Tiancheng Li's Notes on Statistics and Machine Learning

李天成的统计与机器学习笔记

Tiancheng Li

2024.8.9

PREFACE

在硕士阶段，我系统地学习了数据科学、统计方法、机器学习及深度学习相关的理论与实践。在学习过程中，我意识到每种算法背后都蕴含着严密的数学推导和逻辑原理。为此，我常常尝试推导这些算法的理论基础，并以自己的方式阐述推导过程。然而，由于知识积累是一个持续的过程，我对模型的推导和理解也在不断深化。遗憾的是，许多早期纸面记录的推导和证明随着时间的推移逐渐遗失。此外，我的知识来源较为多样，包括课堂学习、书籍、视频以及在线资源等。这些来源的讲解风格各异，有的偏重数学推导，有的偏重学术性描述，也有些倾向于口语化的讲解。这导致我在后续的学习过程中，回顾相关知识点时，难以高效、系统地获取所需信息。

为此，我希望通过这篇笔记系统性地整理和阐述我所学到的统计和机器学习知识。这不仅有助于我构建一个完整的知识体系，也能为日后查阅提供便利。正如爱因斯坦所说，“任何值得做的事，都值得慢慢做。”这份笔记是我持续思考和深入理解这些精妙理论的过程记录，也希望它能帮助我在未来更好地连接和应用这些知识。

This note mainly references the following classical works and textbooks: ”Machine Learning” by Zhihua Zhou, ”Statistical Learning Methods” by Hang Li, ”Deep Learning” by Ian Goodfellow et al., and ”Pattern Recognition and Machine Learning” by Christopher M. Bishop.

Contents

1 绪论

1.1 机器学习想法的产生

在人类产生文明到发明计算机之前，人的判断往往是根据以往的经验做出的。现在，我们希望做的事是用计算机来“学习”人类的经验。在计算机系统中，“经验”通常存储在“数据”中，这些数据本身只是对过往发生过的事件的客观记录。我们希望的从数据中提取出某种“规律”，用这些规律来解释我们未曾见过的新数据。因此机器学习所研究的主要内容是关于在计算机上从数据中产生模型的算法，即“学习算法”。

1.2 对机器学习的感性理解

1.2.1 数据集

=: 取值为

现在我们来从一个非常朴素的想法开始介绍一些专业术语。假设我们现在想要判断一个西瓜的甜不甜，第一步，先要观察一些西瓜并把观察结果记录下来。例如，
西瓜 1 的颜色 = 绿，西瓜藤形状 = 直，敲击声音 = 高音；
西瓜 2 的颜色 = 黑，西瓜藤形状 = 弯，敲击声音 = 低音；
西瓜 3 的颜色 = 白，西瓜藤形状 = 微曲，敲击声音 = 高音;

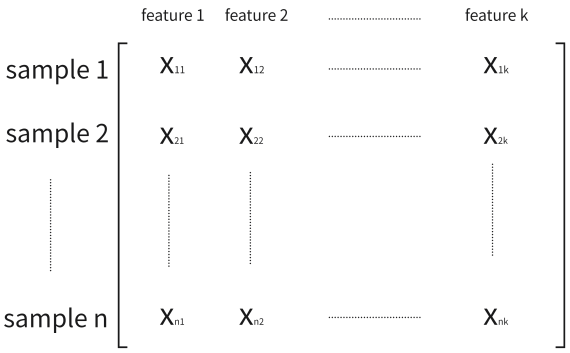
我们将所有观察到的数据的集合称为**数据集 (data set)**, 数据集在数学形式上可以直观的理解为一个矩阵：

