统计学习方法学习笔记

Pionpill [1]

本文档为作者学习《统计学习方法》[2]一书时的笔记，
主要对常用统计机器学习方法做了笔记。

2021 年 9 月 20 日

[1]笔名：北岸，电子邮件：673486387@qq.com，Github：https://github.com/Pionpill
[2]《统计学习方法》: 李航，清华大学出版社，2019 年第二版

前言：

笔者为软件工程系在校本科生，空余时间自学了机器学习相关内容，在撰写该文档时，笔者并没有较深的统计学或数学知识，数学知识体系仅限于理工科必修的《高等数学》《概率统计》《离散数学》《线性代数》[1]四本书。数据方面自学了基本的数据处理方式，包括 python 基础，numpy,pandas 两个重要库以及一些常用数据分析工具，这方面的知识参考《利用 Python 进行数据分析》[2]一书。在模型建立方面，笔者有一定的老本，曾在 2021 年 MCMICM[3]中获 M 奖，主要负责编程，数据分析，排版。以上为笔者的基本介绍与知识系统说明，供各位学习时对比参考。

笔记的书写顺序与《统计学习方法》一致，章节基本一一对应，但笔者只记录认为重要或笔者有疑惑的地方，并不会对原书进行大段抄写，也即这篇笔记只适合各位读者辅助学习，或有一定机器学习知识后将此笔记作为一个备忘录，查阅模型或算法。

原书对于一些名称或公式并没有详细解读，而是跳跃性说明，因此本文对于一些专有名词，会编写脚注，若经常用到或难以理解，会附上本人收集到的认为有用的学习链接，并放在附录中。若有知识超过本科数学 (理工科必修的四本书)，会被标记本科超纲。此外，本文在编写同时，也在对《机器学习》[4]一书做笔记，主要由于部分内容过于深奥，会定期求助于我的一些研究生学长学姐，也可能由于在补充知识体系而长期断更。

由于本人非常不喜欢中式教科书式的死板，行文风格比较自由，还请谅解。

本笔记只是对原书的马虎概括与整理，如有疑问或需求，还请购买原书；本文所在 Github 仓库采用 GPL v3 协议，但请勿将本文商业使用。

由于数学知识较深，本人决定复习一下本科数学。

2021 年 9 月 20 日

[1]笔者撰写时已经一年没有接触相关知识了，在撰写过程中过深的数学知识会格外备注

[2]原书英文名：《Data Analysis for Python》

[3]美国大学生数学建模竞赛

[4]《机器学习》: 周立华，清华大学出版社，2019 年第二版

目录

I 监督学习

附录

I 监督学习

一、统计学习及监督学习的概念

本章节是对后续监督学习的概括，本人在学习时遇到很多不明确的概念，需要结合后续章节的理解。因此本章节只需要了解并对主要概念有印象，不必死磕每一个概念。

1.1 统计学习

统计学习 (statistical learning) 指计算机**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。统计学习也称为统计机器学习 (statistical machine learning)。

1. 统计学习的对象

   统计学习对象是数据，且要求数据具有一定的统计学规律性。此书只探讨利用数据构建模型进行预测与分析，**对数据观测以及收集等问题不做讨论**。

2. 统计学习的方法

   概括说明如下：从给定的，有限的，用于学习的训练数据 (training data) 集合触发，假设数据**独立同分布**[1]产生，并且假设要学习的模型属于某个函数的集合，称为**假设空间** (hypothesis space)[2]；应用某个评价准则 (evaluation criterion), 从假设空间中选取一个最优模型，使它对已知的训练数据及未知的测试数据 (test data) 在给定的评价准则下有最优的预测；最优模型的选取由算法实现。

   简言之，统计学习方法包括模型的假设空间，模型选择的准则以及模型学习的算法。也即统计学习方法的三要素：**模型 (model)，策略 (category) 和算法 (algorithm)**。

3. 统计学习方法的步骤

   (a) 得到有限的训练数据集合。

   (b) 确定包含所有可能的模型假设空间，即学习模型的集合。

   (c) 确定模型选择的准则，即学习策略。

   (d) 实现求解最优模型的算法，即学习的算法。

   (e) 通过学习方法选择最优模型。

   (f) 利用学习的最优模型对新数据进行预测或分析。

[1]独立同分布：独立：每次抽样之间没有关系，不会相互影响；同分布：每次抽样的样本服从同一个分布

[2]假设空间：模型属于由输入空间到输出空间的映射的集合

1.2 统计学习的分类

1.2.1 基本分类

统计学习或机器学习一般包括**监督学习，无监督学习，强化学习**。有时还包括半监督学习，主动学习。

- **监督学习**

监督学习 (supervise learning) 是指从**标注数据**[3]中学习预测模型[4]的机器学习问题。监督学习的本质是学习输入到输出的映射的统计规律。

