



机器学习（研究生）

贝叶斯分类器、降维与 度量学习

Bayesian Classifiers, Dimensionality Reduction & Metric Learning

赵凯 上海大学

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

本讲内容索引

01 贝叶斯决策论

贝叶斯公式 · 条件风险 · 贝叶斯最优分类器 · 生成式与判别式模型

02 极大似然估计

似然与对数似然 · 抛硬币与西瓜的例子 · 与损失函数、正则项的关系

03 朴素贝叶斯分类器

联合分布的参数量 · 属性条件独立性假设 · 完整算例 · 拉普拉斯修正

04 k 近邻学习

后验的局部估计 · 算法 · 错误率分析 · 维数灾难

05 主成分分析

数据的低维结构 · 投影的两个要求 · 目标函数与求解 · 方差贡献率 · 算例 · 特征脸

06 多维缩放

由距离求内积 · 由内积求坐标 · 与主成分分析的关系

07 度量学习

度量学习的对象 · 线性度量学习 · 孪生网络与三元组损失 · 对比学习

08 小结

本讲要点 · 中英文术语对照 · 课后作业

已学习的分类模型

线性模型、决策树、支持向量机、神经网络

模型	学习的对象	预测输出	训练目标
逻辑回归	线性函数 $f(x) = w^\top x + b$	$\sigma(f(x))$	对数损失
决策树	对属性空间的轴平行划分	叶结点的类别	信息增益、基尼指数
支持向量机	最大间隔超平面	$\text{sign } f(x)$	hinge loss 与正则项
神经网络	多层复合函数	softmax 输出	交叉熵

共同点 由训练样本直接学习从 x 到 y 的映射，即判别函数。

记号 $\sigma(\cdot)$: sigmoid 函数; sign : 符号函数; 决策树的划分准则也是一种训练目标，逐结点贪心优化。

共同的学习框架

经验风险最小化 (empirical risk minimization)

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(\mathbf{x}_i), y_i) \quad (\text{另加正则项})$$



- 损失 $\mathcal{L}$ 在训练之前选定，其影响**固化在参数中**
- 输出为类别或分数；分数的数值含义**随模型而异**
- 训练只用到样本，不显式描述样本的分布

记号 $\mathcal{F}$: 候选函数的集合（线性函数、决策树、某一结构的神经网络等）； $\hat{f}$: 训练得到的判别函数。

判别函数难以回答的问题

训练完成后，应用条件发生变化

场景	所需的量	仅有判别函数时
判错代价不对称：漏诊的损失远大于误诊	后验与损失	修改损失后重新训练
类别比例变化：训练集均衡，部署时阳性率 1%	先验 $P(c)$ 的修正	无从修正
置信度与拒识：低把握的样本转交人工	后验 $P(c \boldsymbol{x})$ 的大小	分数没有统一的概率含义
性能上限：错误率还能降低多少	真实分布下的最优判定	无法区分模型不足与类别重叠

要点 四个问题都需要**后验概率** $P(c | \boldsymbol{x})$ ，并把**估计后验与作出判定**分为两步。

记号 逻辑回归与神经网络输出的是 $P(c | \boldsymbol{x})$ 的估计，可用于前三个问题，其可靠性取决于概率是否校准（calibration）。前三个问题的讨论见 Bishop (2006) 第 1.5.4 节。

例：类别比例变化后的后验修正

训练集阳性、阴性各半；部署时阳性率为 1%； $P(\mathbf{x} | c)$ 不变

$$P'(c | \mathbf{x}) = \frac{P(\mathbf{x} | c) P'(c)}{\sum_k P(\mathbf{x} | c_k) P'(c_k)}$$

$$\propto P(c | \mathbf{x}) \frac{P'(c)}{P(c)}$$

- $P(\mathbf{x} | c) \propto P(c | \mathbf{x}) / P(c)$ ，换先验只需重新加权
- 不需要重新训练，也不需要新的标注

模型在 x 处输出 $P(+ | \mathbf{x}) = 0.90$

权重 $P'(c) / P(c)$:

