# 机器学习算法

我们知道，机器学习算法是一种能让机器从数据中学习的算法，由此需要给“学习”一个准确的定义。1997年，Mitchell提出了简洁的定义：“对于某个任务*T*和性能度量*P*，一个计算机程序被认为可以从经验*E*中学习是指，通过经验*E*改进后，它在任务*T*上由性能度量*P*衡量的性能有所提升。”可以看出，机器学习算法由任务*T*、经验*E*和性能度量*P*构建而成，下文将分别从这三个方面对其进行阐述。

## 任务*T*

针对常规具有明确流程的任务所编写的确定性程序并不能称之为机器学习，机器学习通常是用于帮助人们解决一些人为设定以及确定性程序难以解决的问题。

学习本身是一个获得完成任务能力的过程，通常机器学习任务定义为机器学习系统应如何处理样本，而样本是包含各类已量化特征的需要处理的对象或事件。机器可以通过已有的样本，掌握对这些或未知的样本进行处理的能力，以此完成某些任务。

机器学习可以解决诸多类型的任务，常见的任务类型列举如下:

* **分类**：对于某具有*n*个特征的样本，计算机程序需要指定该样本属于*k*类中的哪一类。学习算法通常会返回一个函数。设，代表算法将向量映射到数值所代表的类别。例如在手写数字识别任务中，机器需要学到将所给手写数字图片分类到中某一数字。另一著名应用为ImageNet项目，有超过1400万张带标注图片，机器可利用这些数据学习分类1000种对象以实现图像识别。
* **回归**：在该任务中，计算机需要根据已知输入返回预测值。学习算法通常会返回一个函数。这一类任务的事例有，房价预测、股票预测等，这类预测常用于算法交易中。对比分类问题，两者的区别在于回归任务的输出为连续值，而分类任务的输出为离散值。事实上，回归任务可以用来解决分类任务，常规分类任务的实际输出值为样本归属各类的概率，再选取概率值最大的类作为其分类结果。
* **转录**：在这类任务中，机器学习系统的输入为一些非结构化的数据，如文本图片、语音等，并将其转录为离散文本信息。例如文字识别，将文本图片中的信息转录为文字输出；语音识别，将音频信号转录为所说的字符或单词ID的编码。
* **机器翻译**：机器翻译任务，本质上是将输入的某种语言的符号序列，转化成另一语言的符号序列。通常应用于自然语言处理中，例如科大讯飞的人工智能同声传译软件，网易的有道翻译官等。
* **结构化输出**：结构化输出的结果为向量或是包含多个值的数据结构，且输出的不同元素间有重要联系。事实上，转录、翻译等任务都可以归属于这一范畴内。其他的应用，例如语法分析——将自然语言句子映射到语法结构树，并标记树的节点为动词、名词、副词等。此外，还有图像描述，对输入的一幅图像运用自然语言句子描述该图像。
* **异常检测**：此类任务类似于分类任务，计算机程序在一组事件或对象中筛选，找出并标记非正常或非典型个体。例如，用户的信用卡发生非正常的交易行为，公司可紧急冻结该信用卡，避免可能因被盗窃而产生的财产损失。

上述介绍仅为部分机器学习可完成的任务，并非严格意义上的机器学习任务分类。

## 性能度量*P*

性能度量是一种评估机器学习算法性能的定量度量。需要注意的是，性能度量*P*通常是依据任务*T*所制定的。

通常，我们更关注机器学习算法在未观测数据上的性能，这决定了在实际应用中的性能。因此，区别于已观测过的“训练集”，我们用未观测过的“测试集”对模型性能进行评估。

### 准确率

在分类及转录任务中，常采用“准确率”作为性能度量。机器学习算法*f*在样例集*D*上的分类准确率定义为

## 查准率、查全率与*F*1

在某些场合，例如流行病的检测中，查准或查全往往更总要，因此有时需采用“查准率”或“查全率”作为性能度量，给定分类结果的“混淆矩阵”如下表所示

表1 分类结果混淆矩阵

**Table** 1 **Classification results confusion matrix**

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测结果 | |
| 正例 | 反例 |
| 正例 | *TP* (真正例) | *FN* (假反例) |
| 反例 | *FP* (假正例) | *TN* (真反例) |