1. **输入空间，特征空间，输出空间**

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)。

每个具体的输入是一个实例，通常用特征向量 (feature vector) 表示。所有特征向量存在的空间称为**特征空间** (feature space)。特征空间的每一维对应于一个特征。特征空间与输入空间有时不予区分，有时区分。模型实际上都是定义在特征空间上的。

监督学习中，将输入与输出看作是定义在输入 (特征) 空间与输出空间山上的随机变量的取值。用 X,Y 表示输入输出变量,x,y 表示输入输出变量取值。

书上变量的表示：

$$x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(i)}, \cdots, x^{(n)}\right)^T \tag{1.1.1}$$

其中 $x^{(i)}$ 表示 x 的第 i 个特征。注意 x_i 表示多个输入变量中的第 i 个变量。

$$x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^T \tag{1.1.2}$$

监督学习从训练数据集合中学习模型，对测试数据进行预测。训练数据由输入 (或特征向量) 与输出对[5]组成，通常表示为 T:

样本均为连续变量的预测问题称为**回归问题**，输出变量与有限个离散变量的预测问题称为**分类问题**；输入与输出变量均为变量系列的预测问题称为**标注问题**。

$$T = (x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N) \tag{1.1.3}$$

2. **联合概率分布**

监督学习假设输入与输出随机变量 X,Y 遵循联合概率分布[6] $P(X,Y)$。$P(X,Y)$ 表示分布函数。在学习过程中，假设联合分布概率存在，但其具体定义是未知的。这是监督学习关于数据分基本假设。

---

[3]标注数据：输入输出的对应关系。

[4]预测模型：对给定的输入产生相应的输出。

[5]输入与输出对：也称为样本或样本点

[6]联合概率分布: 两个及以上随机变量组成的随机向量的概率分布

3. **假设空间**

监督学习的目的在于学习一个由输入到输出的映射，也即找到最好的映射模型。模型属于由输入空间到输出空间映射的集合，这个集合被称作假设空间 (hypothesis space)。假设空间的确定就意味着学习范围的确定。

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 $Y = f(x)$ 表示。对具体的输入进行相应的输出预测时，写作 $P(y|x)$ 或 $y = f(x)$。

4. **问题的形式化** 监督学习利用训练数据集学习一个模型，再用模型对测试样本集进行预测。

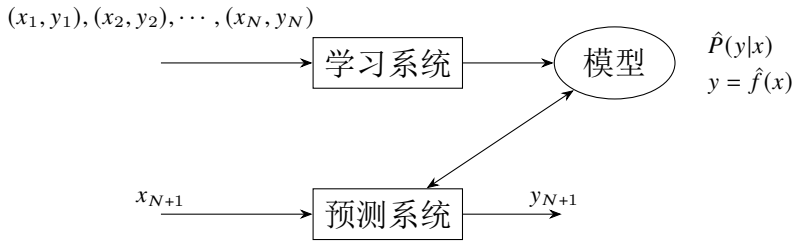

图 1.1.1 监督学习

其中，$\hat{P}(Y|X)$ 称为条件概率分布，$Y = \hat{f}(X)$ 称为决策函数。在预测过程中，预测系统对于给定的测试样本中的输入 x_{N+1}，由模型 $y_{N+1} = \arg\max_y \hat{P}(y|x_{N+1})$[7] 或 $y_{N+1} = \hat{f}(x_{N+1})$ 给出相应输出 y_{N+1}。

- **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型，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[8]。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 $x \in \mathcal{X}, z \in \mathcal{Z}$。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。

无监督学习使用的是无标注数据，训练数据表示为 $U = \{x_1, x_2, \ldots, x_N\}$。

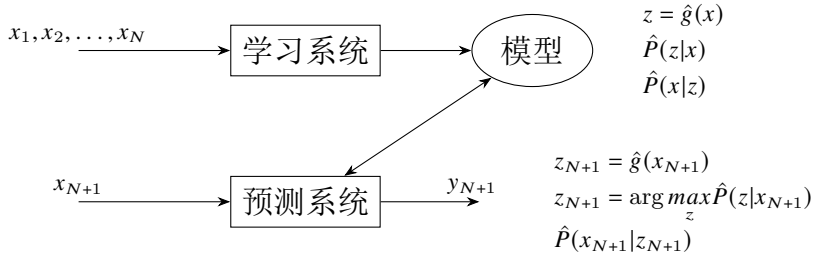

图 1.1.2 无监督学习

[7] arg 表示变元，arg min 表示后面式子达到最小值时变量的取值

[8] 隐式结构空间：类似输出空间

- **强化学习**

强化学习 (reinforcement learning)[9]是指智能系统**在环境的连续互动中**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)[10]，智能系统能观察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**。

