机器学习算法概述及环境安装配置

陈 鑫

2023 年 10 月 6 日

本书配套源码，课件，以及视频均可在
https://github.com/aliasch/Machine-Learning-Fundamentals-and-Practice 下载。

安装 python 和 pycharm

首先到 python 官网下载 python 安装包（建议选择次新版本，因为最新版本可能与第三方软件库存在兼容问题，可能导致一些包无法正常使用），python 的安装过程非常简单，基本上根据提示点击下一步就可以了，注意在安装过程中，勾选将安装路径添加到系统环境变量中。为了节约时间，该软件事先已经下载好了，下面简单地演示一下安装过程。

到 pycharm 官网下载最新版本的 pycharm 社区版，免费，免注册用户，足以满足日常开发使用。如果有教育网邮箱，也可以注册之后使用专业版。其安装过程也非常简单，下面简单地演示一下安装过程。

修改 pip 源为国内源

因为默认的 python 的第三方库安装源在国外，下载速度比较慢，因此，建议修改 pip 源，以加速下载软件包的速度，一般可以使用阿里云源，豆瓣源或者清华源。在 windows 的用户目录C:\Users\aliasch\中，新建目录 “pip”，并在该目录中新建 “pip.ini” 配置文件，并加上配置信息。

```
[global]
index-url = https://mirrors.aliyun.com/pypi/simple/
[install]
trusted-host=mirrors.aliyun.com
```

安装第三方库和使用 jupyterlab

在 pycharm 中，选择 file 菜单，选择 settings，在弹出的 settings，选择 project->python interpreter，点击右边的 + 号，在弹出的 avaliable packages 对话款的搜索栏中，输入想要安装的第三方库，例如 Numpy，下面会显示相应的软件包，双击安装即可。如果需要制定安装的版本，则可在右下角进行选择。安装完 jupyterlab 之后，可以在 pycharm 下方，点击 terminal 输入 jupyter lab 启动 jupyterlab。

在 pycharm 中启动 jupyterlab 可以很方便地以交互式地方式学习 numpy 和 matplotlib。首先我们点击 pycharm 左边的终端按钮，这时下方会出现命令行提示符，我们输入 jupyter lab 即可启动 jupyterlab。

机器学习简介

- 机器学习在许多方面都可以看作是数据科学能力延伸的主要手段。
- 机器学习是用数据科学的计算能力和算法能力去弥补统计方法的不足，其最终结果是为那些目前既没有高效的理论支持，又没有高效的计算方法的统计推理与数据探索问题提供解决方法。
- 机器学习的本质就是借助数学模型理解数据。

当我们给模型装上可以适应观测数据的可调参数时，“学习”就开始了；此时的程序被认为具有从数据中“学习”的能力。一旦模型可以拟合旧的观测数据，那么它们就可以预测并解释新的观测数据。

通俗地讲，机器学习是让机器通过模拟或实现人类的学习行为来获取新的知识或技能，重新组织已有的知识结构，以不断改善自身智能。

Definition (机器学习的定义)

一个被广为引用的抽象定义为：给定任务 T、相关的经验 E 以及关于学习效果的度量 P，机器学习就是通过对经验 E 的学习来优化任务 T 完成效果的度量 P 的一个过程。

机器学习的原理与人类学习十分相似：对已知的经验信息加以提炼，以掌握完成某项任务的方法。在机器学习中，用于学习的经验数据称为训练数据，完成任务的方法称为模型。机器学习的核心就是针对给定任务，设计出以训练数据为其输入，以模型为其输出的算法，然后利用该模型对新数据进行预测。

