Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. -- Arthur Samuel

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. -- Tom Mitchell

第 1 章 绪论

1.1 什么是机器学习

“机器学习”(Machine Learning，简记 ML)就是让计算机具备从大量数据中学习的能力之一系列方法。

机器学习使用很多统计方法，统计学家也称之为“统计学习”(Statistical Learning)，但机器学习在本质上起源于计算机科学的“人工智能”(Artificial Intelligence，简记 AI)领域。

所谓“人工智能”，就是让计算机具备像人类一样的各种智能，比如听说读写与识别图像的能力。例如，人类可轻松识别垃圾邮件，计算机是否也具备这样的能力？

(1) 硬编码 vs. 学习

机器学习的一个早期成功案例是“过滤垃圾邮件”(spam filtering)。随着电子邮件的兴起，垃圾邮件也越来越多。如何自动过滤“垃圾邮件”(spam)，而不错杀“正常邮件”(email 或 ham)？

传统方法将人类关于垃圾邮件的知识直接告诉计算机，将这些规则进行计算机编程，称为“硬编码”(hard coding)；但效果不好。

一个突破性的想法是引入“学习”(Learning)，即无须人类告诉计算机何为垃圾邮件，而由计算机通过学习大量的数据自行判断垃圾邮件。

给予计算机大量的邮件，其中每封邮件都已事先由人类“标注”(labeled)为“正常邮件”或“垃圾邮件”。根据海量邮件的大数据，计算机可统计出不同词汇在正常邮件与垃圾邮件的出现频率。

比如，垃圾邮件经常出现“代开发票”一词，则根据“贝叶斯规则”(Bayes rule)可算出，给定包含“代开发票”一词，该邮件为垃圾邮件的条件概率。然后用数学方法将这些信息综合起来(比如，朴素贝叶斯)，最终算出此邮

件为垃圾邮件的概率。

如果此概率超过某临界值(比如 0.9)，则归类为垃圾邮件。这种方法称为“贝叶斯垃圾邮件过滤”(Bayes spam filtering)。

计算机判断垃圾邮件的能力正是通过学习大量的数据而获得，故名“机器学习”(machine learning)。而上述“贝叶斯垃圾邮件过滤”即为一种“学习机器”(learning machine)或“学习器”(learner)。

(2) 大数据与机器学习

机器学习的效果依赖于“大数据”(big data)。数据量越大，则学习的效果越好。而且，机器学习的能力还可根据最新的数据不断地动态更新。如果只给计算机提供 100 封邮件，可以想象机器学习的效果会很差。

有些机器学习的算法出现得很早，比如“人工神经网络”(Artificial Neural Network，简记 ANN)早在 1960 年代就已提出，但当时既无大数据，电脑运算速度也慢，故停滞不前，直至近年才复兴，发展为炙手可热的“深度学习”(deep learning)。

1.2 机器学习的分类

机器学习主要分为两类，即“监督学习”(supervised learning)与“非监督学习”(unsupervised learning)。

所谓“监督学习”，就是有目标的学习；而“非监督学习”则为无目标的学习。

对于监督学习，第i位个体(或观测值)的数据可写为$(y_i, \mathbf{x}_i)$，而我们的任务是用向量$\mathbf{x}_i$来预测y_i。

在上述过滤垃圾邮件的例子中，$\mathbf{x}_i$为不同词汇在第i封邮件中出现的频率，而y_i则是取值为 0 或 1 的“虚拟变量”或“哑变量”(dummy variable)，表示此封邮件是否为垃圾邮件。

在监督学习中，目标很明确，就是用$\mathbf{x}_i$预测y_i，而y_i起着监督与指导学习过程的作用，故名“监督学习”，有时也称为“预测性建模”(predictive modeling)。

对于非监督学习，数据只是$\mathbf{x}_i$，并没有y_i，而学习过程就是为了在$\mathbf{x}_i$中识别某种模式或规律(pattern recognition)。常见的非监督学习方法包括“主成分分析”(principal component analysis)与“聚类分析”(cluster analysis)等，参见第 16-17 章。

对于监督学习，还可根据y_i的性质进一步细分。如果y_i为连续变量，则称为“回归”(regression)。反之，如果y_i为离散变量(比如，虚拟变量)，则称为“分类”(classification)。

其他类型的机器学习：“半监督学习”(semi-supervised learning)、“强化学习”(reinforcement learning)等，适用于更为专门的场景。