智能系统与环境的交互如下图，在每一步 t，智能系统从环境中观测到一个状态 (state) s_t，与一个奖励 (reward) r_t，采取一个动作 (action) a_t。环境根据系统选择的动作，决定下一步 s_{t+1}, r_{t+1}。要学习的策略表示为给定状态下采取的动作。智能系统的目标不是短期奖励的最大化，二十长期累积奖励的最大化。强化学习过程中，系统不断地试错，以达到学习最优策略的目的。

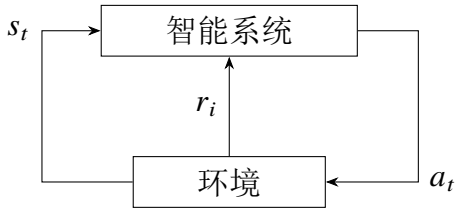

图 1.1.3 强化学习

- **半监督学习与主动学习**

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。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。

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。主动学习的目标是**找出对学习最有帮助实例让教师标注**，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**

统计学习可分为概率模型 (probabilistic model) 和非概率模型 (non-probabilistic) 或者确定性模型 (deterministic model)。在监督学习中，概率模型取条件概率分布形式 $P(y|x)$，非概率模型取函数形式 $y = f(x)$。在无监督学习中，概率模型取条件概率分布形式 $P(z|x)$ 或 $P(x|z)$，非概率模型取函数形式 $z = g(x)$。

条件分布与函数可以相互转化，概率模型与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概率联合分布的形式，而非概率模型则不一定。

[9]强化学习不是书上重点，这里只做简单介绍

[10]马尔可夫决策过程：一种交互决策过程，这里不做过多说明

无论何种模型，均可使用最基本的规则：

$$\begin{aligned} &\text{加法规则：} P(x) = \sum_{y} P(x, y) \\ &\text{乘法规则：} P(x, y) = P(x)P(y|x) \end{aligned} \tag{1.1.4}$$

- **线性模型与非线性模型**

统计学习模型，特别是非概率模型，可以分为线性模型 (linear model) 与非线性模型 (non-liner 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，否则称为非线性模型。

- **参数化模型与非参数化模型**

参数化模型 (parametric model) 假设模型参数的维度固定，模型可以由有限维度参数完全刻画，非参数化模型 (non-parametric) 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

统计学习根据算法，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。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 $\hat{g}(x_t)$，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。

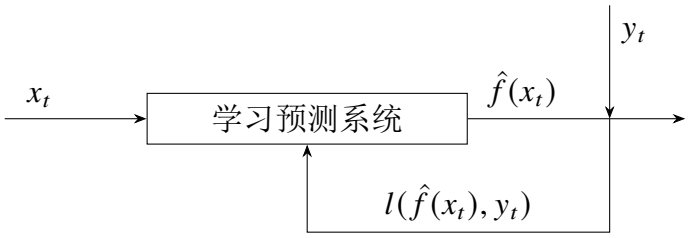

图 1.1.4　在线学习

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)。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[11]，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。将模型，未观测要素及其参数用变量表示，使用模型的先验分布是贝叶斯学习的特点。

假设随机变量 D 表示数据，随机变量 θ 表示模型参数，根据贝叶斯定理，可以用

[11] 贝叶斯定理及相关概念见附录

以下公式计算后验概率 $P(\theta|D)$：

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} \tag{1.1.5}$$

其中，$P(\theta)$ 是先验概率，$P(D|\theta)$ 是似然函数。

模型估计时，估计整个后验概率重分布 $P(\theta|D)$ 。如果需要给出一个模型，通常取后验概率最大的模型。

预测时，计算数据对后验概率分布的期望值 (这里 x 表示样本)：

$$P(x|D) = \int P(x|\theta, D)P(\theta|D)\mathrm{d}\theta \tag{1.1.6}$$

- **核方法**

核方法 (kernel model) 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，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。

把线性模型扩展到非线性模型，直接的做法是显示的定义从输入空间到特征空间的映射，在特征空间中进行内积计算。

1.3 统计学习方法三要素

统计学习方法由模型，策略和算法构成。

$$\text{方法} = \text{模型} + \text{策略} + \text{算法}$$

1.3.1 模型

在监督学习中，模型就是索要学习的条件概率分布或决策函数。模型的假设空间包含所有可能的条件概率分布或决策函数。假设空间用 $\mathcal{F}$ 表示，假设空间可以定义为决策函数的集合：

$$\mathcal{F} = \{f|Y = f(X)\} \tag{1.1.7}$$

这时的 $\mathcal{F}$ 通常是由一个参数向量决定的函数族：

$$\mathcal{F} = \{f|Y = f_\theta(X), \theta \in \mathbf{R}^n\} \tag{1.1.8}$$

参数向量 θ 取之于 n 维欧氏空间 $\mathbf{R}^n$，称为参数空间。

同样，假设条件也可以定义为条件概率的集合，这里不再说明。原书称由决策函数表示的模型为非概率模型，由条件概率表示的模型为概率模型。

1.3.2 策略

统计学习的目标在于从假设空间中选取最优模型。

- **损失函数与风险函数**

损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。

对于输入 X，预测值 $f(X)$ 与真实值 Y 可能一致也可能不一致。需要损失函数 (loss function) 或代价函数[12](cost function) 度量预测错误的程度。损失函数是关于 $f(X)$ 和 Y 的非负实值函数，记作 $L(Y, f(X))$。