查准率*P*与查全率*R*分别定义为

可以看出查准率和查全率彼此矛盾，而实际问题中比较两种机器学习算法的性能，需要同时考虑查准率和查全率，我们常用*F*1度量作为综合考虑这两者的性能度量

在实际应用中，有时更重视查准率，有时更重视查全率，因此给出*F*1度量的一般形式——，定义如下

## ROC与AUC

通常，很多分类器对测试样本预测出一个之间的实数，然后将此值与0.5比较，大于0.5则判为正例，否则为反例。本质上说，这一预测值等价于该样本为正例的概率值，我们可以依据此预测值将测试样本进行排序，“最可能”为正例的排在最前面，以此类推，得到新的样本序列。

针对这个样本序列，可以逐个设置“截断点”将样本分为两个部分，前一部分判作正例，后一部分判作反例。每次计算“真正例率”(TPR) 和“假正例率”(FPR)，两者定义如下

以假正例率为*x*轴，真正例率为*y*轴，依次连接每次计算结果得到“ROC曲线”，以鸢尾花分类为例绘制出ROC曲线如下图所示。

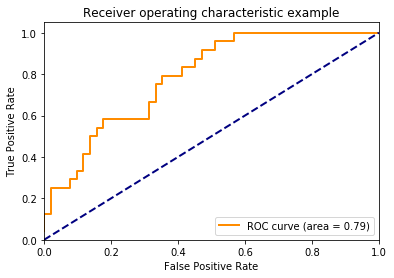


图2 ROC曲线示例图

**Fig.2 ROC curve example diagram**

得到的ROC曲线并不能直接反映机器学习算法的性能好坏，但可通过计算ROC曲线下面积作为判断依据，记为AUC。由图可知，AUC值越大，则模型分类性能越好，计算公式如下

## 经验*E*

目前，大部分机器学习算法是从数据上获取经验。根据学习过程的不同经验，机器学习算法可以大致分为监督学习、无监督学习、半监督学习、迁移学习和强化学习。

### 监督学习

监督学习算法的数据集通常含有多个特征，其本质特征在于每个样本都有一个“标签”或“目标”。机器通过比较预测结果与真实结果对算法进行调整，此时“标签”或“目标”发挥着“监督”机器学习过程的作用，故称之为“监督学习”。

监督学习的典型算法有：线性回归、逻辑回归、多层感知机、卷积神经网络等；典型应用有：回归分析、分类等。

### 无监督学习

与监督学习相反，无监督学习算法的的数据集不显式地有“标签”或“目标”，仅含有多个特征。算法希望通过学习寻求数据间的内在模式和统计规律，从而获得样本数据的结构特征。需要注意的是，无监督学习往往不需要事先进行训练。这种“无中生有”的思想更接近人类的学习方式。

无监督学习的典型算法有：自动编码机、受限玻尔兹曼机，K-means等；典型应用有：聚类和异常检测等。

### 半监督学习

半监督学习算法的数据集结合了前两者算法的数据集，即部分数据集有“标签”或“目标”，而另一部分数据集没有。在现实任务中，未标记样本多、有标记样本少是一个比较普遍的现象。算法希望可利用好大量未标记数据及少量已标记数据提升模型泛化能力。半监督学习算法需假设未标记样本所揭示的数据分布信息与类别标记存在联系，一种是聚类假设而另一种为流形假设。

半监督学习的典型算法有：直推式SVM、图半监督学习、协同训练等；典型应用涵盖了监督学习与无监督学习的应用范畴。

### 迁移学习

迁移学习算法类似于半监督学习，同样是利用某一仅有少量标签样本的数据集进行学习，区别在于后者是运用已有的模型来解决此缺乏标签样本数据的学习问题。通俗地讲，一个人学会了骑自行车，则他很容易学会开摩托车。在实际应用中，例如训练一个识别猫的神经网络，若从头开始训练则需要百万张带标注的数据及大量的显卡资源。而利用迁移学习，可使用VGG16这样成熟的模型，只需用少量数据训练最后一层，即可快速得到较好的识别模型。

### 强化学习

强化学习算法与前四者最大的区别在于，机器的学习过程并不局限于一个固定的数据集上，而会与“环境”进行交互。

机器在于环境交互的过程中，其行为会获得一定的奖惩。机器需要通过在环境中不断尝试而学得一个“策略”，使得长期累积的奖赏最大化。

目前强化学习技术在游戏、机器人控制、参数优化、机器视觉等领域中得到了广泛的应用，并被认为是迈向通用人工智能的重要途径。