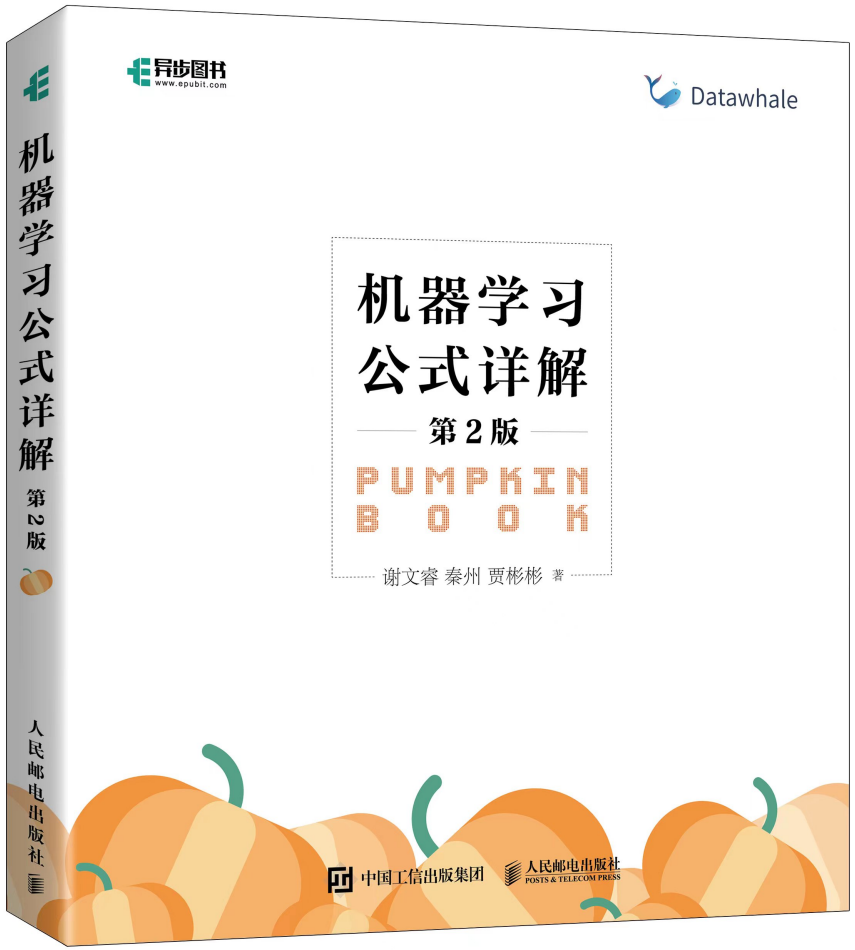

《机器学习公式详解》（南瓜书）

第1章 绪论

本节主讲：谢文睿

简单说明

- 为什么补一二两章?
  - 应广大学习者所需！为学习者服务！
- 第1版和第2版的区别?
  - 解释的更全面、更基础！盖泡面更严实！
- 本套视频能同时兼容第1版和第2版么?
  - 能的，因为第2版新增的内容我都会补录上！缝缝补补又三年！

本节大纲

西瓜书对应章节：第1章

- 机器学习三观：what、why、how
- 假设空间和版本空间：通过一个例子讲解，同时引出基本术语
- 基本术语
- 归纳偏好
- 数据决定模型的上限，而算法则是让模型无限逼近上限

机器学习三观

- what：什么是机器学习?
- why：为什么要学习机器学习?
- how：怎样学机器学习?

机器学习三观

what：什么是机器学习?

研究关于“学习算法”（一类能从数据中学习出其背后潜在规律的算法）的一门学科

PS：深度学习指的是神经网络那一类学习算法，因此是机器学习的子集

机器学习三观

why：为什么要学习机器学习?

