《机器学习》第一章

王钰斐

目录

基本概念 | 机器学习

- 机器学习（machine learning）:

通过计算的手段，利用经验来改善系统自身的性能。（经验→数据）

Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed.

- 学习算法（learning algorithm）:

从数据产生模型的算法。（经验数据→学习算法→模型）

基本术语

例：
（色泽=青绿；根蒂=蜷缩；敲声=浊响）
（色泽=浅白；根蒂=硬挺；敲声=清脆）

示例数据

- 示例（instance),样本（sample）
- 属性（attibute),特征（feature）
- 属性值（attribute value）
- 属性空间（attribute space),样本空间（sample space）

训练过程

- 学习（learning),训练（training）
- 训练数据（training data）
- 训练样本（training sample）
- 训练集（traning set）

预测值

- 分类（classification)：二分类（正类，负类），多分类
- 回归（regression）
- 聚类（clustering）：簇(cluster)

假设空间

演绎

- 从一般到特殊的特化（pecialization)的过程。
- 基本原理 → 具体状况

归纳

- 从特殊到一般的泛化（generalization)的过程。
- 具体的事实 → 一般性规律
- 广义的归纳学习 → 样例中学习
- 狭义的归纳学习 → 训练数据中学得概念。（布尔概念学习）

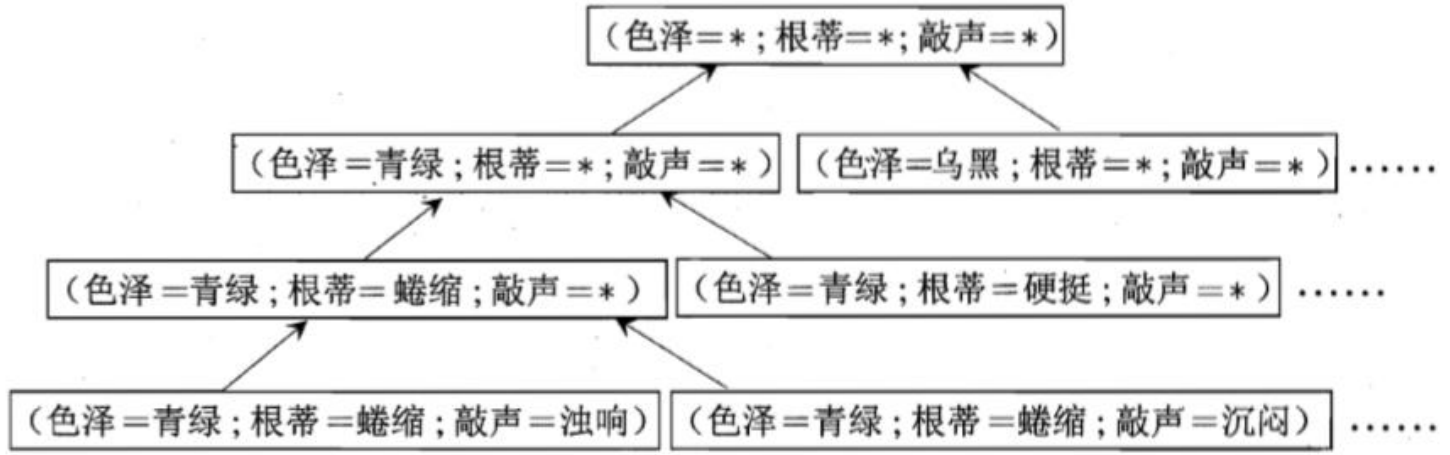
假设空间　布尔概念学习

- 对“是”，“不是”表示为0/1布尔值

例如：

表 1.1　西瓜数据集

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



- 假设空间规模大小：4 × 4 × 4 ＋ 1 ＝ 65
- ＊： 表示无论取任何值都合适
- ϕ： 表示以上概念都不存在

归纳偏好

- 机器学习算法在机器学习过程中对某种类型假设的偏好。

奥卡姆剃刀

- 若有多个假设与观察一致，则选择最简单的那个。

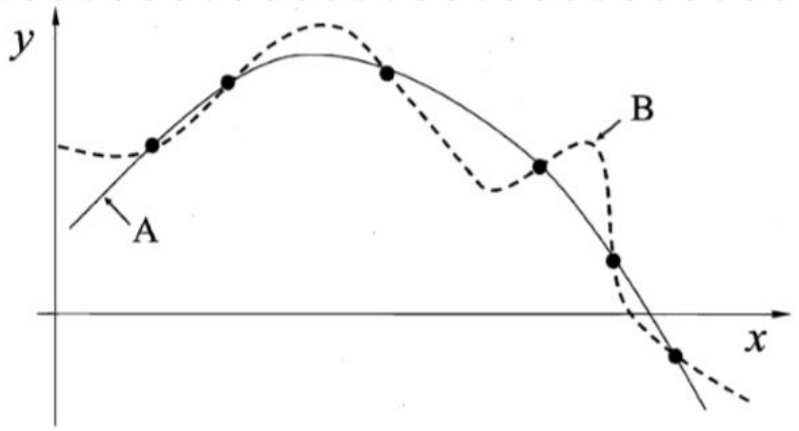


图 1.3 存在多条曲线与有限样本训练集一致

没有免费的午餐

- 对于所有机器学习问题，任何一种算法（包括瞎猜）的期望效果都是一样的。
- 没有一种机器学习算法是适用于所有情况的。