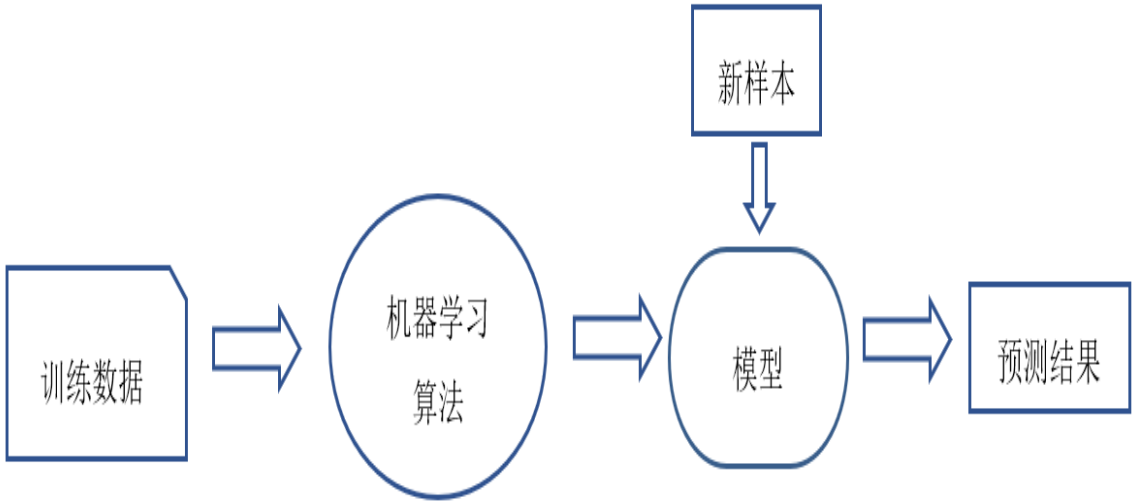

图 1: 机器学习算法工作原理

机器学习的优缺点

机器学习跟人类学习相比，有以下的一些优点：

- 机器学习算法可以从海量数据中提取与任务相关的重要特征。例如，在虹膜识别技术中，机器学习算法能从众多医学数据以及生物特征中选取细丝、冠状、条纹、隐窝等细节特征，来区别任意两个不同的虹膜，人类远无法做到这一点。
- 机器学习算法可以自动地对模型进行调整，以适应不断变化的环境。例如，在房价预测系统中，机器学习算法能自动根据类似的小区的最新交易记录，对某小区的房价预测做出迅速调整，其反应速度也远超人类。

机器学习也存在一些问题:

- 机器学习算法需要大量的训练数据来训练模型。在数据不足的情况下，机器学习算法往往会面临两个挑战。第一，训练数据的代表性不够好。这使得模型在面对完全陌生的任务场景时会不知所措，例如无人驾驶汽车算法的训练数据中没有包含雪天的行驶记录，那么经训练所得到的模型很可能无法在雪天给出正确的驾驶指令。
- 训练数据的一些特殊的特征可能将模型带入过度拟合的误区。过度拟合就是指算法过度解读训练数据，从而失去了模型的可推广性（泛化能力）。在无人驾驶汽车的例子中，如果训练数据不足，例如只有两条数据：遇到红灯停车，遇到红色停止标志停车，在这种情况下，机器学习算法可能会从仅有的这两条数据中提炼出如下模型：前方出现红色物体则停车，这就是过度拟合。机器学习存在的第二个问题是：目前它还没有在创造性的工作领域中取得成效。

机器学习分类

机器学习一般可以分为两类：有监督学习和无监督学习。除此之外，还有介于两者之间的强化学习。

监督学习是指对数据的若干特征与若干标签（类型）之间的关联性进行建模的过程。这类学习过程可根据标签取离散值或连续值，进一步划分为分类任务或回归任务。监督学习的任务是根据对象的特征对标签的取值进行预测。

- 如果标签的取值是有限个可能值，则称相应的监督学习问题为分类问题。例如，手写数字识别是一个经典的监督学习 10 分类问题。
- 如果标签取值于某个区间的实数，则称相应的监督学习为回归问题。例如，在房价预测问题中，训练数据有房屋面积、学区、与地铁站距离等特征，并含有交易价格作为其标签，因为价格取连续值，所以房价预测问题是一个回归问题。

无监督学习是指对不带任何标签的数据特征进行建模，通常被看成是一种“让数据自己介绍自己”的过程。例如，在手写数字识别问题中，忽略训练数据的标签，仅根据特征对训练数据进行分类，机器学习算法也能将数据分成 10 类。每一类具有相同的数字，但它无法识别出具体数字，因为训练数据中并不包含这类信息，这就是一个无监督学习问题。