	颜色	西瓜藤形状	敲击声音
西瓜1	绿	直	高音
西瓜2	黑	弯	低音
西瓜3	白	微曲	中音

这个矩阵中的每一行代表一个**样本 (sample)** 或**示例 (instance)**，在上面的例子中，每一行代表观察到的一个西瓜的数据；矩阵的每一列表示反映这个样本在某些方面的表现或性质的事项，称为**特征 (feature)** 或**属性 (attribute)**，每一个特征的具体取值称为**特征值 (feature value)**. 例如，在上面的数据集中，一共有三个特征，分别是颜色，西瓜藤形状以及敲击声音。西瓜 1 在颜色这个特征上的特征值是“绿”。

数据集矩阵中的一个样本的特征的总和 (即数据集矩阵中的一行) 称为**特征向量 (feature vector)**; 数据集矩阵中的所有特征以及特征值的取值范围构成了这个数据集的**特征空间 (feature space)** 或**样本空间 (sample space)** $\mathcal{X}$，每一个样本对应一个的**标签 (label)** y_i，所有标签的集合称为**标签空间，标记空间 (label space)** $\mathcal{Y}$, 需要注意的是，并不是所有的数据集都有标签，无监督学习的数据集中没有标签。

让我们用更一般的写法来描述一个数据集：令 $D = \{ x_1, x_2, ..., x_n \}$ 表示 D 这个数据集包含了 n 个样本; $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ 是数据集 D 中的第 i 个样本，也就是 k 维样本空间 $\mathcal{X}$ 中的一个向量，



$x_i \in \mathcal{X}$。其中，x_{ij} 是 x_i 在第 j 个特征 (第 j 维) 上的取值，k 称为样本 x_i 的维数。

在数据集带有标签的情况下，每一个样本可以表示为 (x_i,y_i), $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$。

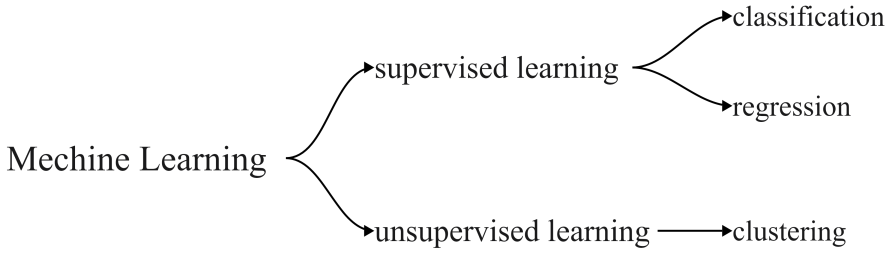
1.2.2 从数据集获得规律

我们认为自然界中存在一个真实的规律，或者模型，这个真实的模型是非数学的，我们永远也无法掌握的。因此我们希望通过一个数学模型来尽可能的模拟出这个真实的模型。这个过程及就是从数据集中获得规律的过程，叫做**学习 (learning)** 或**训练 (training)**。自然界中存在的真实模型叫做**真实 (ground-truth)**, 我们训练出来的数学模型叫做**假设 (hypothesis)**

在这个阶段，我们一般不用整个数据集来训练，因为数据是非常宝贵的，通常是从整个数据集中选择一部分数据，作为**训练集 (training set)**, 训练集中的每一个样本称为**训练样本**

1.2.3 机器学习的主要任务

在机器学习中，根据数据集有没有标签，可以将机器学习的任务分为两大类: **监督学习 (supervised learning)** 和**无监督学习 (unsupervised learning)**, 监督学习的意思是每一个样本有真正的标签对其进行“监督”，训练过程中把一个样本投入建立的数学模型，输出结果与这个样本的真实标签进行比对，如果数学模型的预测结果与真实标签不符，就修改模型参数，直到输出结果与真实标签一致。反之，无监督学习就是数据集中没有真实的标签对训练过程进行“监督”。