阳性 $0.01 / 0.5 = 0.02$

阴性 $0.99 / 0.5 = 1.98$

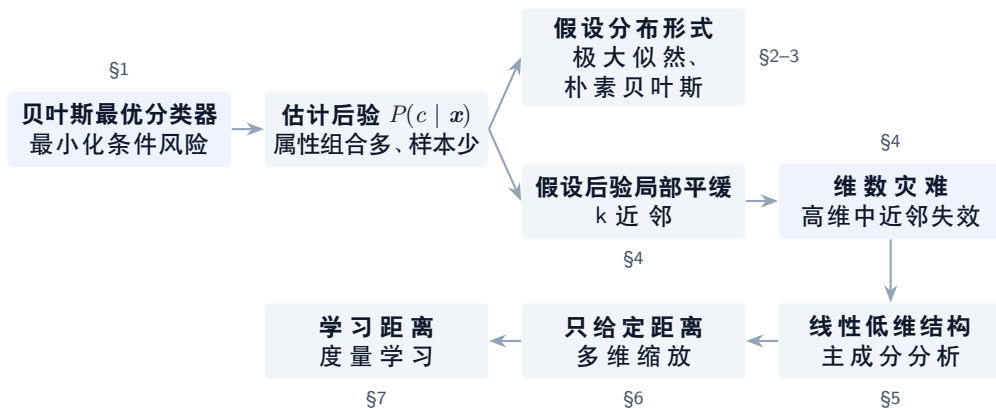
$$P'(+ | \mathbf{x}) = \frac{0.90 \times 0.02}{0.90 \times 0.02 + 0.10 \times 1.98} = 0.083$$

要点 同一个 x ，在阳性罕见的人群中后验由 0.90 降为 0.083：先验决定后验的量级。

记号 先验 $P(c)$ 、 $P'(c)$ ：看到 x 之前各类的比例，分别对应训练集与部署人群；似然 $P(\mathbf{x} | c)$ ：第 c 类中出现 x 的概率，两种场景相同；后验 $P(c | \mathbf{x})$ 、 $P'(c | \mathbf{x})$ ：看到 x 之后属于第 c 类的概率。补充示例（非教材例题）；方法见 Bishop (2006) 第 1.5.4 节 p.45，以及 Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1), 21–41。

从贝叶斯最优分类器到度量学习

后验难以直接估计：两种假设，以及近邻方法在高维中遇到的问题



本讲内容

用概率做分类决策；在合适的低维空间与距离下做近邻判定

Part 1 贝叶斯分类器

由后验概率与损失给出最优决策

贝叶斯决策论

极大似然估计

朴素贝叶斯分类器

Part 2 降维与度量学习

近邻方法依赖的维数与距离

k 近邻学习

主成分分析

多维缩放

度量学习

符号约定

沿用前几讲的记号；新增与概率、降维有关的记号

数据	$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$; $\mathbf{x}_i \in \mathbb{R}^d$, 第 j 个属性的取值记为 x_{ij} (单个样本时简记 x_j)
类别	$y \in \{c_1, \dots, c_K\}$; D_c 为 D 中第 c 类样本的集合, $ D_c $ 为其样本数
概率	$P(c)$ 先验, $P(\mathbf{x} c)$ 类条件概率 (似然), $P(c \mathbf{x})$ 后验; 大写 P 表示概率, 小写 p 表示密度 (连续变量与连续参数)
损失	$\mathcal{L}(c_i, c_j)$: 真实类别为 c_j 时判为 c_i 的损失, 沿用 $\mathcal{L}(\hat{y}, y)$ 的写法
降维	样本矩阵 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m) \in \mathbb{R}^{d \times m}$, 每一列为一个样本 (与周志华《机器学习》第 10 章一致; PRML 按行排列, $\mathbf{X}_{\text{PRML}} = \mathbf{X}^\top$); 低维坐标 $\mathbf{z}_i \in \mathbb{R}^{d'}$, $d' < d$

记号 $\ln$: 自然对数; tr : 矩阵的迹 (对角元素之和); $\mathbf{I}$: 单位矩阵; $\mathbf{1}$: 全 1 列向量; $\mathbf{M} \succeq 0$: $\mathbf{M}$ 半正定; 矩阵用粗体大写。

01

贝叶斯决策论

Bayesian Decision Theory

贝叶斯公式 · 条件风险 · 贝叶斯最优分类器 · 生成式与判别式模型

贝叶斯公式

观察 x 之前，对类别的判断是**先验** $P(c)$ ；观察 x 之后，修正为**后验** $P(c | x)$

$$\underbrace{P(c | x)}_{\text{后验}} = \frac{\overbrace{P(x | c)}^{\text{似然}} \overbrace{P(c)}^{\text{先验}}}{\underbrace{P(x)}_{\text{证据}}}, \quad P(x) = \sum_{k=1}^K P(x | c_k) P(c_k)$$

名称	英文	含义与名称的来由
先验	prior	观察 x 之前的判断；拉丁语 <i>a priori</i> ：“从在先的出发”
似然	likelihood	第 c 类中出现 x 的概率；作为 c 的函数，衡量 c 解释 x 的程度
后验	posterior	观察 x 之后修正的判断；拉丁语 <i>a posteriori</i> ：“从在后的出发”
证据	evidence	观察到的 x 出现的概率；对各类相同，起归一化作用