- 某一个机器学习算法在某个领域好用，在另外一个领域就有可能不好用。在讨论算法的相对优劣，必须针对具体的学习问题。

发展历程

二十世纪七十年代

- 从二十世纪七十年代中期开始，人工智能研究进入了“知识期”，在这一时期，大量专家系统问世，在很多应用领域取得了大量成果。

二十世纪

- 八十年代：主流是符号主义学习，其代表包括决策树和基于逻辑的学习。
- 九十年代中期之前：主要技术是基于神经网络的连接主义学习。
- 九十年代中期：“统计学习”占据主流，代表性技术是支持向量机以及核方法。

二十一世纪

- 掀起了以”深度学习“为名的热潮，在若干测试和竞赛上，尤其是涉及语音、图像等复杂对象的应用中，深度学习技术取得了优越性能。

应用现状

- **机器学习为许多交叉学科提供了重要的技术支撑。**

例如：“生物信息学”试图利用信息技术来研究生命显现和规律，生物信息学研究涉及从“生命现象“到”规律发现“的整个过程，其间包括数据处理、数据管理、数据分析、仿真实验等环节，其中，数据分析是机器学习技术的舞台。

- **机器学习与普通人的生活密切相关。**

例如：天气预报、能源勘测、环境检测等方面。在商业营销中，有效利用机器学习技术对销售数据、客户信息进行分析，帮助商家优化库存，降低成本。

- **机器学习影响人类社会政治生活。**

例如：2012年美国大选期间，奥巴马有一支机器学习团队，他们对各类选情数据进行分析，为奥巴马提示下一步竞选行动。

机器学习 周志华 chap2 chap3 知识点总结

目录

Chap 2 经验误差与过拟合

- 错误率 = 分类错误的样本数/样本总数 $E = a / m$
- 精度 = 1 - 错误率

- 误差：学习器的实际预测输出与样本的真实输出之间的差异
- 经验误差（训练误差）：学习器在训练集上的误差
- 泛化误差：学习器在新样本上的误差
- 学习器的训练目标：希望得到泛化误差小的学习器。

- 过拟合：学习器学习能力过于强大，将训练样本自身的一些特点当做所有潜在样本具有的一般性质，导致泛化性能下降。
- 欠拟合：与“过拟合”相对，指对训练样本的一般性质尚未学好

由于机器学习面临的问题通常是NP或者更难的问题，因此过拟合不可避免只能缓解。

Chap 2 评估方法

- 假设测试样本从样本真实分布中独立同分布采样得到，使用一个“测试集”来测试学习器对新样本的判别能力，然后以测试集上的“测试误差”作为泛化误差的近似。
- 测试集应该尽可能与训练集互斥，即测试样本尽量不在训练集中出现。

- 一个包含m个样例的数据集 $D=\{(\mathbf{x}1,y1),(\mathbf{x}2,y2),...,(\mathbf{x}m,ym)\}$，有以下几种方式来划分训练集 S 和测试集 T:

