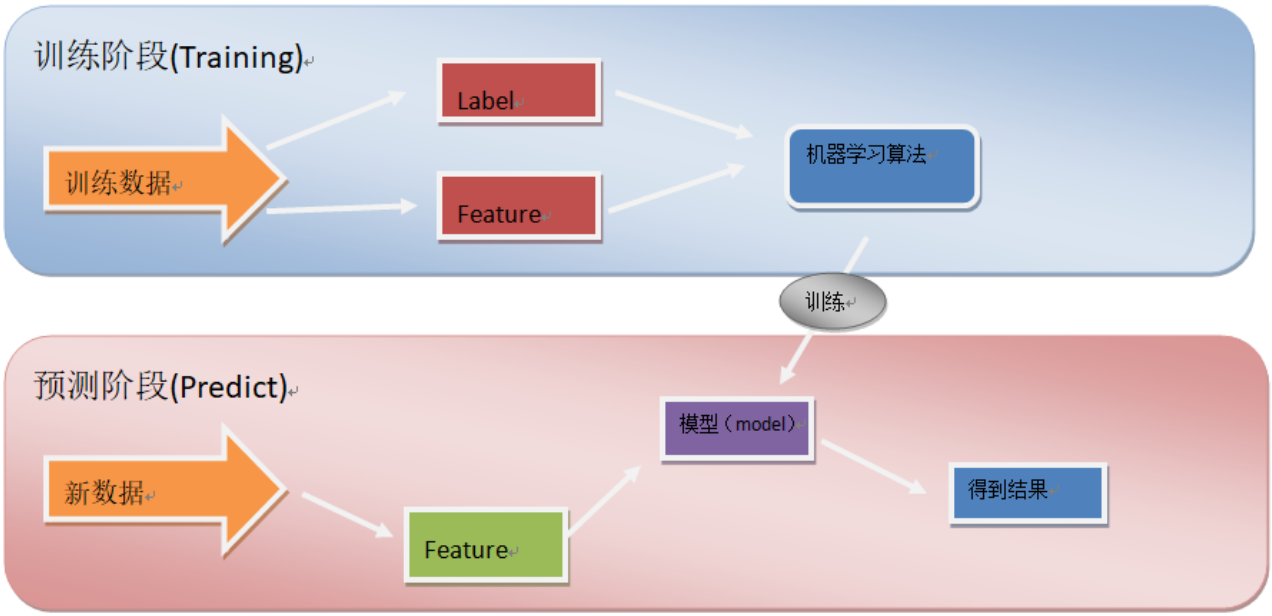
机器学习技术不断的进步，应用相当广泛,例如推荐系统，定向广告，需求预测，垃圾邮件过滤，医学诊断，自然语言处理，搜索引擎，欺诈检测，证券分析，视觉识别，语音识别，手写识别，频率识别等等。

一、机器学习架构

机器学习（Machine Learning）通过算法、使用历史数据进行训练，训练完成后会产生模型。未来当有新的数据提供时，我们可以使用训练产生的模型进行预测。

机器学习训练用的数据是由Feature、Label组成的。

- Feature :数据的特征,也叫做特征列，例如湿度、风向、季节、气压。
- Label：数据的标签，也叫做目标值，例如降雨（0.不会下雨，1.会下雨），天气状况（1.晴天，2.雨天，3.阴天，4.雾天）



（1）训练阶段（Training）

训练数据是过去累计的历史数据，可能是文本文件、数据库文件或者是其它的来源。经过Feature Extraction（特征提取），产生Feature（数据特征）于Label（预测目标），然后经过机器学习算法的训练后产生模型。

