



机器学习（研究生）

学习问题的形式化

Risk, Hypothesis Space & Inductive Bias

赵凯 上海大学

<https://kaizhao.net/teaching/shu-ml-grad>

本讲内容索引

01 监督学习的基本要素

输入与输出 · 数据分布 · 训练样本 · 假设空间 · 损失函数

02 风险与经验风险最小化

期望风险与经验风险 · 经验风险最小化 · 模型复杂度 · 误差分解 · 正则化

03 归纳与归纳偏好

演绎与归纳 · 监督学习中的归纳 · 归纳偏好

04 小结

本讲要点 · 中英文术语对照 · 延伸阅读

本讲内容

把“从数据中确定规则”写成数学问题



问题的形式化

监督学习的基本要素
风险与经验风险最小化

从样本到未见输入

归纳与归纳偏好
小结

符号约定

整门课程统一使用以下记号

样本	$\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^d$, 粗体小写表示向量; 标签 $y \in \mathcal{Y}$
数据分布	$\mathcal{D}$: $(\mathbf{x}, y)$ 的联合分布, 未知
数据集	$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, 共 m 个样本; x_{ij} 为第 i 个样本的第 j 个特征
假设空间	$\mathcal{H}$: 候选模型 $h: \mathcal{X} \rightarrow \mathcal{Y}$ 的集合; $\hat{y} = h(\mathbf{x})$ 为预测
损失	$\mathcal{L}(\hat{y}, y) \geq 0$: 一次预测的代价
风险	期望风险 $R(h)$, 经验风险 $\hat{R}(h)$; 正则项 $\Omega(h)$, 系数 λ

记号 $^\top$: 转置; $\mathbb{R}^d$: d 维实数向量空间; $\mathbb{E}$: 期望; $\mathbb{I}[\cdot]$: 示性函数, 条件成立取 1, 否则取 0。

01

监督学习的基本要素

Elements of Supervised Learning

输入与输出 · 数据分布 · 训练样本 · 假设空间 · 损失函数

监督学习的五个要素

第一讲的“候选规则的范围”与“评判优劣的标准”，在这里分别是假设空间与损失函数

要素	记号	说明
① 输入与输出空间	$\mathcal{X}, \mathcal{Y}$	样本与标签的取值范围
② 数据分布	$\mathcal{D}$, 定义在 $\mathcal{X} \times \mathcal{Y}$ 上	未知，需从有限数据推断
③ 训练样本	$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$	通常假设独立同分布抽取
④ 假设空间	$\mathcal{H} \subseteq \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$	候选函数的集合，由建模者选定
⑤ 损失函数	$\mathcal{L}(\hat{y}, y) \geq 0$	一次预测的代价，由建模者选定

要点 分布未知，样本已给定；建模者需要选定的是假设空间与损失函数。

① 输入与输出空间

房价预测：一套房屋的多种特征 $\rightarrow$ 房屋总价

输入 x ：一套房屋的特征

面积	120 m ²	离地铁距离	1.5 km
卫生间数量	2 个	房龄	10 年
楼层	8 楼	学区	A 学区

$$x = (x_1, x_2, \dots, x_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^d$$

模型 h
将特征映射到总价

预测 $\hat{y} = h(x)$
房屋总价 / 万元

$$y \in \mathcal{Y} = \mathbb{R}$$

要点 x 是一套房屋的特征向量； $\mathcal{X}$ 是所有允许输入的特征向量的集合。

记号 学区等类别特征先编码，如 A、B、C 三个学区编码为 (1, 0, 0)、(0, 1, 0)、(0, 0, 1)，称独热编码 (one-hot encoding)； d 为编码后的特征数。分类任务的 $\mathcal{Y}$ 是有限的类别集合。

② 数据分布

$\mathcal{D}$ 描述：哪些房屋更常出现，以及它们的价格如何变化

输入的分布

小户型与小户型各有多少？不同房龄、学区的房屋各有多少？

$$p(\mathbf{x})$$

给定输入后的价格

已记录的特征相同，总价仍可能受议价等未记录因素影响

$$p(y \mid \mathbf{x})$$

$$(\mathbf{x}, y) \sim \mathcal{D}, \quad p(\mathbf{x}, y) = p(\mathbf{x}) p(y \mid \mathbf{x})$$