方法名称	留出法	交叉验证法	自助法
主要思想	直接将数据集D划分为两个互斥的集合，训练集S和测试集T，在S上训练出模型后用T来评估误差，作为对泛化误差的估计	先将数据集D划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为测试集。进行k次训练和测试，返回k个测试结果的均值	对于包含m个样本的数据集D，进行m次有放回的随机取样（每次取一个样本）放入D’，将D’用于训练集，D-D’作为测试集（D中约有36.8%的样本未出现在D’中）
使用范围	初始数据量充足时	初始数据量充足时	1. 适用于数据集较小，难以有效划分训练/测试集的样本集 2. 能从初始数据集中产生多个不同的训练集，有利于集成学习等方法
缺点	测试集不是全部样本，降低评估保真性	稳定性和保真性取决于k的取值	产生的数据集改变了初始数据集的分布，会引入估计偏差
备注	通常若干次随机划分重复实验评估后取平均值，一般划分比例为2/3~4/5的训练集	特例：留一法，不受随机样本划分方式的影响，结果较为准确，缺点：数据集比较大时开销大	

Chap 2 参数调节

- 参数调节即模型的参数设定。
- 方法：对每个参数选定一个范围和变化步长
- 优点：虽然得到的往往不是最佳值，但是能折中开销和性能

- 最终模型：在学习算法和参数配置已经选定后，要用数据集 D 重新训练模型，从而得到最终模型。

Chap 2 性能度量

- 性能度量是衡量模型泛化能力的评价标准，不同的性能度量会导致不同的评判结果。

- 错误率与精度：**分类任务**中常用的两种性能度量。
  错误率：分类错误的样本数占总样本数的比例。
  精度：分类正确的样本数占总样本数的比例。

二分类问题

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- 查准率、查全率、F_1
  准确率（查准率）：$R = \frac{TP}{TP + FN}$

  召回率（查全率）：$P = \frac{TP}{TP + FP}$

  查准率与查全率相矛盾，一方升高另一方降低。
  当查准率=查全率时，为PR图的平衡点（BEP）。

  根据应用中对查准率和查全率的重视程度不同，得到 F_1度量的一般形式 F_β 度量：

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

β 度量了查全率对查准率的相对重要性，β>1 时查全率有更大影响，β<1时查准率有更大影响。

F_1

- ROC（受试者工作特征）与 AUC

  根据学习器的预测结果对样例进行排序，按此顺序逐个将样例作为正反例的分界点，计算出

  真正例率：

$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

  假正例率：

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

  作为横纵坐标绘制ROC曲线。 ROC曲线下的面积AUC，用于比较分类器的优劣。

F_β

Chap 2 性能度量

	横纵轴	意义	度量方法			比较
P-R曲线	查准率为纵轴，查全率为横轴	反映学习器性能	平衡点（BEP）：查准率=查全率	F1度量：表现对查准率和查全率的偏好	F_β度量	1）若一个学习器的曲线被另一个学习器完全包住，则后者性能一定优于前者；2）若交叉无法判断，比较曲线下面积大小
ROC曲线	纵轴真正例率，横轴假正例率	反应样本预测的排序质量				ROC曲线下面积：AUC
代价曲线	横轴是取值[0,1]的正例概率代价，纵轴是取值[0,1]的归一化代价	不同类型的错误造成的损失				ROC曲线上的每个点可以转化为代价平面上的一条线段，取所有线段下界，围成的面积为在所有条件下学习器的期望总体代价。

Chap 2 比较检验

方法	思想	解决的问题	结果示意图
假设检验	根据测试错误率估推出泛化错误率的分布	对单个学习器泛化性能的检验	二项检验：概率p符合二项分布；t校验：变量τ_t服从自由度为k-1的t分布
交叉验证t检验	若两个学习器的性能相同，则它们使用相同的训练/测试集得到的测试错误率应相同。	对多个学习器的性能进行比较，在一个数据集上比较两个算法的性能	变量τ_t服从自由度为k-1的t分布（k折交叉验证）
McNemar检验	假设两学习器性能相同	针对二分类问题，在一个数据集上比较两个算法的性能	变量τ_{χ^2}服从自由度为1的χ^2分布
Friedman检验	基于算法排序的校验方法，得到算法比较序值表。	在一组数据集上比较多个算法的性能，是基于算法排序的，要求k（算法个数）较大	原始Friedman检验：在k和N都较大时，变量τ_{χ^2}服从自由度为k-1的χ^2分布； 改进的Friedman检验:τ_F服从自由度为k-1和（k-1）(N-1)的F分布
Nemenyi后续检验	计算出平均序值差别的临界值域，若两个算法的平均序值之差超出了临界值域，则以相应的置信度拒绝“两个算法性能相同”这一假设。	在一组数据集上比较多个算法的性能，若“所有算法的性能相同”假设被拒绝，需进行“后续检验”。	Friedman检验图：纵轴显示各个算法，横轴是平均序值，若两个算法的横线段有重叠，则没有显著差别，否则有显著差别