1.3 机器学习的术语

机器学习始于计算机科学的人工智能领域，有一套独特的术语。

由于“样本数据”(sample data)主要用于训练计算机获得学习能力，故一般称为“训练数据”(training data)。事实上，在进行机器学习时，一般将所有数据分为两类，其中大部分数据构成“训练数据”；而少部分数据则作为“测试数据”(test data)、“验证数据”(validation data)或“保留数据”(hold-out data)。

测试数据仅用于检验机器学习的效果，以避免出现“过拟合”(overfit)，即虽然样本内拟合效果好，但外推预测效果差。

统计学一般称$\mathbf{x}_i$为“自变量”(independent variables)、“解释变量”(regressors, explanatory variables)或“协变量”(covariates)，而机器学习则称$\mathbf{x}_i$为“特征”(features)、“特征向量”(feature vector)、“预测变量”(predictors)或“属性”(attributes)。

统计学称y_i为“因变量”、“被解释变量”(dependent variable)或“结果变量”(outcome variable)，而机器学习则称y_i为“响应变量”(response)或“目标”(target)。进一步，对于分类问题，机器学习有时称离散的响应变量为“标签”(label)或“类别”(class)。

统计学称第i个数据为“观测值”(observation)或“样本点”(data point)，而机器学习则通常称之为“样例”(example)或“示例”(instance)。

1.4 机器如何学习

大多数的机器学习问题都是监督学习，因为许多问题都可纳入到此框架中。比如，人脸识别(facial recognition)。

首先，可将传感器捕捉到的人脸相片转换为“像素”(pixel)的矩阵(参见图 1.1)，其中每个像素用一个 0-255 之间的整数表示其“灰度”(grayscale，假设为黑白图片)，0 表示全黑，而 255 表示全白。

其次，将此矩阵的每列依次叠放，构成一个很长的列向量$\mathbf{x}_i$。比如，假设此图片的像素为 100×100，则其特征向量的维度高达 10,000 维。机器学习的任务就是要判断此图像是否为人脸($y_i=1,0$)，或是否为某人的脸($y_i=1,0$)。

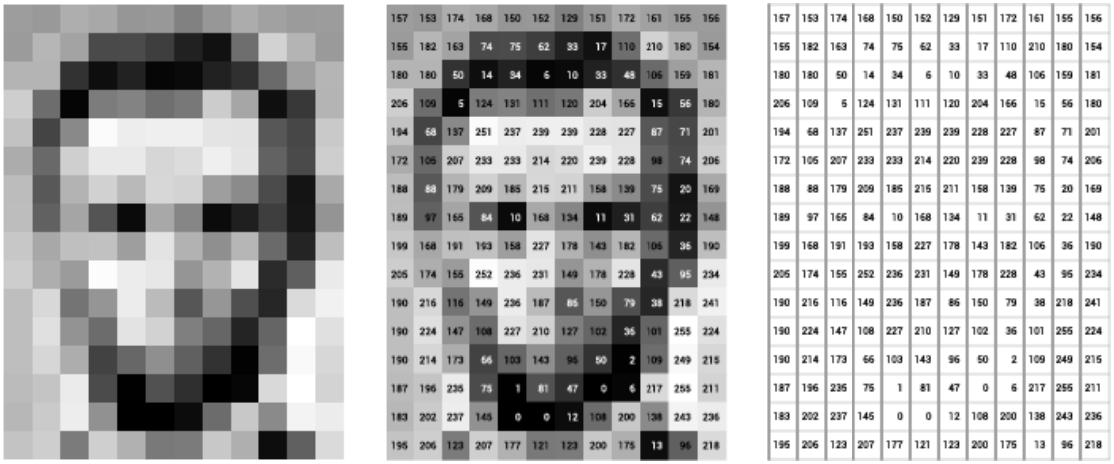

图 1.1 美国总统林肯的灰度图

使用“硬编码”的方法将行不通。我们无法告诉计算机究竟怎样的图像才算人脸。

即使简单如 0-9 的手写数字，如果使用硬编码的方法，计算机也力不从心(比如，邮局为自动分拣而识别手写邮编)。不同人的手写数字千差万别，参见图 1.2。

图 1.2 手写体数字的示例

真正的突破依然来自“学习”的方法，即给予计算机大量的图像，有些包含人脸，而有些不含人脸，让计算机通过学习大量的数据而获得识别人脸的能力。

给定一个未知函数 $f(\mathbf{x}_i;\boldsymbol{\beta})$，机器学习的目标就是通过训练数据 $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ 来学习此未知函数 $f(\mathbf{x}_i;\boldsymbol{\beta})$，其中 $\boldsymbol{\beta}$ 为未知的参数向量。具体来说，希望根据训练数据找到一个函数

$$\hat{y}_i = \hat{f}(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) \qquad (1.1)$$