统计学习常用的损失函数如下：

$$\begin{aligned}
&\text{0-1 损失函数 :} L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases} \\
&\text{平方损失函数 :} L(Y, f(X)) = (Y - f(x))^2 \\
&\text{绝对损失函数 :} L(Y, f(X)) = |Y - f(x)| \\
&\text{对数损失函数 :} L(Y, P(Y|X)) = -\log P(Y|X)
\end{aligned} \tag{1.1.9}$$

期望计算公式 (即风险函数/期望损失)：

$$\begin{aligned}
R_{exp}(f) &= E_p[L(Y, f(X))] \\
&= \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x)) P(x, y) \mathrm{d}x\mathrm{d}y
\end{aligned} \tag{1.1.10}$$

给定一个训练数据集 T，模型 $f(X)$ 关于训练数据集的平均损失称为经验风险 (empirical risk) 或者经验损失 (empirical loss)，记作 R_{emp}:

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.1.11}$$

由大数定理可知，当样本容量 N 趋于无穷时，经验风险 $R_{emp}(f)$ 趋于期望风险 $R_{exp}(F)$。所以可以使用经验风险估计期望风险。但现实中数据量并不足够，需要对经验风险进行矫正。

- **经验风险最小化与结构风险最小化**

经验风险最小化 (empirical risk minimization, ERM) 策略认为，经验风险最小的模型就是最优模型。根据这一策略，按经验风险最小化求最优模型就是求解最优化问题。

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.1.12}$$

当样本容量足够大时，经验风险最小化能保证有很好的学习效果。当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化就等于极大似然估计。

当样本容量很小时，经验风险最小化学习效果未必好，会产生” 过拟合[13]”(over-fitting) 现象。

结构风险最小化 (structural risk minimization，SRM) 是为了防止过拟合而提出的策略。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

[12]代价函数：和风险函数类似，所有样本误差均值

[13]过拟合：一个假设在训练数据上能够获得比其他假设更好的拟合，但是在训练数据外的数据集上却不能很好的拟合数据

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，结构风险的定义式：

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.1.13}$$

其中，$J(f)$ 为模型的复杂度，是定义在假设空间 $\mathcal{F}$ 上的泛函。模型 f 越复杂，复杂度 $J(f)$ 就越大。也即，复杂度表示了对复杂模型的惩罚。$\lambda \geqslant 0$ 是系数，用以权衡经验风险和模型复杂度。

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

$$\min_{f \in \mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.1.14}$$

1.3.3 算法

算法是指学习模型中具体的计算方法。即同什么样的计算方法从假设空间中求解最优模型。统计学习的问题归结为最优化问题。

1.4 模型评估与模型选择

1.4.1 训练误差与测试误差

训练误差的大小，对判定给定的问题是不是一个容易学习的问题是有意义的，但本质上并不重要。假设学习到的模型是 $Y = \hat{f}(X)$，训练误差是模型关于训练数据集的平均损失，N 是训练样本容量：

$$R_{emp}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i)) \tag{1.1.15}$$

测试误差反映了学习方法对未知的测试数据集的预测能力，十分重要。测试误差是模型 $Y = \hat{f}(X)$ 关于测试数据集的平均损失，$N^{'}$ 是测试样本容量：

$$e_{test} = \frac{1}{N^{'}}\sum_{i=1}^{N^{'}} L(y_i, \hat{f}(x_i)) \tag{1.1.16}$$

