机器学习习方法

李航著

Stephen CUI

April 5, 2023

Contents

Part I

监督学习

Chapter 1

机器学习及监督学习概论

1.1 机器学习

机器学习的特点

Definition 1 (机器学习) 机器学习（*machine learning*）是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。机器学习也称为统计机器学习 *(statistical machine learning)*。

机器学习的对象

机器学习关于数据的基本假设是同类数据具有一定的统计规律性，这是机器学习的前提。这里的同类数据是指具有某种共同性质的数据。

机器学习的方法

实现机器学习方法的步骤如下：

1. 得到一个有限的训练数据集合；
2. 确定包含所有可能的模型的假设空间，即学习模型的集合(model)；
3. 确定模型选择的准则，即学习的策略(strategy)；
4. 实现求解最优模型的算法，即学习的算法(algorithm)；
5. 通过学习方法选择最优模型；
6. 利用学习的最优模型对新数据进行预测或分析。

1.2 机器学习的分类

1.2.1 基本分类

监督学习

Definition 2 (监督学习) 监督学习（*supervised learning*）是指从标注数据中学习预测模型的机器学习问题。标注数据表示输入输出的对应关系，预测模型对给定的输入产生相应的输出。监督学习的本质是学习输入到输出的映射的统计规律。

在监督学习中，将输入与输出所有可能取值的集合分别称为输入空间（input space）与输出空间（output space）。输入空间与输出空间可以是有限元素的集合，也可以是整个欧氏空间。每个具

体的输入是一个实例（instance），通常由特征向量（feature vector）表示。这时，所有特征向量存在的空间称为特征空间（feature space）。有时假设输入空间与特征空间为相同的空间，对它们不予区分；有时假设输入空间与特征空间为不同的空间，将实例从输入空间映射到特征空间。模型实际上都是定义在特征空间上的。

监督学习假设输入与输出的随机变量 X 和 Y 遵循联合概率分布 $P(X,Y)$。$P(X,Y)$ 表示分布函数或分布密度函数。

监督学习的目的在于学习一个由输入到输出的映射，这一映射由模型来表示。监督学习的模型可以是概率模型或非概率模型，由条件概率分布 $P(Y|X)$或决策函数（decision function）$Y = f(X)$ 表示，随具体学习方法而定。对具体的输入进行相应的输出预测时，写作 $P(y|x)$或 $y = f(x)$。

无监督学习

Definition 3 (无监督学习) *无监督学习（unsupervised learning）是指从无标注数据中学习预测模型的机器学习问题。无标注数据是自然得到的数据，预测模型表示数据的类别、转换或概率。无监督学习的本质是学习数据中的统计规律或潜在结构。*

假设 $\mathcal{X}$ 是输入空间，$\mathcal{Z}$ 是隐式结构空间。要学习的模型可以表示为函数 $z = g(x)$、条件概率分布 $P(z|x)$ 或者条件概率分布 $P(x|z)$ 的形式，其中 $x \in \mathcal{X}$ 是输入，$z \in \mathcal{Z}$ 是输出。包含所有可能的模型的集合称为假设空间。无监督学习旨在从假设空间中选出在给定评价标准下的最优模型。

强化学习

Definition 4 (强化学习) *强化学习（reinforcement learning）是指智能系统在与环境的连续互动中学习最优行为策略的机器学习问题。假设智能系统与环境的互动基于马尔可夫决策过程（Markov decision process），智能系统能观测到的是与环境互动得到的数据序列。强化学习的本质是学习最优的序贯决策。*

半监督学习与主动学习

Definition 5 (半监督学习) *半监督学习（semi-supervised learning）是指利用标注数据和未标注数据学习预测模型的机器学习问题。通常有少量标注数据、大量未标注数据。半监督学习旨在利用未标注数据中的信息，辅助标注数据，进行监督学习，以较低的成本达到较好的学习效果。*

Definition 6 (主动学习) *主动学习（active learning）是指机器不断主动给出实例让教师进行标注，然后利用标注数据学习预测模型的机器学习问题。通常的监督学习使用给定的标注数据，往往是随机得到的，可以看作是“被动学习”，主动学习的目标是找出对学习最有帮助的实例让教师标注，以较小的标注代价达到较好的学习效果。*

1.2.2 按模型分类

概率模型与非概率模型

机器学习的模型可以分为概率模型（probabilistic model）和非概率模型（non-probabilistic model）或者确定性模型（deterministic model）。