使得预测的 $\hat{y}_i$ 与实际的 y_i 之间差距最小，比如在测试数据(test data)中的

“均方误差”(mean squared errors)最小：

$$\min \ \mathrm{E}\left[(y_i - \hat{y}_i)^2\right] \qquad (1.2)$$

甚至“无人驾驶汽车”(driverless cars)也可纳入此一般的机器学习框架。此时，特征向量$\mathbf{x}_t$为汽车上各种传感器在时刻t实时输送的各种数据(参见图 1.3)，而y_t为是否在时刻t刹车($y_t = 1,0$)。例如，若预测$\hat{y}_t = \hat{f}(\mathbf{x}_t;\hat{\boldsymbol{\beta}}) > 0$，则刹车；反之，则不刹车。当然，$y_t$也可以是连续变量，比如方向盘的角度。

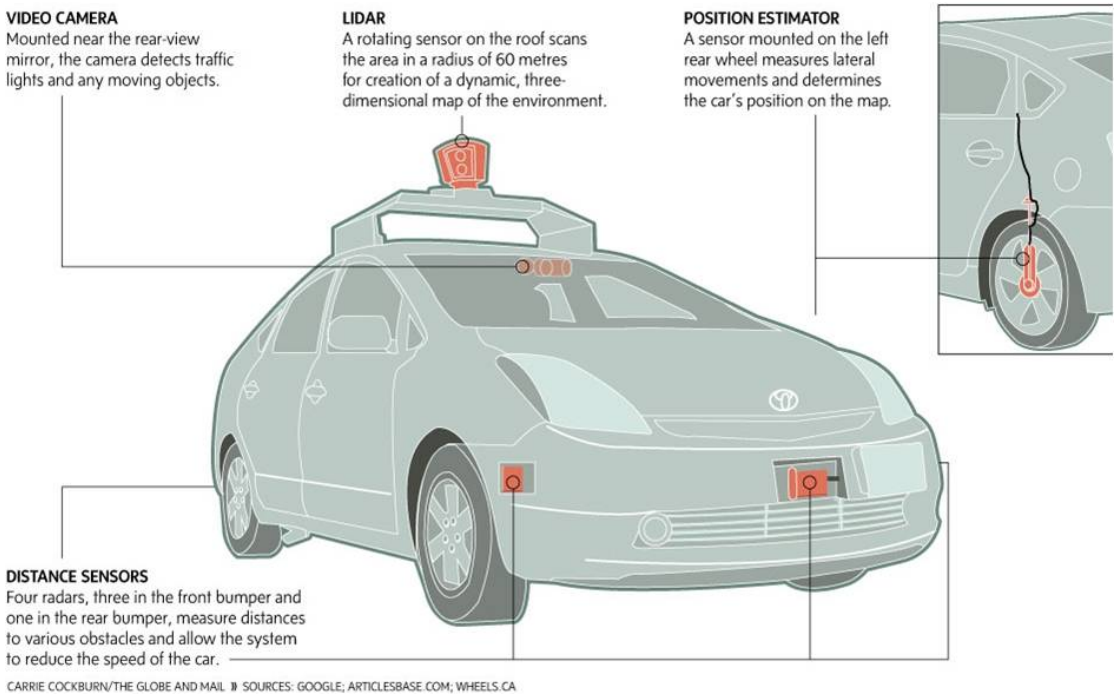

图 1.3 谷歌提供的无人驾驶示意图

1.5 机器学习与统计学、计量经济学的关系

2018 年 9 月，麻省理工学院名誉校长 Eric Grimson 在接受澎湃新闻采访时即表示，机器学习在未来“会变得像使用 Word、PowerPoint 或者 Excel 一样”。既然如此，机器学习与传统的统计学，以及广泛用于社会科学的计量经济学有何关系呢？

(1) 研究目标的不同

在表面上，机器学习通常使用大数据(样本容量很大或变量很多)，而统计学与计量经济学则一般样本较小。但这种区别正变得日益模糊，因为统计学与计量经济学也越来越多地使用大数据。