- 从事机器学习理论的研究
- 从事机器学习系统的开发
- 将机器学习中的算法迁移应用到自己的研究领域
- 从事AI应用方向的研究：自然语言处理（NLP）、计算机视觉（CV）、推荐系统等

机器学习三观

how：怎样学机器学习？

- 从事纯机器学习理论的研究：
  - 本课程讲的所有内容都要听懂
  - 进一步可阅读周志华老师的《机器学习理论导引》
  - 机器学习还很年轻，当前正处于工程领先理论阶段，还有很多未解之谜
- 从事机器学习系统的开发：
  - 进阶学习：https://ucbrise.github.io/cs294-ai-sys-sp22 、https://openmlsys.github.io
- 将机器学习中的算法迁移应用到自己的研究领域
  - 看下一页
- 从事AI应用方向的研究：自然语言处理（NLP）、计算机视觉（CV）、推荐系统等
  - 看下一页

机器学习三观

学习方法：

1. 基础数学知识自己补：
    - 时间充裕：打好高等数学、线性代数、概率论基础
    - 时间紧张：直接开啃，哪里不懂补哪里
        - 推荐课程：张宇考研数学基础班视频课
2. 高阶数学知识由南瓜书+本套视频承包

学习程度：

1. 在学的过程中能看懂每一步推导过程即可，不用达到熟稔于心的地步
2. 会调scikit-learn库即可，不用自行实现
3. 时间紧张的同学，学完前5章即可开始学深度学习

假设空间和版本空间

举个栗子：假设现已收集到某地区近几年的房价和学校数量数据，希望利用收集到的数据训练出能通过学校数量预测房价的模型，具体收集到的数据如下表所示：

表 1-1 房价预测

年份	学校数量	房价
2020	1 所	1 万/m^2
2021	2 所	4 万/m^2

假设空间：一元一次函数，算法：线性回归，模型：$y = 3x - 2$
假设空间：一元二次函数，算法：多项式回归，模型：$y = x^2$

所有能够拟合训练集的模型（假设）构成的集合称为“版本空间”。

基本术语

“算法”是指从数据中学得“模型”的具体方法，例如后续章节中将会讲述的线性回归、对数几率回归、决策树等。

“算法”产出的结果称为“模型”，通常是具体的函数或者可抽象地看作为函数，例如一元线性回归算法产出的模型即为形如$f(x) = wx + b$的一元一次函数。

不过由于严格区分这两者的意义不大，因此多数文献和资料会将其混用，当遇到这两个概念时，其具体指代根据上下文判断即可。

基本术语

样本：也称为“示例”，是关于一个事件或对象的描述。

一个“色泽青绿，根蒂蜷缩，敲声清脆”的西瓜用向量来表示即为

$$\boldsymbol{x} = (\text{青绿}; \text{蜷缩}; \text{清脆})$$

向量中的各个维度称为“特征”或者“属性”

向量中的元素用分号“;”分隔时表示此向量为列向量，用逗号“,”分隔时表示为行向量

解释一下“特征工程”

基本术语

标记：机器学习的本质就是在学习样本在某个方面的表现是否存在潜在的规律，我们称该方面的信息为“标记”

标记通常也看作为样本的一部分，因此，一个完整的样本通常表示为$(\boldsymbol{x}, y)$

一条西瓜样本：$\boldsymbol{x} = (\text{青绿}; \text{蜷缩}; \text{清脆}), y = \text{好瓜}$
一条房价样本：$\boldsymbol{x} = (1\text{所}), y = 1\text{万}/m^2$

基本术语

样本空间：也称为“输入空间”或“属性空间”。由于样本采用的是标明各个特征取值的“特征向量”来进行表示，根据线性代数的知识可知，有向量便会有向量所在的空间，因此称表示样本的特征向量所在的空间为样本空间，通常用花式大写的$\mathcal{X}$表示