条件概率分布 $P(y|x)$ 和函数 $y = f(x)$ 可以相互转化（条件概率分布 $P(z|x)$ 和函数 $z = g(x)$ 同样可以）。具体地，条件概率分布最大化后得到函数，函数归一化后得到条件概率分布。所以，概率模型和非概率模型的区别不在于输入与输出之间的映射关系，而在于模型的内在结构。概率模型通常可以表示为联合概率分布的形式，其中的变量表示输入、输出、隐变量甚至参数。而非概率模型不一定存在这样的联合概率分布。

线性模型与非线性模型

机器学习模型，特别是非概率模型，可以分为线性模型（linear model）和非线性模型（non-linear model）。如果函数 $y = f(x)$ 或 $z = g(x)$ 是线性函数，则称模型是线性模型，否则称模型是非线性模型。

参数化模型与非参数化模型

机器学习模型又可以分为参数化模型（parametric model）和非参数化模型（non-parametric model）。参数化模型假设模型参数的维度固定，模型可以由有限维参数完全刻画；非参数化模型假设模型参数的维度不固定或者说无穷大，随着训练数据量的增加而不断增大。

1.2.3 按算法分类

机器学习根据算法，可以分为在线学习（online learning）与批量学习（batch learning）。在线学习是指每次接受一个样本，进行预测，之后学习模型，并不断重复该操作的机器学习。与之对应，批量学习一次接受所有数据，学习模型，之后进行预测。

在线学习中，学习和预测在一个系统，每次接受一个输入 x_t，用已有模型给出预测 $\hat{f}(x_t)$，之后得到相应的反馈，即该输入对应的输出 y_t；系统用损失函数计算两者的差异，更新模型，并不断重复以上操作。

1.2.4 按技巧分类

贝叶斯学习

贝叶斯学习（Bayesian learning）又称为贝叶斯推理（Bayesian inference），是统计学、机器学习中重要的方法。其主要想法是：在概率模型的学习和推理中，利用贝叶斯定理，计算在给定数据条件下模型的条件概率，即后验概率，并应用这个原理进行模型的估计，以及对数据的预测。

核方法

核方法（kernel method）是使用核函数表示和学习非线性模型的一种机器学习方法，可以用于监督学习和无监督学习。有一些线性模型的学习方法基于相似度计算，更具体地，向量内积计算。核方法可以把它们扩展到非线性模型的学习，使其应用范围更广泛。

把线性模型扩展到非线性模型，直接的做法是显式地定义从输入空间（低维空间）到特征空间（高维空间）的映射，在特征空间中进行内积计算。

核方法的技巧在于不显式地定义这个映射，而是直接定义核函数，即映射之后在特征空间的内积。这样可以简化计算，达到同样的效果。

假设 x_1 和 x_2 是输入空间的任意两个实例（向量），其内积是 $\langle x_1, x_2 \rangle$。假设从输入空间到特征空间的映射是 φ，于是 x_1 和 x_2 在特征空间的映像是 $\varphi(x_1)$ 和 $\varphi(x_2)$，其内积是 $\langle \varphi(x_1), \varphi(x_2) \rangle$。核方法直接在输入空间中定义核函数 $K(x_1, x_2)$，使其满足 $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$。表示定理给出核函数技巧成立的充要条件。

1.3 机器学习方法三要素

1.3.1 模型

模型的假设空间（hypothesis space）包含所有可能的条件概率分布或决策函数。假设空间中的模型一般有无穷多个。

1.3.2 策略

有了模型的假设空间，机器学习接着需要考虑的是按照什么样的准则学习或选择最优的模型。机器学习的目标在于从假设空间中选取最优模型。

失函数度量模型一次预测的好坏，风险函数度量平均意义下模型预测的好坏。

损失函数和风险函数

损失函数是 $f(X)$ 和 Y 的非负实值函数，记作 $L(Y, f(X))$。

- 0-1 损失函数（0-1 loss function），主要针对分类问题

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases} \tag{1.1}$$

- 平方损失函数（quadratic loss function），主要针对回归问题

$$L(Y, f(X)) = (Y - f(X))^2 \tag{1.2}$$

- 绝对损失函数（absolute loss function），主要针对回归问题

$$L(Y, f(X)) = |Y - f(X)| \tag{1.3}$$

- 对数损失函数（logarithmic loss function）或对数似然损失函数（log-likelihood loss function），主要针对概率问题

$$L(Y, P(Y|X)) = -logP(Y|X) \tag{1.4}$$

由于模型的输入、输出 (X, Y) 是随机变量，遵循联合分布 $P(X, Y)$，所以损失函数的期望是

$$\begin{aligned} R_{exp}(f) &= E_P[L(Y, f(X))] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x))P(x, y)dxdy \end{aligned} \tag{1.5}$$