1.4.2 过拟合与模型选择

如果一味追求对训练数据的预测能力，所选模型的复杂度则往往比真模型更高。这种现象称为过拟合。过拟合是指学习时旋转的模型所办喊的参数过多，以至出现这一模型对已知

数据预测很好，但对位置数据预测很差的现象。可以说模型旋转旨在避免过拟合并提高模型的预测能力[14]。

下图描述了训练误差和测试误差与模型复杂度之间的关系。当选择的模型复杂度过大时，过拟合现象就会发生。

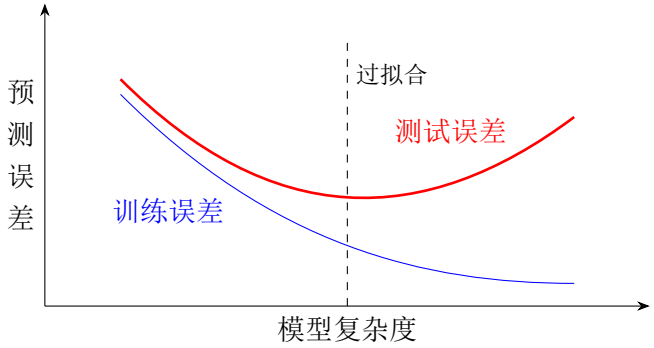

图 1.1.5　过拟合

1.5 正则化与交叉验证

选择复杂度适当的模型，以达到使测试误差最小的学习目的，避免过拟合现象发生。有两种常用的模型选择方案：正则化与交叉检验。

1.5.1 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或罚项。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**。正则化的一般形式如下 (第一项是经验风险，第二项是正则化项，$\lambda \geqslant 0$ 为调整两者之间的关系系数)：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.1.17}$$

第一项的经验风险较小的模型可能较复杂 (有多个非零参数)，这是第二项的模型复杂度会较大。**正则化的作用是选择经验风险与模型复杂度同时较小的模型**。

正则化符合奥卡姆剃刀原理[15]。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。从贝叶斯估计[16]角度来看，正则化项对于与模型的先验概率。可以假设复杂的模型有较小的先验概率，简单的模型有较大的先验概率。

[14] 书上 P20-22 给出了具体的例子，这里不对其进行说明

[15] 奥卡姆剃刀原理-个人理解：在两个方案同时能达到目标时，选择更为简单的方案。

[16] 贝叶斯估计：从参数的先验知识和样本出发。后文会专门讲解。

1.5.2 交叉验证

如果给定的样本数据充足，进行模型选择的一种简单方法是随机地将数据集切分成三部分，分别为**训练集** (training set)，**验证集** (validation set) 和**测试集** (test set)。训练集用于训练模型，验证集用于模型的选择，而测试集用于最终对学习方法的评估。

但实际应用中，数据往往不充分，为了选择好的模型，采用交叉检验的方法。交叉检验的基本方法是重复地使用数据；把给定的数据进行切分，将切分的数据集组合为训练集与测试集，在此基础上反复地进行训练，测试以及模型的选择。

- **简单交叉验证**

  首先随机地将数据分为两个部分，一部分作为训练集，另一部分作为测试集 (如 70% 的数据为训练集，30% 的数据为测试集); 然后用训练集在各种条件下 (如参数个数不同) 训练模型，从而得到不同的模型；在测试集上评价各个模型的测试误差，选出测试误差最小的模型。

- **S 折交叉验证**

  S 折交叉验证应用最多。首先随机地将数据切分为 S 个互不相交，大小相同地子集；然后利用 S-1 个子集地数据训练模型，利用余下的子集测试模型；将这一过程对可能的 S 种选择重复进行；最后选出 S 次评测中平均测试误差最小的模型。

- **留一次交叉验证**

  S 折交叉验证地特殊情况，S=N，称为留一次交叉验证。往往在数据缺乏时使用。

1.6 泛化能力

1.6.1 泛化误差

学习方法的泛化能力 (generalization ability) 是指由该方法学习到的模型对未知数据的预测能力，是学习方法本质上重要的性质。现实中采用最多的办法是通过测试误差来评价学习方法的泛化能力。这种方法依赖于数据，但数据集是有限的，评价结果很可能不可靠。

泛化误差的定义：如果学到的模型是 $\hat{f}$，那么同这个模型**对未知数据预测的误差即为泛化误差** (generalization error)：

$$\begin{aligned} R_{exp}(\hat{f}) &= E_p[L(Y, \hat{f}(X))] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{f}(x)) P(x, y) \mathrm{d}x\mathrm{d}y \end{aligned} \tag{1.1.18}$$

泛化误差反映了学习方法的泛化能力，如果一种方法学习的模型比另一种方法学习的模型具有更小的泛化误差，那么这种方法就更有效。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，简称泛化误差上界 (generalization error bound)[17]。

泛化误差上界具有以下性质[18]：

1. 泛化误差上界是样本容量的函数，当样本容量增加时，泛化上界趋于 0;
2. 泛化误差上界是假设空间容量 (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。

1.7 生成模型与判别模型

监督学习方法可分为生成方法 (generative approach) 和判别方法 (discriminative approach)。所学习到的模型分别称为生成模型 (generative model) 和判别模型 (discriminative model)。