要点 $\mathcal{D}$ 通常未知；有限样本只是关于它的观测。

记号 $\mathcal{D}$ 是输入与输出的联合分布，不是某个数据文件；连续变量用密度，离散变量用概率。明确目标是上海还是东京的房屋，才能判断哪些样本有代表性。

③ 训练样本

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^\top$$

样本 i	面积 x_{i1} / m^2	距离 x_{i2} / km	总价 y_i / 万元
1	60	0.4	294
2	120	0.4	522
3	180	2.4	680
$\vdots$	$\vdots$	$\vdots$	$\vdots$

要点 一行是一套房屋； $\mathbf{x}_i$ 是输入， y_i 是真实观测。下标 i 给样本编号， j 给特征编号。

记号 补充示例（非教材例题），表中只展开面积与距离两个特征，数值为构造； m ：样本数； d ：特征数。

③ 训练样本：独立同分布假设

记 $D \sim \mathcal{D}^m$ 时，已假定样本独立同分布（independent and identically distributed, i.i.d.）

失效方式	场景	后果
跨地区分布偏移	上海房屋训练，东京房屋测试	特征分布与定价规律都可能变化
标签偏移（label shift）	疾病流行率随季节变化	先验错配，需要重新校准
样本不独立	时间序列；同一病人的多次采样	随机拆分可能高估性能
分布漂移	推荐系统、金融、用户行为	原有评估可能不再适用

常见错误 把同一病人的不同切片同时放进训练集与测试集：模型可借助个体身份得到乐观的成绩。数据划分应以**独立单元**为准。

记号 经典分析还要求训练与测试样本来自同一分布：分布偏移（distribution shift）。

④ 假设空间：所有候选模型

以两个输入为例：面积 x_1 、离地铁距离 $x_2 \rightarrow$ 总价 y

$$\mathcal{H}_1 = \{ h(x_1, x_2) = w_{00} + w_{10}x_1 + w_{01}x_2 \mid w_{00}, w_{10}, w_{01} \in \mathbb{R} \}$$

候选模型（示意）	函数
$h_A \in \mathcal{H}_1$	$h_A(x_1, x_2) = 60 + 4x_1 - 20x_2$
$h_B \in \mathcal{H}_1$	$h_B(x_1, x_2) = 80 + 3x_1 - 10x_2$

- **定次数**：确定候选范围；**学系数**：在范围内选出具体模型
- 次数上限改为 2，得到更大的 $\mathcal{H}_2 \supseteq \mathcal{H}_1$

问题 候选模型有无穷多个。用什么标准评判一个模型的预测好坏？

记号 假设空间 (hypothesis space)；系数为实数，可连续变化，因此候选模型有无穷多个。表中系数为补充示例（非教材例题）；教材中的离散例子见下页。

例：西瓜问题的假设空间

假设“好瓜”由色泽、根蒂、敲声三个属性完全确定：好瓜 $\leftrightarrow$ (色泽 =?) $\wedge$ (根蒂 =?) $\wedge$ (敲声 =?)

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

- 每个属性取一个具体值，或取通配符“*”（取什么值都合适）
- 设色泽、根蒂、敲声分别有 3、2、2 种可能取值，各加一个“*”
- 再加上“世界上没有好瓜”的假设 $\emptyset$

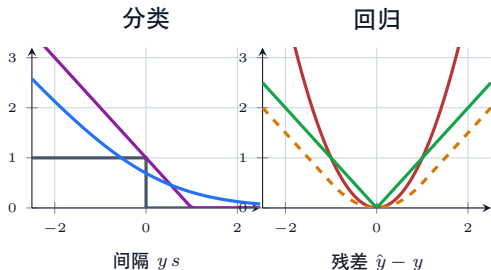
$$|\mathcal{H}| = 4 \times 3 \times 3 + 1 = 37$$

要点 学习就是在假设空间中搜索与训练集“匹配”的假设。

记号 例题：周志华《机器学习》第 1 章 1.3 节，表 1.1，图 1.1，pp.4-5；此处假定训练样本不含噪声，并不考虑“非青绿”这样的 $\neg A$ 操作。