在监督学习中，又分为**分类 (classification)** 和**回归 (regression)**。分类和回归的主要区别在于标签空间是否是连续的。若标签空间中的标签是离散的，这类学习任务就是分类。简单的说，分类就是将输入的样本预测成不同的标签，建立一个从样本空间 $\mathcal{X}$ 到标签空间 $\mathcal{Y}$ 的映射 $f : \mathcal{X} \to \mathcal{Y}$，对于二分类任务，一般 $\mathcal{Y} = \{-1, +1\}$ 或 $\{0, 1\}$, 对于多分类任务，$|\mathcal{Y}| > 2$。比如判断图片中的人物判断性别，则是二分类问题，将一张照片中的形象判断为狗，猫，或者鸟，则是多分类问题。

回归任务的标签空间则是连续的，即 $\mathcal{Y} \in \mathbb{R}$，比如输入一些特征来预测房价，房价是一个连续的标

签。

无监督学习所使用的数据集是没有标签的，只有每个样本的特征，因此我们希望的是通过对这些特征的学习，将具有相同或相似特征的样本聚合成一个“**簇 (cluster)**”, 不同的簇之间意味着一些潜在概念的划分。

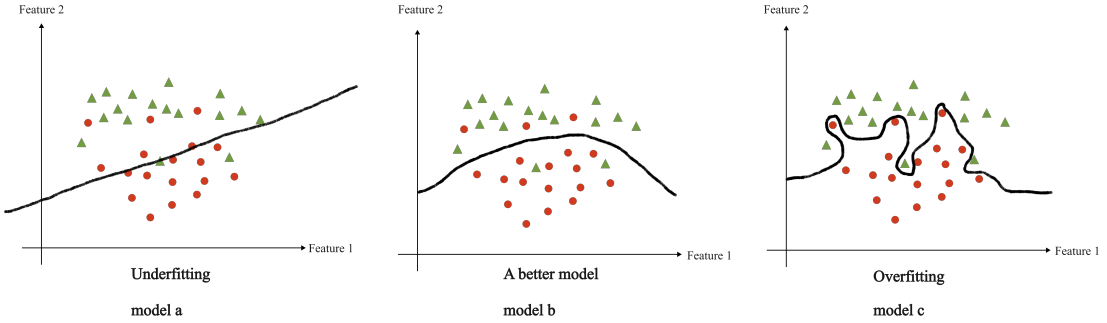
1.2.4 机器学习的目标

无论是分类，回归还是聚类，机器学习的目标是使训练好的模型能够在未见过的样本上表现良好。模型适应新样本的能力，称为**泛化 (generaliazation) 能力**。如何寻找到泛化能力最强的那个模型，我们需要弄清一些概念。

机器学习的训练阶段，可以理解为在这个数据集的假设空间中，寻找最优假设的过程。在模型框架确定的情况下 (就是我们确定了模型是 SVM, 决策树或者神经网络)，寻找最优的假设即寻找最优的数学模型参数。一个朴素的愿望是：我们希望，最终确定的数学模型可以在没有见过的新样本上依然可以正确的预测结果。这个朴素的愿望包含两个隐藏的条件：
1. 最终的数学模型可以在训练集中表现良好。
2. 最终的数学模型在没有见过的样本中依然表现良好。
让我们分析这两个要求。

在模型训练之后，我们会发现，不仅仅是唯一确定的一组参数可以将不同类别的样本点区分开，因为样本点中间不是紧密排列的。(如下图所示)。我们可以有多组甚至无数组参数将不同类别的样本点区分开。因此，第一个要求是容易满足的，即我们的数学模型很容易就可以在训练集上表现良好。我们只需要在训练阶段根据预测结果与真实标签差别不断地调整参数就可以满足要求。但是第二个要求是不容易达到的，即在没有见过的样本点上模型依然可以表现良好。为了达到这个要求，就需要对满足条件 1 的多组数学参数中选择最优的一组。