- **生成方法**

  生成方法原理上由数据学习联合概率分布 $P(X,Y)$，然后求出条件概率分布 $P(Y|X)$ 作为预测的模型，即生成模型：

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \tag{1.1.19}$$

  这样的方法之所以称为生成方法，是因为模型表示了给定输入 X 产生输出 Y 的生成关系。

  生成方法的特点：

  – **生成方法可以还原出联合概率分布** $P(X,Y)$(判别方法不能)；

  – 生成方法的学习收敛速度更快；

  – 存在隐变量时，仍可以使用生成方法学习 (判别方法不能)；

- **判别方法**

  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，即判别模型。判别方法关系的是对给定的输入 X ，应该预测什么样的输出 Y。

  判别方法的特点：

  – 判别方法直接学习的是条件概率 $P(Y|X)$ 或决策函数 $f(X)$，直接面对预测，往往学习的准确率更高。

  – 由于直接学习 $P(Y|X)$ 或 $f(X)$，可以对数据进行各种程度上的抽象，定义特征并使用特征，可以简化学习问题。

[17]注意这里的英文名是 bound，也就是说没有 (没有必要) 定义泛化误差下界

[18]书上 P25-27 有一个例子，这里不做说明

1.8 监督学习应用

监督学习主要应用在三个方面：分类问题，标注问题和回归问题。

1.8.1 分类问题

在监督学习中，当输出变量 Y 取有限个离散值时，预测问题就变成了分类问题。此时输入变量 X 可以是离散的，也可以是连续的。

分类问题的一些概念与特有名词：

- 分类器 (classifier)：监督学习从数据学习中学习一个分类模型或分类决策函数；
- 分类 (classification)：分类器对新的输入进行输出的预测；
- 类别 (class)：可能的输出；
- 多分类问题：分类的类别为多个[19]。

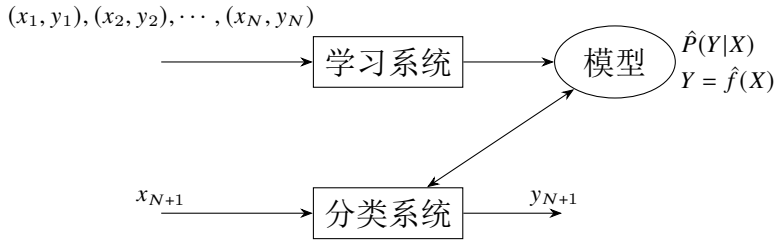

图 1.1.6　分类问题

评价分类器性能的指标一般是分类准确率 (accuracy)，其定义是：对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。

对于二分类问题常用的评价指标是精确率 (precision) 与召回率 (recall)。通常以关注的类为正类，其他类为负类，分类器在测试数据集上预测正确或不正确，4 种情况出现的总数分别记作：

- TP ——将正类预测为正类数;
- FN ——将正类预测为负类数;
- FP ——将负类预测为正类数;
- TN ——将负类预测为负类数;

于是有精确率 P，召回率 R 定义如下：

$$\begin{aligned} P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \end{aligned} \tag{1.1.20}$$

此外，还有 F_1 值，是精确率和召回率的调和均值。

[19] 书上主要讨论二分类问题

$$\begin{aligned}\frac{2}{F_1} &= \frac{1}{P} + \frac{1}{R} \\ F_1 &= \frac{2TP}{2TP + FP + FN}\end{aligned} \tag{1.1.21}$$

1.8.2 标注问题

可以认为标注问题是分类问题的一个推广，标注问题有时更复杂的结构预测问题的简单形式。标注问题的输入是一个观测序列，输出是一个标记序列或状态序列。标注问题的目标在于学习一个模型，使它能够对观测序列给出标记序列作为预测。

标记序列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，$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}), i = 1, 2, \cdots, N$；$y_i = (y_i^{(1)}, y_i^{(2)}, \cdots, y_i^{(n)})$ 是相应的输出标记序列。n 是序列的长度，对不同样本可以有不同的值。学习系统基于训练数据构建一个模型，表示为条件分布概率：

$$P(Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)} | X^{(1)}, X^{(2)}, \cdots, X^{(n)})$$

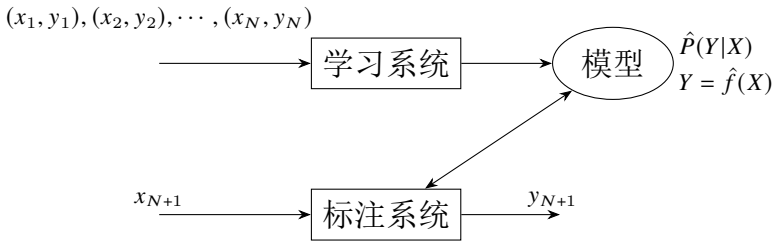

图 1.1.7　标注问题

评价标注模型的指示与评价分类模型的知识一样，常用的有标注准确率，精确率和召回率。

1.8.3 回归问题

回归用于预测输入变量和输出变量之间的关系，特别是当输入变量的值发生变化时，输出变量的值随之发生的变化。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。回归问题的学习等价于函数拟合。