- 线性回归的**基本思想**是采用对输入样例各个特征进行线性加权的方式得到预测的输出，并将预测的输出和真实值的均方误差最小化。

- **均方误差**即函数值与平均数的方差，它是回归任务最常用的度量，它采用的是欧几里得(欧式)距离。基于均方误差来进行模型求解的方法，称为“**最小二乘法**”。在线性回归中，“最小二乘法”就是找到一条直线，使所有样本到该直线的欧式距离之和最小。

- 求解线性方程 $E_{(w,b)}=\sum_{i=1}^{m}(y_i-wx_i-b)^2$ 中的 ω 和b的过程，称为最小二乘“参数估计”。分别对 ω 和b求偏导，当两个偏导数均为0时(极值点处)，得到的 ω 和b为最优解。

- 对于有多个属性的情况，我们可以用多元线性回归来实现问题的求解。

- 考虑单调可微函数g(.)，令 $y=g^{-1}\left(\omega^T x+b\right)$，

使得线性模型推广为广义线型模型。对数线性回归即是广义线性模型在g(.)=ln(.) 时的特例。

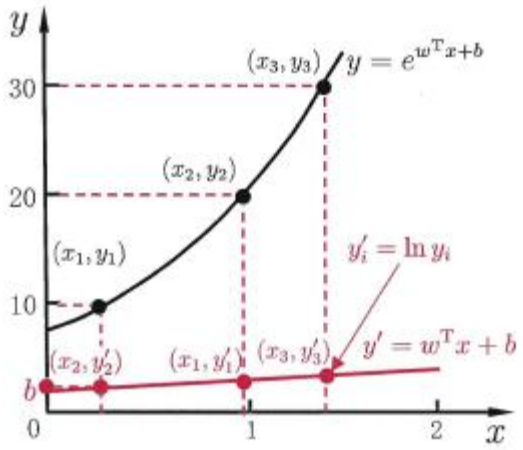


图 3.1 对数线性回归示意图

- 通过一个单调可微函数将分类任务的真实标记y与线性回归模型的预测值z联系起来。

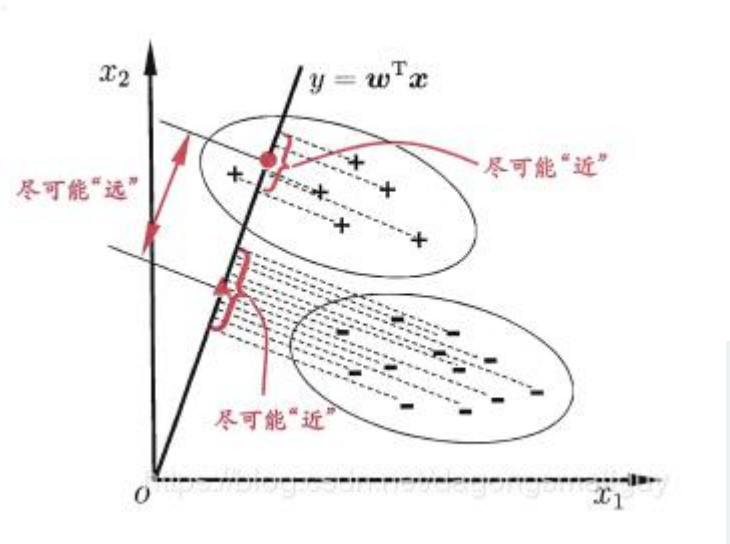
  若预测值z大于零就判为正例，小于零则判为反例，预测值为临界值零则可任意判别。因为单位阶跃函数不连续，用对数几率函数来替代。

- 将对数几率函数作为可微函数g(.)，推导出：

$$\ln \frac{y}{1-y} = \omega^T + b$$

- 把*y*看做是正例的可能性，1－*y*看成是反例的可能性，则两者的比值 $\frac{y}{1-y}$ 称为几率，反映了*x*作为正例的相对可能性对几率取对数则得到"对数几率"：$\ln \frac{y}{1-y}$