下图所示的是假设空间中表现较好的三组参数，然我们用上面的两个要求来选出最好的一组参数。Model a 的参数模型不能在训练集中表现良好，我们称之为**欠拟合 (Underfitting)**, 不满足条件 1，因此最先被淘汰。剩下的模型参数需要用条件 2 来区分。这里需要引入“**奥卡姆剃刀 (Occam's razor)**”原则，即“若有多个假设与观察一致，则选择最简单的那个”。model b 和 c 都可以把训练集中的样本点比较好的区分开，但是 model c 的规则肉眼可见的过于复杂，过于复杂带来的问题是，虽然它在训练集中的表现相当完美，正确率甚至达到了 100%, 但有很大的概率在未见过的样本中的正确率会非常的低，我们称之为**过拟合 (Overfitting)**。相比之下，model b 用比较简单的规则就实现了不错的效果。因此 model b 是假设空间中的最优假设。



需要注意的是，在实际的机器学习任务中，哪一种模型更好并不像上面图片中的规则那样清晰。

2 模型评估与选择

根据前面机器学习的目标，我们想要得到一个泛化能力最强的模型，接下来我们详细展开如何评估模型的泛化能力。

2.1 经验误差与过拟合

抛开机器学习，用最朴素的想法来想象一下，当我希望评价一个可以用来区分男性和女性的“黑箱”的时候，我们期待的结果是什么？我想是这样一个情况：当我把任意一种男性的照片，无论图像中的这个男人是长发，短发，戴不戴眼镜，有没有胡子，什么姿势，经过这个黑箱之后，黑箱给你的答案都是“男性”。这是我们想要的结果，因为这个黑箱满足了我们的预期：即预测的结果（分类器的预测）和我们的预期（实际的标签）是一致的，即预测正确。反过来当预测的结果和实际这个图片中的人物性别不同时，比如照片中是一个男性，但黑箱分类成了女性，则我们认为这个结果是错误的。

通过这个非常朴素的思想，我们可以引出在机器学习领域中，去评判一个分类器的好坏的一个重要标准。学习器的实际预测结果与样本的真实标签的之间的差异称为“**误差 (error)**”。在实际的机器学习中，预测结果和标签往往转化成数值或向量，通过作差来进行比较。学习器在训练集上的误差称为” **训练误差 (training error)** 或**经验误差 (empirical error)**” 。在新样本上的误差称为” **泛化误差 (generalization error)**”.

因此，在对一个学习器的训练或应用中，误差的数量就是评价一个模型好坏的最直接的标准，令 m 为总的样本数量，可以是整个数据集的样本数量，也可以是训练中的一个**批次 (batch)** 的样本数量。批次可以理解为一个完整的训练集是很大的，为了提高训练效率，我们往往把一个完整的训练集分解成若干个子集，这些子集就是“批次”。a 为 m 个样本通过模型之后预测错误的样本个数，则**错误率 (error rate)** 为

$$E = \frac{a}{m}$$

正确率、精度 (accuracy) $= 1 - \frac{a}{m}$。

过拟合是模型评估中永恒的问题，现在我尝试用错误率的视角来描述一下过拟合问题。在 1.2.4 中我总结了人们对一个表现良好的学习器的 2 个朴素的要求：最终的数学模型可以在训练集中表现良好；最终的数学模型在没有见过的样本中依然表现良好。结合这一章的概念，这两个要求可以表述为：我们希望得到一个经验误差和泛化误差都很小的学习器。经验误差和泛化误差是同一个判定公式在不同的数据集中得到的不同结果，我们用这两个结果的差异来判定学习器是否出现了欠拟合和过拟合。欠拟合的一般表现为学习器在训练集和新样本上的表现都不佳；过拟合的一般表现为学习器在训练集上的表现非常好，但是在新样本上的表现很差。欠拟合通常是由于学习器未能有效捕捉样本的共性特征，导致其对数据的表达能力不足；而过拟合则表现为学习器将训练集中样本的个体特征误认为是普遍规律。例如，在构建区分男性和女性的学习器时，若模型只捕捉到过于简单的特征，如“长头发的都是女性”，这会导致模型处于欠拟合状态，因为它忽略了数据中更复杂的性别差异。而过拟合的表现则是模型学习了过多与特定样本相关的细节规则，例如“有胡子、短头发、戴眼镜、穿白色衣服且坐在椅子上的才是男性”，这些规则虽然在训练集中特定样本上可能成立，但缺乏普遍性，无法有效泛化到新的样本。