在无监督学习中，聚类算法和降维算法是两类应用最为广泛算法。

机器学习中的一些的基本术语

萼片长度	萼片宽度	花瓣长度	花瓣宽度	种类
5.1	3.5	1.4	0.2	0
4.9	3	1.4	0.2	0
7	3.2	4.7	1.4	1
6.4	3.2	4.5	1.5	1
6.3	3.3	6	2.5	2
5.8	2.7	5.1	1.9	2

图 2: 鸢尾花（部分）数据集

- 样本：每一行数据称为一个样本（也称为训练数据），每一个样本都含有特征和标签。图2中一共包含 6 个样本。样本的所有可能取值称为样本空间。
- 特征：除了最后一列，每一列数据都表示一个特征（属性），图2的 4 个特征分别是鸢尾花的萼片长度、萼片宽度、花瓣长度和花瓣宽度，单位为 cm。特征的所有可能取值称为特征空间。
- 标签：在回归问题中，训练数据含有一个数值标签 $y \in \mathbb{R}$; 在 k 元分类问题中，训练数据含有一个向量标签 $y \in \{0,1\}^k$，标签的所有可能取值，称为标签空间。表 1 中最后一列表示的就是标签，种类 0 表示 setosa（山鸢尾），1 表示 versicolor（变色鸢尾花），2 表示 virginica（弗吉尼亚鸢尾）。在分类问题中，标签表示对象的类别，例如山鸢尾；在回归问题中，标签表示对象的一个数值属性，例如在房价预测中表示价格。
- 数据集：随机抽取观测对象采集到的数据的整体称为数据集。例如，鸢尾花的 150 条样本就构成了一个数据集。

机器学习中的几个基本概念

- 模型：机器学习中的“模型”是运行在数据上的机器学习算法的输出。模型表示机器学习算法所学到的内容，通常用函数 $h(x)$ 来描述。
- 损失函数：损失函数通常用来评价模型的预测值和真实值不一样的程度。
- 经验损失：将模型在训练数据上的平均损失称为经验损失。在理想情况下，一个监督式学习算法应该选择期望损失最小的模型，但通常是做不到的，当训练数据的规模足够大的时候，经验损失能够很好的近似期望损失。
- 模型假设：在不对模型作任何限制时，一个模型过多地拟合训练数据会影响到模型的可推广性，即泛化能力。因此，通常需要根据通过对数据的观察以及对问题背景的理解，对模型的结构做出合理的假设（例如线性模型），从而降低过度拟合。

正式一点的定义

具备特征的对象是机器学习问题的一个基本单位，每个对象可能有各种各样的特征，所以数学上用向量来表示全体对象特征。

- 特征组：在一个监督式学习问题中，将每个对象的 n 个特征构成的向量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$ 称为该对象的特征组。设 $\boldsymbol{X} \subseteq \mathbb{R}^n$ 是特征组的所有可能取值构成的集合，称 $\boldsymbol{X}$ 为样本空间。
- 标签：在回归问题中，训练数据含有一个数值标签 $\boldsymbol{y} \in \mathbb{R}$；在 k 元分类问题中，训练数据含有一个向量标签 $y \in \{0,1\}^k$。设 $\boldsymbol{Y}$ 为全体可能的标签取值，称 $\boldsymbol{Y}$ 为标签空间。
- 特征分布：设 X 为样本空间。监督式学习假定特征组是由一个 X 上的概率分布 D 生成的。称 D 为特征分布。用 $x \sim D$ 表示 x 为依特征分布 D 的一个随机采样。
- 标签分布：设 $\boldsymbol{Y}$ 为标签空间。对任意的 $\boldsymbol{x} \in \boldsymbol{X}$，监督式学习假定存在一个由 $\boldsymbol{x}$ 决定的 $\boldsymbol{Y}$ 上的概率分布 $D_{\boldsymbol{x}}$，使得 $\boldsymbol{x}$ 对应的标签 $\boldsymbol{y}$ 服从 $D_{\boldsymbol{x}}$。称 $D_{\boldsymbol{x}}$ 为特征组 $\boldsymbol{x}$ 的标签分布。用 $\boldsymbol{y} \sim D_{\boldsymbol{x}}$ 表示 $\boldsymbol{y}$ 为依分布 D_x 随机生成的标签。