标记空间：标记所在的空间称为“标记空间”或“输出空间”，数学表示为花式大写的$\mathcal{Y}$

基本术语

根据标记的取值类型不同，可将机器学习任务分为以下两类：

- 当标记取值为离散型时，称此类任务为“分类”，例如学习西瓜是好瓜还是坏瓜、学习猫的图片是白猫还是黑猫等。当分类的类别只有两个时，称此类任务为“二分类”，通常称其中一个为“正类”，另一个为“反类”或“负类”；当分类的类别超过两个时，称此类任务为“多分类”。由于标记也属于样本的一部分，通常也需要参与运算，因此也需要将其数值化，例如对于二分类任务，通常将正类记为1，反类记为0，即$\mathcal{Y}=\{0,1\}$。这只是一般默认的做法，具体标记该如何数值化可根据具体机器学习算法进行相应地调整，例如第6章的支持向量机算法则采用的是$\mathcal{Y}=\{-1,+1\}$
- 当标记取值为连续型时，称此类任务为“回归”，例如学习预测西瓜的成熟度、学习预测未来的房价等。由于是连续型，因此标记的所有可能取值无法直接罗列，通常只有取值范围，回归任务的标记取值范围通常是整个实数域$\mathbb{R}$，即$\mathcal{Y}=\mathbb{R}$。

基本术语

根据是否有用到标记信息，可将机器学习任务分为以下两类：

- 模型训练阶段有用到标记信息时，称此类任务为“监督学习”，例如第3章的线性模型
- 在模型训练阶段没用到标记信息时，称此类任务为“无监督学习”，例如第9章的聚类

基本术语

无论是分类还是回归，机器学习算法最终学得的模型都可以抽象地看作为以样本$\boldsymbol{x}$为自变量，标记y为因变量的函数$y = f(\boldsymbol{x})$，即一个从输入空间$\mathcal{X}$到输出空间$\mathcal{Y}$的映射。

例如在学习西瓜的好坏时，机器学习算法学得的模型可看作为一个函数$f(\boldsymbol{x})$，给定任意一个西瓜样本$\boldsymbol{x}_i = (\text{青绿}; \text{蜷缩}; \text{清脆})$，将其输入进函数即可计算得到一个输出$y_i = f(\boldsymbol{x}_i)$，此时得到的$y_i$便是模型给出的预测结果，当$y_i$取值为$1$时表明模型认为西瓜$\boldsymbol{x}_i$是好瓜，当$y_i$取值为$0$时表明模型认为西瓜$\boldsymbol{x}_i$是坏瓜。

预测房价的栗子同理

基本术语

数据集：数据集通常用集合来表示，令集合$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$表示包含$m$个样本的数据集，一般同一份数据集中的每个样本都含有相同个数的特征，假设此数据集中的每个样本都含有d个特征，则第i个样本的数学表示为d维向量：$\boldsymbol{x}_i = \left(x_{i1}; x_{i2}; \ldots; x_{id}\right)$，其中$x_{ij}$表示样本$\boldsymbol{x}_i$在第$j$个属性上的取值。

基本术语

模型：机器学习的一般流程如下：首先收集若干样本（假设此时有100个），然后将其分为训练样本（80个）和测试样本（20个），其中80个训练样本构成的集合称为“训练集”，20个测试样本构成的集合称为“测试集”，接着选用某个机器学习算法，让其在训练集上进行“学习”（或称为“训练”），然后产出得到“模型”（或称为“学习器”），最后用测试集来测试模型的效果。执行以上流程时，表示我们已经默认样本的背后是存在某种潜在的规律，我们称这种潜在的规律为“真相”或者“真实”，例如样本是一堆好西瓜和坏西瓜时，我们默认的便是好西瓜和坏西瓜背后必然存在某种规律能将其区分开。当我们应用某个机器学习算法来学习时，产出得到的模型便是该算法所找到的它自己认为的规律，由于该规律通常并不一定就是所谓的真相，所以也将其称为“假设”。通常机器学习算法都有可配置的参数，同一个机器学习算法，使用不同的参数配置或者不同的训练集，训练得到的模型通常都不同。