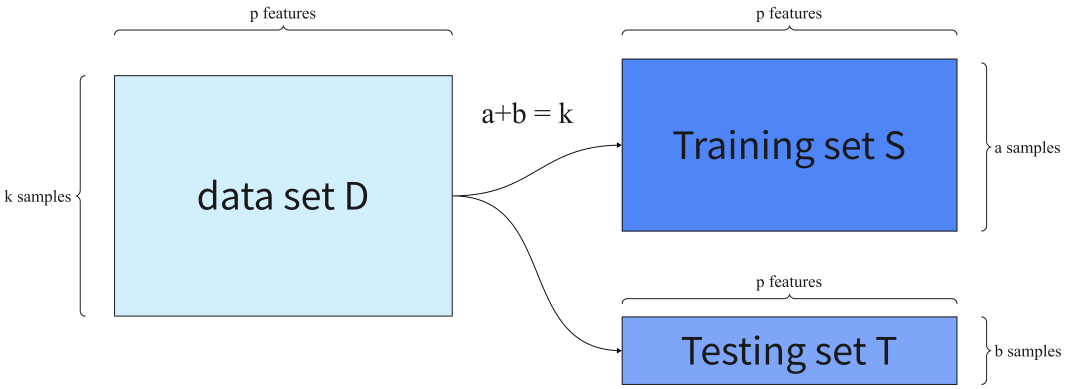
2.2 分割数据集的方法

现实中数据集是一种很宝贵的资源，我们没有充足的新样本最为评估泛化误差的数据集。在这里，我将包含所有样本的集合称为数据集，并假设数据集中的所有样本都是独立同分布的。我们不得不对数据集进行分割，一部分样本作为训练集来训练模型，另一部分样本作为测试集来验证模型的泛化能力，测试集中产生的**测试误差 (testing error)** 被视为泛化误差的替代。训练集和测试集之间不应该有交集，道理非常简单，如果一个样本即出现在训练集，又出现在测试集，那么这个学习器在训练过程中就见过这个样本，以及学习了这个样本的特征，那么当学习器在测试集中再次见到这个样本时，很轻松就可以判断正确，因此容易造成测试误差偏低，导致结果“欺骗”了人们，使人们决定这个学习器的表现好。

将数据集分为训练集和测试集是一个初步的想法，具体如何分数据集为训练集和测试集有多重不同的方法。主要有留出法，交叉验证法和自助法

2.2.1 留出法

留出法 (hand-out) 是最符合直觉的分割方法，即直接将数据集 D 分为两个互斥的集合，其中一个最为训练集 S，另一个最为测试集 T,$D = S \cup T, S \cap T = \emptyset$, 在 S 上训练出模型后，用 T 来评估其测试误差，作为对泛化误差的估计。