  监督式学习的任务是计算从样本空间 $\boldsymbol{X}$ 到标签空间 $\boldsymbol{Y}$ 的映射。这样的映射称为一个模型。
- 模型：设 $\boldsymbol{X}$ 为样本空间，$\boldsymbol{Y}$ 为标签空间，Φ 为全体 $\boldsymbol{X} \to \boldsymbol{Y}$ 的映射集合。称 Φ 为模型空间。任意模型空间中的映射 $h \in \Phi$ 都称为一个模型。
- 监督式学习任务：在一个监督式学习问题中，给定样本空间 $\boldsymbol{X}$、标签空间 $\boldsymbol{Y}$、未知的特征分布 D 与标签分布 $\{D_{\boldsymbol{x}} | \boldsymbol{x} \in \boldsymbol{X}\}$，监督式学习任务是计算一个模型 $h : \boldsymbol{X} \to \boldsymbol{Y}$，并对任意特征组 $\boldsymbol{x}$、以 $h(\boldsymbol{x})$ 作为对标签期望值 $E_{\boldsymbol{y} \sim D_{\boldsymbol{x}}}[\boldsymbol{y}]$ 的预测。也将 $h(\boldsymbol{x})$ 简称为对 $\boldsymbol{x}$ 的标签的预测。
- 损失函数：设 Y 为标签空间。损失函数是一个从 $Y \times Y$ 映射到正实数的函数 $\ell : Y \times Y \to \mathbb{R}^+$，并且要求其具备如下性质：对任意 $y \in Y$，有 $\ell(y, y) = 0$。损失函数有两个参数：第一个参数值表示标签的真实值，第二个参数表示标签的预测值。损失函数用于度量标签的真实值和预测值的误差。

从理论上来说，损失函数的期望就是对模型效果的度量

- 期望损失：给定样本空间 X、特征分布 D、标签分布 $\{D_x : x \in X\}$ 以及损失函数 ℓ。对任意模型 h，定义 $L_E(h) = E_{x\sim D, y\sim D_x}[\ell(h(x), y)]$ 为模型 h 的期望损失。在实际应用中，由于特征分布 D 与标签分布 $\{D_x : x \in X\}$ 是未知的，所以无法直接计算期望损失，由此引入测试数据的概念。
- 测试数据与模型度量：在一个监督学习问题中，给定样本空间 X、标签空间 Y、未知的特征分布 D 与标签分布 $\{D_x : x \in X\}$。假定一个监督式学习算法输出模型 h。给定一组数据 $T = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots, (\boldsymbol{x}^{(t)}, y^{(t)})\}$。其中，$x^{(1)}, x^{(2)}, \cdots, x^{(t)} \sim D$ 为 X 中 t 个依特征分布 D 的独立采样，并且对任意 $1 \leq i \leq t$ 有 $y^{(i)} \sim D_x^{(i)}$。将 T 称为测试数据。用 h 在测试数据 T 上的平均损失 $L_T(h) = \frac{1}{t}\sum_{i=1}^{t} \ell(h(x^{(i)}), y^{(i)})$ 作为模型 h 效果的度量。测试数据是用来模拟期望损失的，当测试数据的规模足够大时，概率论中的 Hoeffding 不等式保证了测试损失集的平均损失能良好地近似期望损失。

无约束经验损失最小化算法因其模型的选择不受任何约束，可以输出模型空间 Φ 中的任何模型，因此由拉格朗日插值法知道，它可以精确地拟合训练数据。当一个模型过多地拟合训练数据中的特例而影响了它的可推广性时，就认为该模型是过度拟合的。因此，如果对模型的结构不做任何假设而一味优化算法在训练数据上的损失，则无法避免过度拟合，由此引入一个非常重要的概念–**模型假设：通过对训练数据的观察以及对问题背景的理解，可以对模型的结构做出合理的假设，从而降低过度拟合。**