由于联合分布 $P(X, Y)$ 是未知的，$R_{exp}(f)$ 不能直接计算。

给定一个训练数据集

$$T = (x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$$

模型 $f(X)$ 关于训练数据集的平均损失称为经验风险（empirical risk）或经验损失（empirical loss），记作 R_{emp}：

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.6}$$

经验风险最小化与结构风险最小化在假设空间、损失函数以及训练数据集确定的情况下，经验风险函数式 **??** 就可以确定。经验风险最小化（**empirical risk minimization，ERM**）的策略认为，经验风险最小的模型是最优的模型。根据这一策略，按照经验风险最小化求最优模型就是求解最优化问题：

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.7}$$

结构风险最小化（**structural risk minimization，SRM**）是为了防止过拟合而提出来的策略。结构风险最小化等价于正则化（regularization）。结构风险在经验风险上加上表示模型复杂度的

正则化项（regularizer）或罚项（penalty term）。在假设空间、损失函数以及训练数据集确定的情况下，结构风险的定义是：

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.8}$$

结构风险最小化的策略认为结构风险最小的模型是最优的模型，所以求最优模型就是求解最优化问题：

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.9}$$

1.3.3 算法

算法是指学习模型的具体计算方法。机器学习基于训练数据集，根据学习策略，从假设空间中选择最优模型，最后需要考虑用什么样的计算方法求解最优模型。这时，机器学习问题归结为最优化问题，机器学习的算法成为求解最优化问题的算法。

1.4 模型评估与模型选择

1.4.1 训练误差与测试误差

注意，机器学习方法具体采用的损失函数未必是评估时使用的损失函数。当然，让两者一致是比较理想的。

1.4.2 过拟合与模型选择

Definition 7 (过拟合) 过拟合是指学习时选择的模型所包含的参数过多，以至出现这一模型对已知数据预测得很好，但对未知数据预测得很差得现象。

在学习时就要防止过拟合，进行最优的模型选择，即选择复杂度适当的模型。

1.5 正则化与交叉验证

1.5.1 正则化

Definition 8 (正则化) 正则化是结构风险最小化策略的实现，是在经验风险上加一个正则化项（*regularizer*）或罚项（*penalty term*），即 ??。正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大。比如，正则化项可以是模型参数向量的范数。

正则化项可以取不同的形式。例如，在回归问题中，损失函数是平方损失，正则化项可以是参数向量的 L_2 范数，也可以是参数向量的 L_1 范数。

正则化符合奥卡姆剃刀（Occam's razor）原理。奥卡姆剃刀原理应用于模型选择时变为以下想法：在所有可能选择的模型中，能够很好地解释已知数据并且十分简单才是最好的模型，也就是应该选择的模型。

1.5.2 交叉验证

样本量多的话可以选择训练、验证和测试。

样本量少的话可以选择简单交叉验证、K 折交叉验证、留一交叉验证。

1.6 泛化能力

1.6.1 泛化误差

Definition 9 (泛化误差) 如果学到的模型是 $\hat{f}$，那么用这个模型对未知数据预测的误差即为泛化误差（*generalization error*）

$$\begin{aligned} R_{exp}(\hat{f}) &= E_P[L(Y, \hat{f}(X))] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{f}(x)) P(x, y) dxdy \end{aligned} \tag{1.10}$$

事实上，泛化误差就是所学习到的模型的期望风险。

1.6.2 泛化误差上界

学习方法的泛化能力分析往往是通过研究泛化误差的概率上界进行的，简称为泛化误差上界（generalization error bound）。泛化误差上界通常具有以下性质：

- 它是样本容量的函数，当样本容量增加时，泛化误差上界趋于 0；
- 它是假设空间容量（capacity）的函数，假设空间容量越大，模型就越难学，泛化误差上界就越大。