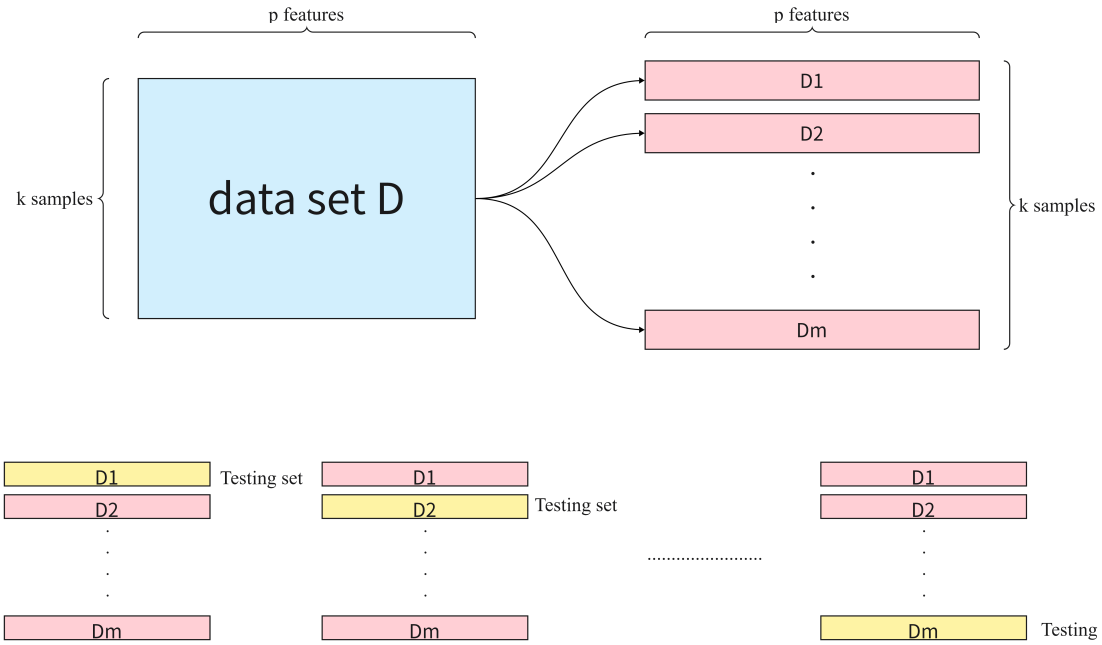


需要注意的是，将不同的样本划分到训练集/测试集可能会对结果产生不同的影响，比如，在一个判断人类性别的分类器中，如果把数据集中所有的男性图片放到训练集，所有的女性图片放到测试集中，那么显然不会有良好的训练效果，所以一般采用**分层抽样 (stratified sampling)**, 一般是将三分之二或者五分之四的样本用于训练，剩余的用于测试。并且重复随机抽样若干次，将结果取平均值来避免因为数据集划分的随机性而导致的偏差。

2.2.2 交叉验证法

交叉验证法（cross validation） 先将数据集 D 划分为 k 个大小相似的互斥子集，即 $D = D_1 \cup D_2 \cup \cdots \cup D_k$，$D_i \cap D_j = \emptyset (i \neq j)$。每个子集 D_i 都尽可能保持数据分布的一致性，即从 D 中通过分层采样得到。然后，每次用 $k-1$ 个子集的并集作为训练集，余下的那个子集作为测试集；这样就可获得 k 组训练/测试集，从而可进行 k 次训练和测试，最终返回的是这 k 个测试结果的均值。通常把交叉验证法称为“k-折交叉验证”（k-fold cross validation）。

k 常用的取值有 5,10,20 等，当 k= 数据集中样本数量时，相当于每一次只留出一个样本作为测试集，这样的分割方式称为**留一法 (Leave-One-Out),LOO**



2.2.3 自助法 (有放回抽样)

自助法就是从原本的数据集 D 做有放回抽样，假设数据集 D 中有 m 个样本，做 m 次有放回抽样，将每一次抽出来的样本记录之后再把样本放回 D，m 次之后就会有一个新的数据集 D' 根据这个假设，新数据集 D' 中肯定会有一些重复的样本，在 m 次采样中始终不被采到的概率是 $(1-\frac{1}{m})^m$, 取极限

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m \to \frac{1}{e} \approx 0.368$$

也就是说，每一次生成新的数据集 D'，都会有大约三分之一的 D 中的样本没有被选中，那么 D' 被视为训练集，$D-D'$ 的样本作为测试集。

2.3 性能度量

2.3.1 均方误差

2.3.2 错误率与精度

2.3.3 查准率、查全率与 F1

2.3.4 ROC 与 AUC

2.3.5 代价敏感错误率与代价曲线

2.4 比较检验

2.5 偏差与方差权衡