这三个学科的主要区别在于研究目标有所不同。机器学习的主要目的在于“预测”(prediction)，统计学侧重于“统计建模与推断” (statistically modeling and inference），而计量经济学则着重于“因果推断”(causal inference)，参见表 1.1。

表 1.1 不同学科的比较

学科	预测	因果推断	可解释性	主要方法
机器学习	***	*	*	最优化、算法
统计学	**	**	***	渐近理论
计量经济学	*	***	***	渐近理论

注：“***”表示强，“**”表示中等，“*”表示弱。

机器学习的主要目标在于预测，即根据$\mathbf{x}_i$预测y_i。为达到此目的，可使用任何函数$f(\mathbf{x}_i;\boldsymbol{\beta})$，甚至是难以解释的黑箱方法(比如神经网络)；只要预测结果$\hat{y}_i$接近y_i就好。因此，机器学习方法的“可解释性”(interpretability)一般比较差。

机器学习的模型中，即使有参数$\hat{\boldsymbol{\beta}}$，也只是作为预测的中间手段与桥梁而已。机器学习的关注重点为$\hat{y}_i$，几乎完全生活在$\hat{y}_i$的世界，成功与否就看$\hat{y}_i$的预测效果。

计量经济学的主要目标则在于“因果推断”(causal inferences)，即推断$\mathbf{x}_i$对y_i的“因果效应”(causal effects)。为了识别并便于解释此“因果关系”(causality)，经济学家通常需对$f(\mathbf{x}_i;\boldsymbol{\beta})$的函数形式作很强的假定，比如假设线性回归模型(因为线性模型最易解释参数$\boldsymbol{\beta}$的含义)：

$$f(\mathbf{x}_i;\boldsymbol{\beta})=\mathbf{x}_i'\boldsymbol{\beta} \qquad (1.3)$$

将精力集中于得到未知参数$\boldsymbol{\beta}$的估计量$\hat{\boldsymbol{\beta}}$，并针对$\hat{\boldsymbol{\beta}}$进行统计推断。计量经济学关注的重点为$\hat{\boldsymbol{\beta}}$，几乎总是生活在$\hat{\boldsymbol{\beta}}$的世界。由于计量经济学对于函数形式作了较强假定，可能与现实不符，故预测效果可能不理想。

对于统计学而言，则十分注重对于$\hat{\boldsymbol{\beta}}$的统计推断(这是统计学的核心方法)，但所建模型可能只是相关关系，而不像计量经济学那样专注于因果关系。

Breiman(2001a)认为在统计建模(statistically modeling)中存在“两种文化”(two cultures)。占主流的“数据建模文化”(data modeling culture)，首先假设从$\mathbf{x}$到y的一个随机数据模型(a stochastic data model)，然后进行参数估计，而模型验证(model validation)则通过样本内的拟合优度与残差来检验。

占少数派的“算法建模文化”(algorithmic modeling culture)，在寻找从$\mathbf{x}$到y的映射$f(\mathbf{x})$时，并不对$f(\mathbf{x})$作任何假设；而模型验证则通过样本外的预测准确率来检验，即机器学习或统计学习的方法。

(2) 方法论的区别

在方法论上，机器学习主要使用“最优化”(optimization)方法，经常表现为最小化某个“目标函数”(objective function)或“损失函数”(loss function)。一般需要通过某种“迭代算法”(iterative algorithm）寻找近似的“数值解”(numerical solution)。

机器学习的目标是让预测结果$\hat{y}_i$尽量接近y_i，而y_i可以观测，故度量机器学习的效果非常简单，直接比较$\hat{y}_i$与y_i的接近程度即可(比如，均方误差、预测错误率等)，并不需要使用渐近理论。

对于统计学与计量经济学而言，虽也常作最优化，但由于关注重点为不可观测的参数$\boldsymbol{\beta}$，故在估计$\hat{\boldsymbol{\beta}}$之后，无法直接比较$\hat{\boldsymbol{\beta}}$与$\boldsymbol{\beta}$的接近程度。

只能使用概率统计的“渐近理论”(asymptotics)，也称为“大样本理论”(large sample theory)，证明当样本容量趋向无穷大($n \to \infty$)时，估计量$\hat{\boldsymbol{\beta}}$(依概率)收敛到真实参数$\boldsymbol{\beta}$，以及$\hat{\boldsymbol{\beta}}$服从渐近正态分布等性质(以便进行统计推断)。