基本术语

泛化：由于机器学习的目标是根据已知来对未知做出尽可能准确的判断，因此对未知事物判断的准确与否才是衡量一个模型好坏的关键，我们称此为“泛化”能力。

分布：此处的“分布”指的是概率论中的概率分布，通常假设样本空间服从一个未知“分布”$\mathcal{D}$，而我们收集到的每个样本都是独立地从该分布中采样得到，即“独立同分布”。通常收集到的样本越多，越能从样本中反推出$\mathcal{D}$的信息，即越接近真相。

归纳偏好

在“房价预测”的例子中，当选用一元线性回归算法时，学得的模型是一元一次函数，当选用多项式回归算法时，学得的模型是一元二次函数，所以不同的机器学习算法有不同的偏好，我们称为“归纳偏好”。

对于当前房价预测这个例子来说，这两个算法学得的模型哪个更好呢？著名的“奥卡姆剃刀”原则认为“若有多个假设与观察一致，则选最简单的那个”，但是何为“简单”便见仁见智了，如果认为函数的幂次越低越简单，则此时一元线性回归算法更好，如果认为幂次越高越简单，则此时多项式回归算法更好，因此该方法其实并不“简单”，所以并不常用，而最常用的方法则是基于模型在测试集上的表现来评判模型之间的优劣。

归纳偏好

例如在房价预测问题中，通常会额外留有部分未参与模型训练的数据来对模型进行测试。假设此时额外留有1条数据：$(\text{年份}：2022\text{年}; \text{学校数量}：3\text{所}; \text{房价}：7\text{万}/m^2)$用于测试，模型$y = 3x - 2$的预测结果为$3 * 3 - 2 = 7$，预测正确，模型$y = x^2$的预测结果为$3^2 = 9$，预测错误，因此，在当前房价预测问题上，我们认为一元线性回归算法优于多项式回归算法。

机器学习算法之间没有绝对的优劣之分，只有是否适合当前待解决的问题之分，例如上述测试集中的数据如果改为$(\text{年份}：2022\text{年}; \text{学校数量}：3\text{所}; \text{房价}：9\text{万}/m^2)$则结论便逆转为多项式回归算法优于一元线性回归算法。

归纳偏好

没有免费的午餐定理（NFL）：众算法生而平等（详细推导参见南瓜书）

实际应用：哪个算法训出来的模型在测试集上表现好哪个算法就nb

数据决定模型的上限，而算法则是让模型无限逼近上限

数据决定模型效果的上限：其中数据是指从数据量和特征工程两个角度考虑。从数据量的角度来说，通常数据量越大模型效果越好，因为数据量大即表示累计的经验多，因此模型学习到的经验也多，自然表现效果越好。例如以上举例中如果训练集中含有相同颜色但根蒂不蜷缩的坏瓜，模型a学到真相的概率则也会增大；从特征工程的角度来说，通常对特征数值化越合理，特征收集越全越细致，模型效果通常越好，因为此时模型更易学得样本之间潜在的规律。例如学习区分亚洲人和非洲人时，此时样本即为人，在进行特征工程时，如果收集到每个样本的肤色特征，则其他特征例如年龄、身高和体重等便可省略，因为只需靠肤色这一个特征就足以区分亚洲人和非洲人。

算法则是让模型无限逼近上限：是指当数据相关的工作已准备充分时，接下来便可用各种可适用的算法从数据中学习其潜在的规律进而得到模型，不同的算法学习得到的模型效果自然有高低之分，效果越好则越逼近上限，即逼近真相。

预告

下一节：模型评估与选择

西瓜书对应章节：第2章

结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 扫描下方二维码关注公众号，然后发送【南瓜书】三个字获取机器人二维码
2. 添加机器人为微信好友，然后给机器人发送【南瓜书】三个字即可获取进群链接