（2）预测阶段（Predict）

新输入的数据，经过Feature Extraction（特征提取）产生Feature（数据特征），使用训练完成的模型进行预测，最后产生预测结果。

二、机器学习的分类

（1）有监督学习

对于有监督的学习（Supervised Learning），从现有数据我们希望预测的答案有下列分类。

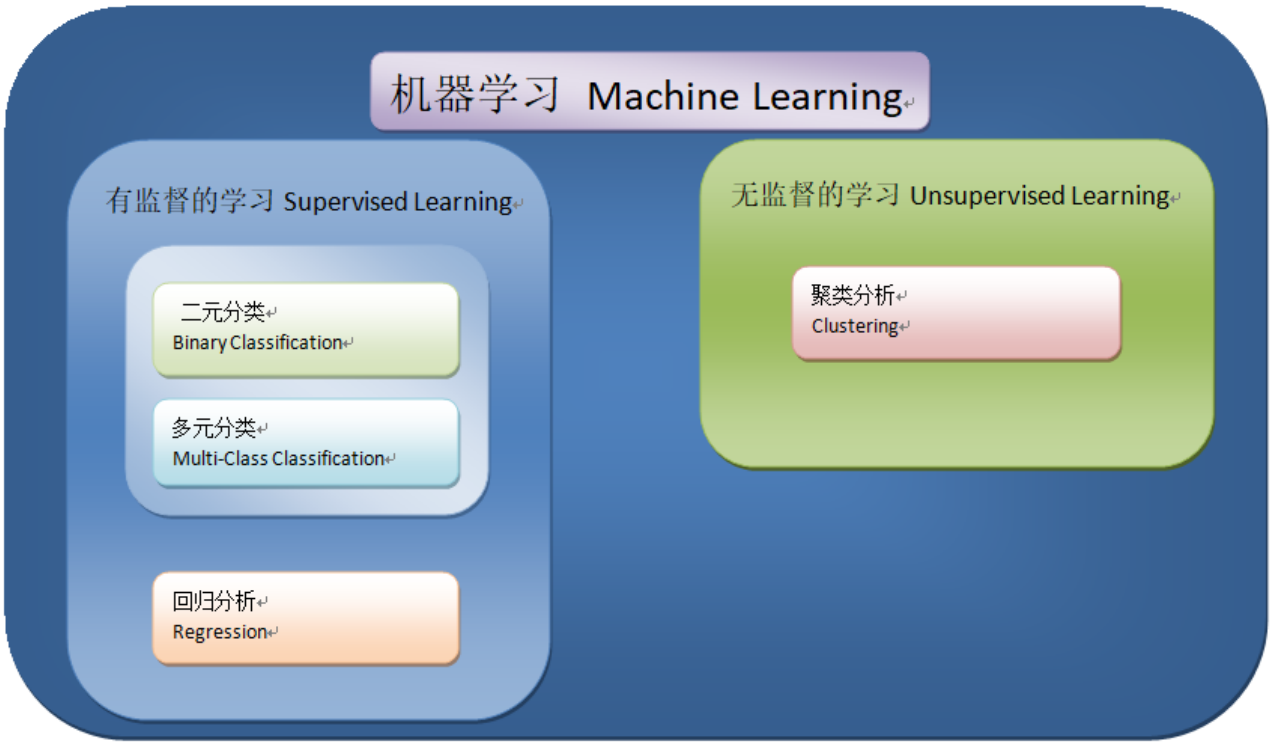
- 二元分类：我们已知湿度、风向、风速、季节、气压等数据特征，希望预测当天是否会下雨（0.不会下雨，1.会下雨）。目标Label只有两种选项。
- 多元分类：我们已知湿度、风向、风速、季节、气压等数据特征，希望预测当天的天气（1.晴天，2.雨天，3.阴天，4.雾天）。目标Label有多个选项。
- 回归分析：我们已知湿度、风向、风速、季节、气压等数据特征，希望预测当天的气温。目标Label是一个连续值，是一种方程的计算方法。

