# 机器学习（Machine Learning）

简单纲要，蒋学渊整理

机器学习的基本概念

机器学习是一种特定的人工智能方案，使用相关的算法，让计算机使用一定的模型，通过对一定量的数据训练集进行学习。通过训练，计算机可以获得一个训练后模型，从而获得对海量同类数据的处理能力。评价结果的优劣可以由对这些海量数据处理的正确性及资源占用情况等入手。

机器学习的常规分类

* 监督学习（Supervised learning）

通过有预先人工判断标注目标结果的训练集训练模型，使模型获得对未来数据的处理能力。典型案例是分类器模型（Classification）、回归模型（Regression）。

* 非监督学习（Unsupervised learning）

通过对未作人工标注的训练集训练模型，使模型获得对未来数据的处理能力。典型案例是聚类（Clustering）。

* 半监督学习（Semi-supervised learning）

训练集分为两部分，通常是较小规模的人工标注训练集和较大规模的未标注训练集组合。核心思想是先用标注的小规模训练集进行监督学习，得到的初步模型再用未标注集合作调整提高普适能力。

机器学习的通常流程

1. 数据采集及预处理

从海量数据中，使用合理的抽样方法，抽样出分布符合一定要求的原始数据，并对数据进行必要的正规化等预处理。

1. 数据初析及特征选择

对抽样的数据结果进行分析，如计算特征相关性等，决定需保留的数据特征。

1. 模型选择及训练

根据特征规模及训练目标是否容易人工判断等，决定训练使用的模型，及是否进行人工标注等。用选择的训练模型和训练方法训练数据

1. 应用模型及交叉验证

用训练的模型处理一般数据，验证结果的正确率和资源使用情况等。

常见训练模型

* 分类器模型（Classification）

给定特征向量和分类集合，分类器模型需要完成的任务是将特征向量与正确的类别完成映射。

* + 朴素贝叶斯（Naive Bayes）

朴素贝叶斯是一种估值方法，基于随机变量之间的条件独立假设和贝叶斯定理计算随机变量的条件概率估值，可以基于这个概率估值构建分类器。

核心公式：

常用场合：文本分类、垃圾邮件识别分拣

* + K近邻算法（K Nearest Neighbor，KNN）

K近邻算法是以全部的训练样本作为代表点，计算测试样本与所有训练样本之间的距离，选取前K近邻者类别投票决定测试样本的类别，据此构建分类器。

核心概念：欧几里德距离（Euclidean Distance）用于连续空间的距离计算，汉明距离（Hamming Distance）用于离散空间的距离计算

常用场合：文本分类

主要优点：思路直观，实现简单

主要不足：随着样本空间增大，由于需要计算测试样本到所有已标识样本的距离，计算量和存储量会迅速增大

* + 支持向量机（Support Vector Machine，SVM）

支持向量机所解决的核心问题是，将高维度的输入向量非线性的映射到向量空间，并给出最佳定义的线性边界。

经典模式识别函数：（m为均值向量，M为相关矩阵）

* + 决策树（Decision Tree）

决策树是一种树型的分类模型，把样本从根节点沿各节点的判断条件逐步到某个叶子节点来实现分类，叶子节点即为样本所属的分类。处理复杂问题的时候，还可以把多个决策树构成决策森林，分别对样本分类后，综合分类结果。

* 回归模型（Regression）

给定一个自变量向量和一个因变量向量，两组变量之间有一批已知对应关系的数据组，需要找到一个函数来描述这个对应关系，计算这个函数的过程称为回归。在计算回归的过程中，通常会对自变量和因变量数据进行预先分析，此后得到一个预判的函数关系。在这个函数关系中，可能存在一些尚未确定的参数，这个带有未定参数的函数关系称为回归模型。

* + 线性回归（Linear Regression）

分为一元和多元，模型公式：。系数的计算方法根据所给训练数据相关性程度的不同和优化目标的不同，一般可以采用的有最小二乘法、偏最小二乘法、主成分回归计算法等。

模型优点：普适性好，通过合理的适用不同计算方法训练系数，对多数问题都能给出较好的模型结果；计算简单，预测阶段计算仅为向量点积计算，运算速度较快，适合性能要求较严格的场合。

* + 逻辑回归（Logistic Regression）

逻辑回归的模型公式：。系数的计算方法通常使用最大似然估计方法。

模型优点：对特定概率类问题，可以获得最佳的预测结果。

* 聚类模型（Clustering）

给出若干样本点和目标的类别数N，通过算法，把相似样本点聚合，形成N个类别模型，即聚类模型。

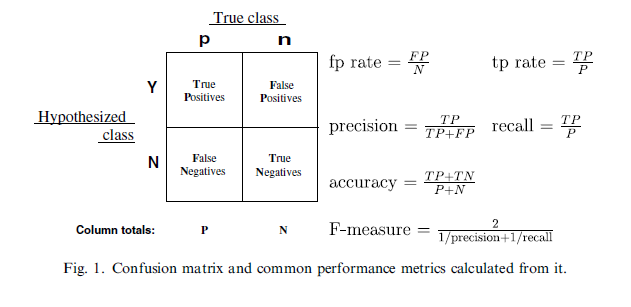
* + K平均聚类算法（K-means Clustering）

K平均聚类算法是将给定样本点分成K聚类的算法。初始时，随机选定K个点作为类别核心，计算其他点与这K个点的距离，为其他点决定类别；所有点完成后，计算所有类别的平均点，作为下一次类别核心；迭代计算，直至类别核心稳定，即可以得到K 个聚类。

机器学习的验证标准

* 基本标准ROC/AUC

ROC（Receiver Operating Characteristics）是对分类器进行整理归纳，可视化进行对比的一种工具。基本的分类模型是P-N模型，目标是将数据点分为Positive或者Negative两类，较复杂的分类模型可能输出在连续空间上，这时一般是使用截断阈值（threshold）的方式来实现最终的分类模型。

根据训练集合P-N分类情况和分类器回答Y-N情况，总体可以将所有数据点分作4类：  
TP-True Positives，训练集合的数据点是Positive，分类器划入Yes集合，属于第一种正确的分类；TN-True Negatives，训练集合的数据点是Negative，分类器划入No集合，属于第二种正确的分类；FP-False Positives，训练集合的数据点是Negative，分类器划入Yes集合，属于第一种错误的分类；FN-False Negatives，训练集合的数据点是Positive，分类器划入No集合，属于第二种错误的分类；

根据模型使用不同参量可以得到的FP率及TP率，可以得到不同的数据点，将数据点连接起来的曲线称为ROC曲线，ROC曲线下方面积称为AUC（Area Under Curve），AUC数值越大，说明分类器的总体效果越好。

* MAP

MAP（Mean Average Precision）适用于判断正例位置的优劣，当得到一组计算结果时，排名靠前的正例可以获得更大的权重，得到更高的得分。MAP值高，说明模型计算结果正例排名靠前，召回情况较好。

核心公式：

* DCG/NDCG

DCG（Discounted Cumulative Gain）与MAP相似，同样是判断正例位置优劣的一种计算标准，核心公式：

IDCG（Ideal DCG），在同等条件下，最佳排序得到的DCG值

NDCG（Normalized DCG），用DCG/IDCG的方法对DCG值正规化的结果