Theorem 1 (泛化误差上界) 对于二分类问题，当假设空间时有限个函数的集合$\mathcal{F} = \{f_1, f_2, \cdots, f_d\}$时，对$\forall f \in \mathcal{F}$，至少一概率$1-\delta$，$0 < \delta < 1$，以下不等式成立：

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta) \tag{1.11}$$

其中，

$$\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N}(\log d + \log \frac{1}{\delta})} \tag{1.12}$$

1.7 生成模型和判别模型

Definition 10 (生成模型) 有数据学习联合分布概率$P(X, Y)$，然后求出$P(Y|X)$作为预测模型，即生成模型*(Generatice Model)*:

$$P(Y|X) = \frac{P(X, Y)}{P(X)} \tag{1.13}$$

典型的生成模型有朴素贝叶斯法和隐马尔可夫模型。

Definition 11 (判别方法) 由数据直接学习决策函数 $f(X)$ 或者条件概率分布 $P(Y|X)$ 作为预测的模型。

1.8 监督学习的应用

1.8.1 分类问题

许多机器学习方法可以用于分类，包括 k 近邻法、感知机、朴素贝叶斯法、决策树、决策列表、逻辑斯谛回归模型、支持向量机、提升方法、贝叶斯网络、神经网络、Winnow 等。

Table 1.1: 生成模型与判别模型特点

生成模型	判别模型
所需数据量较大	所需样本的数量少于生成模型
可还原联合概率密度分布 $P(X,Y)$	可直接面对预测，准确率更高
收敛速度更快	可简化学习问题（不需要面面俱到，仅考虑数据的特性）
能反映同类数据本身的相似度	不可以反映数据本身的特性
隐变量存在时，仍可用生成模型	

1.8.2 标注问题

标注问题分为学习和标注两个过程，首先给定一个训练数据集，

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

这里$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T, i = 1, 2, \cdots, N$是输入观测序列，$y_i = (y_i^{(1)}, y_i^{(2)}, \cdots, y_i^{(n)})^T$是相应的输入输出标记序列，$n$是序列的长度，对不同样本可以用不同的值。学习系统基于训练数据集构建一个模型，表示为条件概率分布：

$$P(Y^{(1)}, Y^{(1)}, \cdots, Y^{(n)} | X^{(1)}, X^{(1)}, \cdots, X^{(1)}) \tag{1.14}$$

标注常用的机器学习方法有隐马尔可夫模型、条件随机场。

标注问题在信息抽取、自然语言处理等领域被广泛应用，是这些领域的基本问题。例如，自然语言处理中的词性标注（part of speech tagging）就是一个典型的标注问题：给定一个由单词组成的句子，对这个句子中的每一个单词进行词性标注，即对一个单词序列预测其对应的词性标记序列。

举一个信息抽取的例子。从英文文章中抽取基本名词短语（base noun phrase）。为此，要对文章进行标注。英文单词是一个观测，英文句子是一个观测序列，标记表示名词短语的“开始”、“结束”或“其他”（分别以 B，E，O 表示），标记序列表示英文句子中基本名词短语的所在位置。信息抽取时，将标记“开始”到标记“结束”的单词作为名词短语。例如，给出以下的观测序列，即英文句子，标注系统产生相应的标记序列，即给出句子中的基本名词短语。

输入：At Microsoft Research, we have an insatiable curiosity and the desire to create new technology that will help define the computing experience.

输出：At/O Microsoft/B Research/E, we/O have/O an/O insatiable/B curiosity/E and/ O the/O desire/BE to/O create/O new/B technology/E that/O will/O help/O define/ O the/O computing/B experience/E.

1.8.3 回归问题

回归模型正是表示从输入变量到输出变量之间映射的函数。回归问题学习等价于函数拟合：选择一条函数曲线使其很好地你和已知数据并很好地预测未知数据。

Chapter 2

朴素贝叶斯法

贝叶斯思维：先验概率（主观判断）←调整因子（添加新信息）→后验概率（最终结论）

朴素贝叶斯（Naive Bayes）法是基于贝叶斯定理与**特征条件独立性假设**的分类方法[1]。对于给定的数据集，首先基于特征条件独立假设学习输入输出的联合概率分布，然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。

2.1 朴素贝叶斯法的学习与分类

[1] 朴素贝叶斯法与贝叶斯估计是不同的概念