（2）无监督学习

对于无监督的学习（Unsupervised Learning），从现有的数据我们不知道要预测的答案，所以没有Label（预测的目标）。

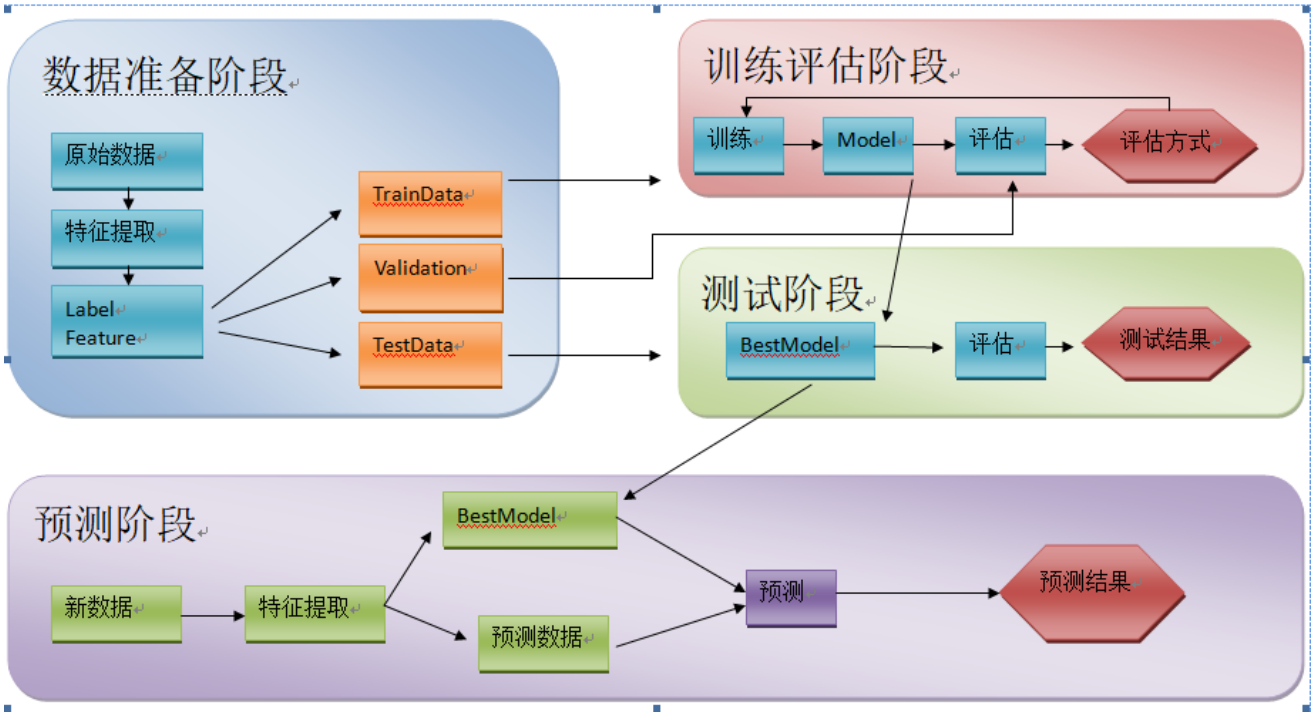
- cluster聚类分析：的目的是将数据分成几个相异性最大的群组，而群组内的相似度最高。

（3）机器学习算法类别概括图表：



分类	算法	Features（特征）	Label（预测目标）
有监督学习	二元分类（Binary Classification）	风向、风速、季节、气压等数据特征	只有0,1的选项
有监督学习	多元分类（Multi-Class Classification）	风向、风速、季节、气压等数据特征	有多个选项
有监督学习	回归分析（Regression）	风向、风速、季节、气压等数据特征	值是一个范围（-10～30）度的范围
无监督学习	聚类分析（Clustering）	风向、风速、季节、气压等数据特征	无Label，物以类聚是

三、机器学习的四个阶段



（1）数据准备阶段

原始数据（可能是文本文件、数据库或其它来源）经过数据转类，提取特征字段与标签字段，产生机器学习所需要的格式，然后将数据以随机方式分为3部分（trainData、validationData、testData）并返回数据，供下一阶段训练评估使用。

（2）训练评估阶段

我们将使用 trainData数据进行训练，并产生模型，然后使用validationData验证模型的准确率。这个过程要重复很多次才能够找出最佳的参数的组合。评估方式：二元分类使用AUC、多元回归使用accuracy、回归分析使用RMSE。训练评估完成后，会成产生一个最好的模型bestModel。

（3）测试阶段

之前阶段产生了最佳模型bestModel，我们会使用另外一组数据testData再次测试，以避免overfitting(过拟合)的问题。如果训练评估阶段准确度很高，但是测试阶段的准确度很低，代表可能有overfitting的问题。如果测试与训练评估阶段的结果准确度差异不大，代表没有没有overfitting问题。

（4）预测阶段

新输入的数据，经过Feature Extraction（特征提取）产生Feature（特征），使用训练完成的最佳模型，也就是bestModel进行预测，最后产生比较不错的预测结果。