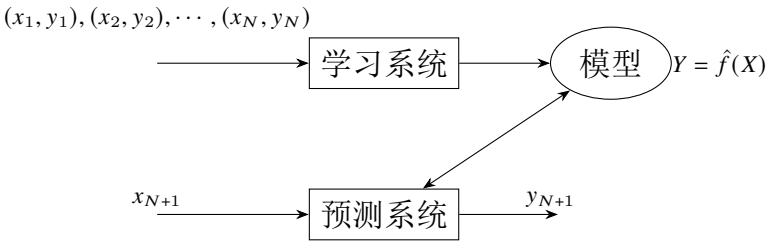

图 1.1.8　回归问题

回归问题按照输入变量的个数，分为一元回归和多元回归；按照输入变量和输出变量之间的关系的类型，分为线性回归和非线性回归。

回归学习最常用的方法是平方损失函数，也即用最小二乘法求解。

1.9 本章小结

- 统计学习可分为监督学习，无监督学习，强化学习。
  - 监督学习：从**标注数据**中学习预测模型的机器学习问题。
  - 无监督学习：从**无标注数据**中学习预测模型的机器学习问题。
  - 强化学习：在环境的**连续互动**中学习最优行为策略的机器学习问题。
- 统计学习三要素：模型，策略，算法。
- 过拟合现象的解决方案：正则化，交叉检验。
- 评估模型能力：泛化能力。
- 监督学习应用：分类问题，标注问题，回归问题。

二、感知机

感知机 (perceptron) 是**二类分类**的**线性分类**模型，其输入为实例的特征向量，输出为实例的类别，取 +1 和 -1 二值。感知机是神经网络与支持向量机的基础。

2.1 感知机模型

定义 1.2.1 (感知机). 假设输入空间 (特征空间) 是 $\mathcal{X} \subseteq \mathbf{R}^n$，输出空间是 $y = \{+1, -1\}$。输入 $x \in \mathcal{X}$ 表示实例的特征向量，对应输入空间 (特征空间) 的点；输出 $y \in \mathbf{Y}$ 表示实例的类别。由输入空间到输出空间的如下函数：

$$f(x) = sign(\omega \cdot x + b) \tag{1.2.1}$$

称为感知机，其中，ω 和 b 为感知机模型参数，$\omega \in \mathbf{R}^n$ 叫作权值 (weight) 或权值向量 (weight vector)，$b \in \mathbf{R}$ 叫作偏置 (bias)，$\omega \cdot x$ 表示内积。$sign$ 是符号函数，即

$$sign(x) = \begin{cases} +1, x \geqslant 0 \\ -1, x < 0 \end{cases} \tag{1.2.2}$$

感知机是一种**线性分类模型**，属于判别模型。感知机的假设空间是定义在特征空间中的所有线性分类模型或线性分类器。即函数集合 $\{f | f(x) = \omega \cdot x + b\}$。

感知机有如下几何解释：线性方程

$$\omega \cdot x + b = 0 \tag{1.2.3}$$

对应于特征空间 $\mathbf{R}^n$ 中的一个超平面 S，其中 ω 是超平面的法向量，b 是超平面的截距。这个超平面被分为将特征空间划分为两个部分。位于两部分的点 (特征向量) 分别被分为正负两类。因此，超平面 S 称为分离超平面：

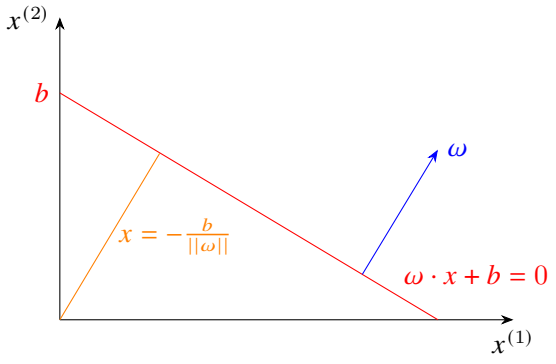

图 1.2.1　感知机模型

2.2 感知机学习策略

2.2.1 数据集的线性可分性

定义 1.2.2 (数据集的线性可分性). 给定一个数据集：

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

其中，$x_i \in \mathcal{X} = \mathbf{R}^n, y_i \in \mathbf{Y} = +1, -1$ 如果存在某个超平面 S 能够将数据集的正实例点和负实例点完全正确地划分到超平面的两侧，即对所有 $y_i = +1$ 的实例 i，有 $\omega \cdot x + b > 0$，对所有 $y_i = -1$ 的实例 i，有 $\omega \cdot x + b < 0$，则称数据集 T 为线性可分数据集；否则，称数据集 T 线性不可分。