然后，并辅之以小样本的“蒙特卡洛模拟”(Monte Carlo simulation)进行验证。

由于研究目标不同，故机器学习与统计学、计量经济学在研究范式上有着本质的区别，参见图 1.4。

一般认为机器学习使用大量的统计方法。但事实上，机器学习几乎不进行统计推断，而只是使用统计方法估计函数 $f(\mathbf{x}_i;\boldsymbol{\beta})$，比如“最大似然估计”(Maximum Likelihood Estimation，简记 MLE)。

由于机器学习可直接比较预测值 $\hat{y}$ 与实际值 y，故无须使用高深的渐近理论(依赖于大数定律与中心极限定理等)来证明预测效果；在这个意义上，机器学习反而比统计学或计量经济学更为简单！

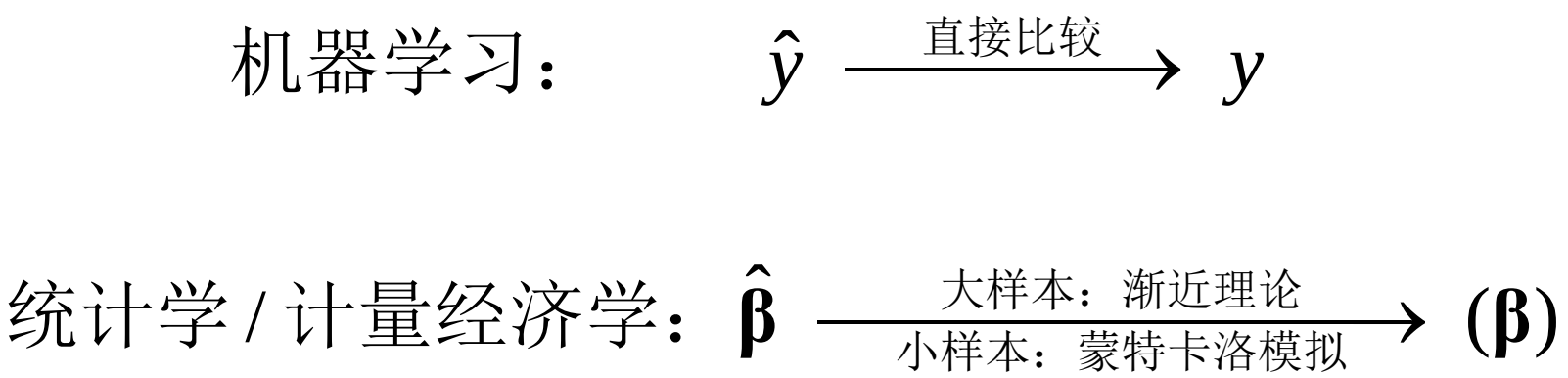

图 1.4 不同学科的研究范式

注：图中$(\boldsymbol{\beta})$的括号表示$\boldsymbol{\beta}$不可观测

(3) 学科间的融合

2011 年图灵奖得主、人工智能先驱 Judea Pearl 即主张将因果推断引入人工智能领域。既然因果推断是人类智能的重要体现，未来的“机器人”怎么能缺少因果推断的能力呢？

业界人士可能认为，做商业预测只需要变量之间的相关关系即可，并不一定需要因果关系。比如，你看到街上有些人带伞，就可预测可能下雨；但人们带伞显然并不导致下雨。

许多商业问题事实上都涉及因果效应。例如，你想预测某个公司政策的效应，比如将排名第一的搜索结果放到排名第三，将会对其点击量有多少影响？此预测其实是在估计该公司政策的因果效应。

又比如，假设你收集了关于宾馆房价与入住率的数据，想预测宾馆房价对入住率的影响。如果直接根据相关关系进行预测，会发现宾馆入住率与房价显著正相关。但这并非因果关系，因为在旅游旺季，宾馆爆满而房价也很高。对于考察公司政策效应的这一类重要预测问题，其本质也是在做因果推断。

另一方面，因果推断也离不开预测。事实上，因果推断本质上恰恰是在做预测。比如，某地区实施了扶贫政策，你想评估此政策的效应。此时，该地区扶贫之后的状态可以度量，但最关键的信息却不可观测，即此地区如果没有实施扶贫会怎么样？对于这种“反事实的结果”(counterfactual outcome)，一般只能进行估计或预测。这也正是“鲁宾因果模型”(Rubin's Causal Model)的核心思想(Rubin, 1974)。由于机器学习擅长作预测，故机器学习方法在因果推断方面也大有用武之地。