5 常用损失函数

损失 $\mathcal{L}(\hat{y}, y)$ 衡量一次预测的代价



损失	形式	用于
0-1	$\mathbb{I}[\text{sign}(s) \neq y]$	分类的真目标
hinge	$\max(0, 1 - ys)$	支持向量机
对数	$\ln(1 + e^{-ys})$	逻辑回归
平方	$(\hat{y} - y)^2$	回归
绝对值	$ \hat{y} - y $	抗离群点
Huber	分段	两者折中

问题 损失只评价一次预测。如何评价模型在所有可能的输入上的表现？

记号 分类时 $y \in \{-1, +1\}$, $s = h(x) \in \mathbb{R}$ 为分类得分, 预测类别 $\text{sign}(s)$ (约定 $s = 0$ 时判为 $+1$); 回归时 $\hat{y}$ 为预测值。0-1 损失不连续, 难以用梯度优化, hinge 与对数损失是它的凸替代。Huber 损失图中取阈值 1。

02

风险与经验风险最小化

Risk & Empirical Risk Minimization

期望风险与经验风险 · 经验风险最小化 · 模型复杂度 · 误差分解 · 正则化

期望风险与经验风险

同一个模型 h 、同一个损失 $\mathcal{L}$ ，区别在于在哪些样本上取平均

期望风险 expected risk

面向总体分布 $\mathcal{D}$

$$R(h) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\mathcal{L}(h(\mathbf{x}), y)]$$

从 $\mathcal{D}$ 抽取新样本时的平均损失
分布未知 → **无法直接计算**
希望降低的泛化目标

经验风险 empirical risk

面向已有训练样本 D

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(h(\mathbf{x}_i), y_i)$$

在 m 个训练样本上的平均损失
样本已知 → **可以直接计算**
训练时的优化依据

要点 经验风险看“做过的题”，期望风险看“新题平均做得怎样”。

记号 期望风险亦称泛化误差（generalization error），本课程统一称期望风险；期望按 $\mathcal{D}$ 中样本出现的概率加权。

经验风险最小化

期望风险无法计算，用可以计算的**经验风险**代替

- ① **给定训练集** $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$
- ② **逐个算损失** 模型预测 $h(\mathbf{x}_i)$ 与真实值 y_i 比较: $\mathcal{L}(h(\mathbf{x}_i), y_i)$
- ③ **求平均损失** $\hat{R}(h) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(h(\mathbf{x}_i), y_i)$

$$\textcircled{4} \quad \hat{h} \in \arg \min_{h \in \mathcal{H}} \hat{R}(h)$$

设定候选范围
模型类型、复杂度

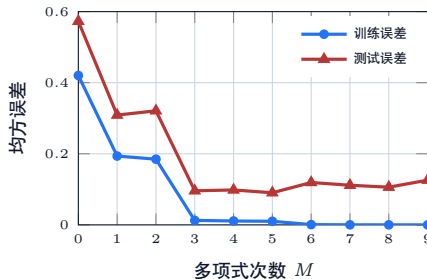
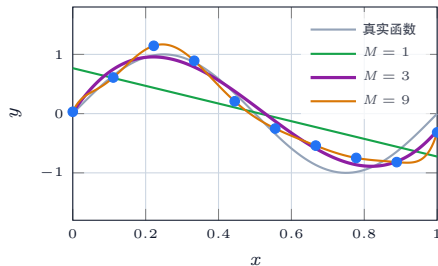
最小化经验风险
学习模型参数

验证集比较
选择类型与复杂度

记号 经验风险最小化 (Empirical Risk Minimization, ERM); $\hat{h}$ 只是给定 $\mathcal{H}$ 、给定数据、给定损失下的经验最优，不是真实分布上的最优。

模型复杂度与两条误差曲线

$y = \sin(2\pi x) + \varepsilon$: 固定 10 个训练样本, 分别拟合 $M = 0, 1, \dots, 9$ 次多项式



$$\hat{w}_M \in \arg \min_{w \in \mathbb{R}^{M+1}} \frac{1}{m} \sum_{i=1}^m (w_0 + w_1 x_i + \dots + w_M x_i^M - y_i)^2$$