2.2.2 感知机学习策略

假设训练数据集是线性可分的，感知机学习目标是求得能够将数据集的正实例点和负实例点完全正确地划分的超平面。也即求得感知机的模型参数 ω, b。需要确定一个学习策略，即**定义 (经验) 损失函数并将损失函数极小化**。

附录

一、符号说明

1.1 通用符号说明

表 A.1 通用输入输出符号

	符号	含义	符号	含义
监督学习	X	输入变量集合	x	输入变量取值
	Y	输出变量集合	y	输出变量取值
	$x^{(i)}$	输入变量 x 的第 i 个特征值	x_i	输出变量集的第 i 个变量
	T	输入与输出对/样本点		
无监督学习	$\mathcal{X}$	输入空间	x	输入变量取值
	$\mathcal{Z}$	输出空间	z	输出变量取值

表 A.2 监督学习模型符号

符号	含义	符号	含义
$\mathcal{F}$	假设空间	$L(Y, f(X))$	损失函数
R_{exp}	风险函数	R_{emp}	经验风险
R_{srm}	结构风险	$J(f)$	模型复杂度

二、重要概念

2.1 贝叶斯定理

2.1.1 基础概率论

条件概率公式

设 A, B 是两个事件，且 $P(B) > 0$, 则在事件 B 发生的条件下，事件 A 发生的条件概率为：

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{A.2.1}$$

条件概率中，事件 A, B 一般是有交集的，否则条件概率为 0 。

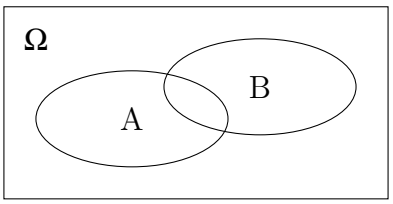

图 A.2.1　条件概率

乘法公式

1. 由条件概率公式得乘法公式：

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{A.2.2}$$

2. 乘法公式推广：对于任意正整数 $n \geqslant 2$，当 $P(A_1A_2 \dots A_{n-1}) > 0$ 时，有：

$$P(A_1A_2 \dots A_{n-1}A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2A_1) \dots P(A_n|A_1A_2 \dots A_{n-1}) \tag{A.2.3}$$

全概率公式

如果事件组 $B_1, B_2, \dots$ 满足如下条件，则称为事件组为样本空间的恶一个划分：

1. $B_1, B_2, \dots$ 两两互斥，即 $B_i \cap B_j = \varnothing \quad i \neq j; i, h = 1, 2, \dots; P(B_i) > 0$
2. $B_1 \cup B_2 \cup \dots = \Omega$

则有：

$$P(A) = \sum_{n=1}^{\infty} P(B_i)P(A|B_i) \tag{A.2.4}$$

2.1.2 贝叶斯定理

贝叶斯定理是英国数学家贝叶斯提出的，其主要用于解决” 逆概率” 问题，即在得知概率后，判断条件[1]。

将条件概率公式变形得到如下公式：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(A)\frac{P(B|A)}{P(B)} \tag{A.2.5}$$

定义以下名词：

1. 先验概率：$P(A)$，即在不知道 B 事件的前提下，对 A 事件概率的主观判断。
2. 可能性函数：$P(B|A)/P(B)$，作为调整因子，作用是使得先验概率更接近真实值。
   - 若 $P(B|A)/P(B) > 1$：先验概率增强，事件 A 发生概率可能性变大。
   - 若 $P(B|A)/P(B) = 1$：事件 B 无助于判断事件 A 的可能性。
   - 若 $P(B|A)/P(B) < 1$：先验概率削弱，事件 A 发生概率可能性变小。
3. 后验概率：$P(A|B)$，即在事件 B 发生后对事件 A 的概率进行重新评估。

[1]瞎编的

2.1.3 参考文献

此篇引用的文章如下:

- CSDN：基础概率论
- CSDN：贝叶斯公式及其应用

2.2 最小二乘法

2.2.1 最小二乘法原理

最小二乘法核心思想：如果误差是随机的，应该围绕真值上下波动。

假设 y 表示真值，y_i 表示样本值，就有最小二乘法：

$$\epsilon = \sum (y - y_i)^2 \text{最小} \Longrightarrow \text{真值} y \tag{A.2.6}$$

2.2.2 最小二乘的应用

在机器学习领域中，可以将真值看作模型函数 $y = f(x_i)$, 样本看作输出 y_i，那么就有：

$$\epsilon = \sum (f(x_i) - y_i)^2 \tag{A.2.7}$$

再结合多元微积分[2]的知识，就可以求解。

2.2.3 参考文献

此篇引用的文章如下:

- CSDN：如何理解最小二乘法

[2]这部分内容可能超过了理工科通修的数学体系