- 接下来就可以用“极大似然法”来对 ω 和b进行估计。



- 优点：
  - 直接对分类可能性进行建模，无需事先假设数据分布，避免了假设分布不准确带来的问题；
  - 可以得到类别近似概率预测；
  - 对率函数式任意阶可导的凸函数，易于求最优解。

- 线性判别分析（LDA）是一种经典的线性学习方法：给定一个训练样本，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近、异类样例的投影点尽可能远离;在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别：

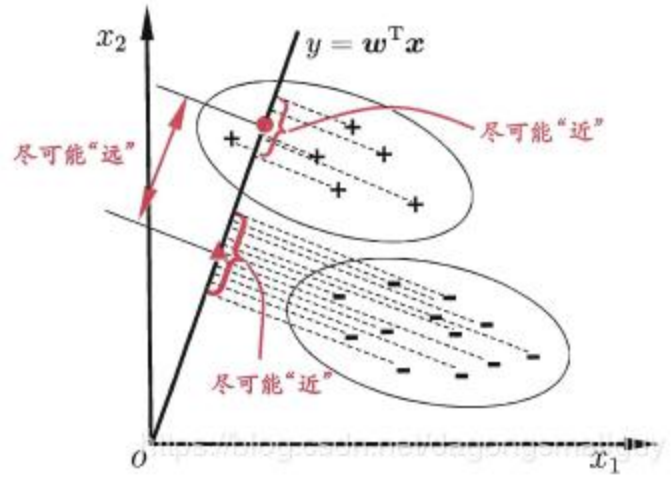
  投影： $\omega^T\mu_0$和$\omega^T\mu_1$

  协方差： $\omega^T\sum_0\omega$和$\omega^T\sum_1\omega$

- 欲使同类样例的投影点尽可能接近，可以让同类样例投影点的协方差尽可能小;
  而欲使异类样例的投影点尽可能远离，可以让类中心之间的距离尽可能大：

$$J=\frac{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}}.$$

这是 LDA 欲最大化的目标，即S_b与S_ω的广义瑞利商。



F_1

- 多分类学习主要是讲述如何对数据集进行拆分。
- 有些二分类学习方法可直接推广到多分类，但在更多情形下是基于一些基本策略，利用二分类学习器来解决多分类问题。

对于考虑 N 个类别 $C_1, C_2, C_3 \ldots C_N$

C1, C2, C3. . . . CN， 多分类学习的基本思路是"拆解法"即将多分类任务拆为若干个二分类任务求解.

拆分策略	思路	结果预测	预测开销
一对一（OvO）	将N个类别两两配对，产生N(N-1)/2个二分类任务和分类结果。	被预测得最多的类别作为最终分类结果。	存储开销和测试时间开销大，但在类别很多时，训练时间开销比OvR小
一对其余（OvR）	每次将一个类的样例作为正例，所有其他类的样例作为范例来训练N个分类器。	若测试时仅有一个分类器预测为正类，则对应的类别标记作为最终分类结果；若有多个分类器预测为正类，则选择置信度最大的类别标记为分类结果。	存储开销和测试时间开销小
多对多（MvM）	每次将若干个类作为正类，若干个其他类作为反类（OvO和OvR是MvM的特例），采用ECOC(纠错输出码)技术。		

Chap 3 类别不平衡问题

- 指分类任务中不同类别训练样例数目差别很大的情况。
- 基本策略：再缩放/再平衡，令 $\frac{y'}{1-y'}=\frac{y}{1-y}\times\frac{m^-}{m^+}$

当训练集中正、反例的数目不同时，我们直接拿预测几率和观测几率进行比较就可以得出结论。如正例数目为 m^+，反例数目为 m^-，则观察几率为 $\frac{m^+}{m^-}$，当分类器的预测几率高于观测几率便可以判断为正例：$\frac{y}{1-y}>\frac{m^+}{m^-}$。

实现方法	含义	代表算法	开销
过采样（上采样）	正例过多，增加一些正例	SMOTE：通过对训练集里的正例进行插值来产生额外的正例	开销大
欠采样（下采样）	反例过多，去除一些反例	EasyEnsemble：利用集成学习机制，将反例划分为若干个集合供不同学习器使用。	开销小
阈值移动	在用训练好的分类器进行预测时，将嵌入到决策过程中		