记号 实验设置同 Bishop (2006) 第 1.1 节, 图 1.2–1.5, pp.4–8 (多项式次数记 M); 书中未给出样本数值, 此处样本为按该设置生成的补充数据: 训练输入取 $[0, 1]$ 上的等距网格, 测试样本 100 个, $\varepsilon \sim \mathcal{N}(0, 0.25^2)$ 。 $M = 9$ 时训练误差为 0, 测试误差为 0.126。

泛化间隙与误差分解

经验风险最小化只控制训练误差

泛化间隙（generalization gap）： $R(\hat{h}) - \hat{R}(\hat{h})$

与最优解的差距可以拆成两项：

$$R(\hat{h}) - R(h^{\text{Bayes}}) = \underbrace{R(\hat{h}) - R(h^*)}_{\text{估计误差}} + \underbrace{R(h^*) - R(h^{\text{Bayes}})}_{\text{近似误差}}$$

$\mathcal{H}$ 变大（能表示的函数更多）	近似误差	估计误差
----------------------------	------	------

不增

可能增大

问题 能否在不扩大 $\mathcal{H}$ 的前提下，表达“更偏好某些模型”？

记号 h^* ： $\mathcal{H}$ 内期望风险最小的模型； h^{Bayes} ：全体函数中期望风险最小者（第 5 讲的贝叶斯最优分类器）；假定二者存在。

同一个模板，不同的算法

在经验风险上加正则项：损失衡量拟合程度，正则项表达对模型的偏好

$$\min_{h \in \mathcal{H}} \hat{R}(h) + \lambda \Omega(h)$$

损失 $\mathcal{L}$	正则项 Ω	得到的方法
平方	无	最小二乘线性回归
平方	$\ w\ _2^2$	岭回归
对数	$\ w\ _2^2$	正则化逻辑回归
hinge	$\ w\ _2^2$	支持向量机

问题 正则项、假设空间都是建模者的选择。训练数据本身能否唯一确定模型？

记号 $\lambda \geq 0$ ：正则化系数； $\|w\|_2^2 = \sum_j w_j^2$ 。

03

归纳与归纳偏好

Induction & Inductive Bias

演绎与归纳 · 监督学习中的归纳 · 归纳偏好

演绎与归纳

演绎 (deduction) 一般 $\rightarrow$ 个别

所有天鹅都是白色的。
这是一只天鹅。

$\therefore$ 它是白色的。

前提真、推理有效 $\Rightarrow$ 结论必真

归纳 (induction) 个别 $\rightarrow$ 一般

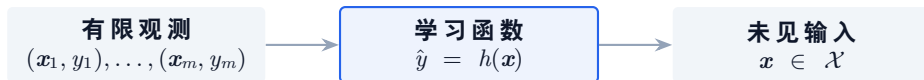
见过的天鹅都是白色的。

$\therefore$ 所有天鹅都是白色的。

结论可能被新的观察推翻

记号 澳大利亚的黑天鹅就是反例 (Australian Museum: Black Swan)。左侧大前提是为演示逻辑形式而假定的。

监督学习中的归纳



有限套房屋的成交记录 $\rightarrow$ 未成交房屋的价格预测

要点 从“样本上成立”到“未见输入上也适用”，是**归纳的跨越**：记住训练标签足以复现观测，预测未见输入还需要额外的假定。

记号 连续输入空间中，有限样本之外还有无穷多个输入。

例：西瓜问题中的归纳偏好

表 1.1 的四个训练样本，有三个假设与之完全一致

与训练集一致的假设	新瓜（青绿；蜷缩；沉闷）
好瓜 $\leftrightarrow$ (色泽 = *) $\wedge$ (根蒂 = 蜷缩) $\wedge$ (敲声 = *)	好瓜
好瓜 $\leftrightarrow$ (色泽 = *) $\wedge$ (根蒂 = *) $\wedge$ (敲声 = 浊响)	不是好瓜
好瓜 $\leftrightarrow$ (色泽 = *) $\wedge$ (根蒂 = 蜷缩) $\wedge$ (敲声 = 浊响)	不是好瓜