- 训练数据与经验损失：给定损失函数 ℓ 以及一组数据 $S = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots, (\boldsymbol{x}^{(m)}, y^{(m)})\}$。其中，$x^{(1)}, x^{(2)}, \cdots, x^{(m)} \sim D$ 为 X 中 m 个依特征分布 D 的独立采样，并且对任意 $1 \leq i \leq t$ 有 $y^{(i)} \sim D_x^{(i)}$。将 S 称为训练数据。将 h 在 S 所有数据的平均损失称为 h 的经验损失，用 $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x^{(i)}), y^{(i)})$ 表示。当训练数据的规模足够大时，概率论中的 Hoeffding 不等式保证了经验损失能良好地近似期望损失。

因为监督式学习假设标签是一个随机变量，所以取值总与期望有一定的偏差。由于模型预测的是标签的期望，所以无须与训练数据标签完全拟合。同时，标签不是无规律地出现，而是服从某个未知的标签分布，因此对标签分布或模型结构做出恰当假设是一个合理的方法。合理选择模型假设是算法设计者经验的重要体现。

- 模型假设：模型空间 Φ 的任意一个子集 H 都称为一个模型假设，一个带有模型假设的经验损失最小化算法的任务是计算在设定的模型假设中经验损失最小的那个模型。

算法 **1** 带模型假设的经验损失最小化算法架构

给定样本空间 X、标签空间 Y 以及损失函数 $\ell : Y \times Y \to \mathbb{R}^+$。取定模型假设 H

Input: m 条训练数据 $S = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots, (\boldsymbol{x}^{(m)}, y^{(m)})\}$

Output: $h_S = \arg\min_{h \in H} L_S(h)$

根据具体问题，选择具体的模型假设与损失函数，可得到相应的算法。

模型假设、损失函数、目标函数的关系

基于式(2.7)可以提出如下的模型假设：取定 $\boldsymbol{w}=(w_1,w_2)\in\mathbb{R}^2$ 以及 $b\in\mathbb{R}$，定义

模型假设

$$h_{\boldsymbol{w},b}(\boldsymbol{x})=\text{Sign}(\langle\boldsymbol{w},\boldsymbol{x}\rangle+b) \tag{2.9}$$

并定义模型假设为 $H=\{h_{\boldsymbol{w},b}:\boldsymbol{w}\in\mathbb{R}^2,b\in\mathbb{R}\}$。

由于本问题的目标是准确地区分正负采样，可以采用例 2.1 中的 0-1 损失函数。通过计算不难发现，模型 $h_{\boldsymbol{w},b}$ 在训练数据 $(\boldsymbol{x}^{(i)},y^{(i)})$ 上的 0-1 损失有式(2.10)中的形式：

损失函数

$$l(h_{\boldsymbol{w},b}(\boldsymbol{x}^{(i)}),y^{(i)})=\frac{1-y^{(i)}\,\text{Sign}(\langle\boldsymbol{w},\boldsymbol{x}^{(i)}\rangle+b)}{2} \tag{2.10}$$

可见山鸢尾预测问题的经验损失最小化算法的目标应当为

$$\min_{\boldsymbol{w},b}\frac{1}{m}\sum_{i=1}^{m}\frac{1-y^{(i)}\,\text{Sign}(\langle\boldsymbol{w},\boldsymbol{x}^{(i)}\rangle+b)}{2} \tag{2.11}$$

式(2.11)等价于如下优化问题：

目标函数

$$\max_{\boldsymbol{w},b}\frac{1}{m}\sum_{i=1}^{m}y^{(i)}\,\text{Sign}(\langle\boldsymbol{w},\boldsymbol{x}^{(i)}\rangle+b) \tag{2.12}$$

将一条能够准确分离正采样与负采样的直线称为分离直线。如果直线 $\langle\boldsymbol{w},\boldsymbol{x}\rangle+b$ 是一条分离直线，则式(2.12)达到最大值 1。由此可见，在山鸢尾识别问题中，经验损失最小化任务实际上就是计算训练数据的分离直线。

图 3: 模型假设、损失函数、目标函数的关系