- 偏好“尽可能特殊”（适用情形尽可能少）：选第三个假设
- 偏好“尽可能一般”且更相信根蒂：选第一个假设

要点 训练集无法区分三个假设；学习算法对某类假设的偏好，称为**归纳偏好**（inductive bias）。

记号 例题：周志华《机器学习》第 1 章 1.4 节，图 1.2，p.6。

归纳偏好：回归中的例子

选择模型范围、损失或正则项时带入的额外假定与选择倾向

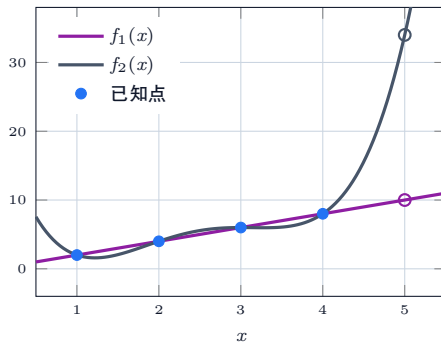
$$f_1(x) = 2x$$

$$f_2(x) = 2x + (x-1)(x-2)(x-3)(x-4)$$

x	1	2	3	4	5
$f_1(x)$	2	4	6	8	10
$f_2(x)$	2	4	6	8	34

已知四点上相同，未见输入 $x=5$ 上不同

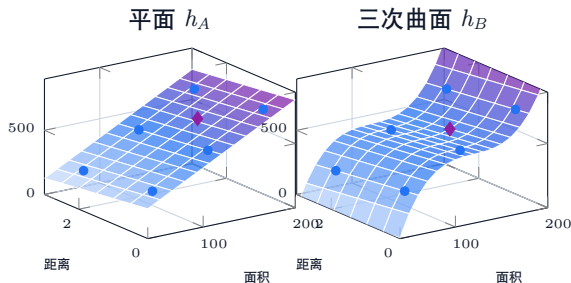
偏好更简单的规律 → 选择 f_1 → 预测 10



记号 $x=5$ 时乘积项为 $4 \times 3 \times 2 \times 1 = 24$ ，故 $f_2(5) = 34$ 。补充示例（非教材例题），对应周志华《机器学习》第1章1.4节图1.3 (p.7) 中两条都穿过训练点的曲线 A、B。简单性（奥卡姆剃刀）只是归纳偏好的一种；“bias”在此译作归纳偏好，与统计偏差区分。

房价预测中的归纳偏好

六套训练房屋；两个曲面都精确通过训练点，对新房屋却给出不同预测



$$h_A = 80 + 3.8x_1 - 35x_2$$

$$h_B = h_A + \frac{(x_1 - 60)(x_1 - 120)(x_1 - 180)}{1000}$$

新房屋：150 m²，距地铁 1.5 km

h_A 的预测 597.5 万元

h_B 的预测 516.5 万元

偏好平面关系 → 限定一次模型
→ 选择 h_A

记号 补充示例（非教材例题）：面积 60、120、180 m² 与距离 0.4、2.4 km 两两组合，总价由 h_A 生成；训练点处乘积项为 0，两者训练误差都为 0。◆ 为新房屋的预测。

04

小结

Summary

本讲要点 · 中英文术语对照 · 延伸阅读

本讲要点

五个要素 输入输出空间、数据分布、训练样本、假设空间、损失函数；分布未知

两种风险 训练时最小化**经验风险**，希望降低的是**期望风险**

评估 训练表现可能偏乐观，须用独立的数据评估

统一模板 许多方法是 $\min_h \hat{R}(h) + \lambda \Omega(h)$ 的实例

归纳偏好 数据不足以唯一确定模型，须引入额外的假定

问题 训练误差偏乐观，如何用数据可靠地估计模型在新样本上的表现，并据此选择模型？

中英文术语对照

英文全称	缩写	中文
independent and identically distributed	i.i.d.	独立同分布
Empirical Risk Minimization	ERM	经验风险最小化

延伸阅读

教材对应章节

- 周志华《机器学习》第 1 章 1.3–1.4：
假设空间与归纳偏好
- 李航《统计学习方法》第 1 章 1.3–1.7：
三要素、模型评估、正则化与交叉验证

进阶方向

估计新样本上的表现
泛化间隙的定量分析
模型如何表达偏好

评估指标、数据划分与交叉验证
Stanford CS229 讲义：学习理论
线性模型、决策树、支持向量
机、神经网络