记号 似然与后验的条件方向相反： $P(x | c)$ 对 x 求和为 1，对 c 求和不一定为 1；比较各类后验时只需比较分子。

例：纹理清晰的西瓜

判断一个瓜是不是好瓜：观察纹理之前与之后

先验 尚未观察

$$P(\text{好}) = 8/17 = 0.47$$

似然 观察到纹理清晰

$$P(\text{清晰} | \text{好}) = 7/8$$

$$P(\text{清晰} | \text{坏}) = 2/9$$

后验 修正后的判断

$$P(\text{好} | \text{清晰}) = 7/9 = 0.78$$

$$P(\text{好} | \text{清晰}) = \frac{7/8 \times 8/17}{7/8 \times 8/17 + 2/9 \times 9/17} = \frac{7/17}{9/17} = \frac{7}{9}$$

- 观察之前：好瓜的可能性为 0.47，略低于坏瓜
- 观察之后：好瓜的可能性修正为 0.78
- 证据 $P(\text{清晰}) = 9/17$ ：17 个瓜中 9 个纹理清晰

记号 数据：周志华《机器学习》第 4 章 4.4.1 节，表 4.3（西瓜数据集 3.0），p.84：好瓜 8 个、坏瓜 9 个；纹理清晰的好瓜 7 个、坏瓜 2 个。贝叶斯公式见第 7 章 7.1 节，式 (7.8)，p.148。

条件风险

后验描述“是哪一类”的不确定性；损失描述“判错的代价”

判为 \ 真实	好瓜	坏瓜
好瓜	0	$\mathcal{L}(\text{好}, \text{坏})$
坏瓜	$\mathcal{L}(\text{坏}, \text{好})$	0

买回坏瓜与错过好瓜的代价可以不同

$$R(c_i | \mathbf{x}) = \sum_{j=1}^K \mathcal{L}(c_i, c_j) P(c_j | \mathbf{x})$$

- **条件风险**：在 $\mathbf{x}$ 处判为 c_i 的期望损失
- **判定准则** $h: \mathbf{x} \mapsto c$ 的**总体风险**

$$R(h) = \mathbb{E}_{\mathbf{x}}[R(h(\mathbf{x}) | \mathbf{x})]$$

记号 $\mathcal{L}(c_i, c_j)$ ：真实类别为 c_j 时判为 c_i 的损失（周志华《机器学习》记 λ_{ij} ；Bishop 记 L_{kj} ，行为真实类别、列为判定，与此处转置）； $R(c_i | \mathbf{x})$ ：条件风险。周志华《机器学习》第 7 章 7.1 节，式 (7.1)–(7.3)，p.147；Bishop (2006) 第 1.5.2 节，p.41。

贝叶斯判定准则

逐个样本最小化条件风险，总体风险随之最小

$$h^*(\boldsymbol{x}) = \arg \min_{c \in \{c_1, \dots, c_K\}} R(c | \boldsymbol{x})$$

- h^* 称为贝叶斯最优分类器
- $R(h^*)$ 称为贝叶斯风险
- $R(h^*)$ 是给定损失下可达到的最小期望损失；0-1 损失下即最小错误率， $1 - R(h^*)$ 为最优精度

理由： $R(h) = \mathbb{E}_x[R(h(x) | x)]$ 中，每个 x 处的判定互不影响；每一处都取最小，期望也最小。

要点 贝叶斯最优分类器需要**真实后验** $P(c | \boldsymbol{x})$ ；机器学习要做的，是用有限样本尽可能准确地估计它。

贝叶斯判定准则与经验风险最小化

两者使用同一个损失 $\mathcal{L}$ ，最小化同一个目标：期望风险

$$R(h) = \mathbb{E}_{(x,y)} [\mathcal{L}(h(\mathbf{x}), y)] = \mathbb{E}_{\mathbf{x}} \left[\sum_j \mathcal{L}(h(\mathbf{x}), c_j) P(c_j | \mathbf{x}) \right]$$

经验风险最小化（前几讲）

期望的计算	用训练集上的样本平均近似
求解的对象	模型参数，训练时求一次
损失的作用	决定参数，固化在模型中

贝叶斯判定准则

已知后验，按后验精确求期望
每个 x 的类别，预测时逐个求
决定判定，可随时更换

要点 后验已知时，贝叶斯判定准则是期望风险的**精确最小解**；后验未知时，经验风险最小化用样本近似。

记号 $\mathbb{E}_{(x,y)}$ ：对样本与标签的联合分布求期望；第二个等号先对 y 按后验求期望，再对 x 求期望。

例：0-1 损失下的贝叶斯判定

判对损失为 0，判错损失为 1：最小化分类错误率

$$\mathcal{L}(c_i, c_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$

$$R(c | \mathbf{x}) = \sum_{c_j \neq c} P(c_j | \mathbf{x}) = 1 - P(c | \mathbf{x})$$

$$h^*(\mathbf{x}) = \arg \max_c P(c | \mathbf{x})$$

- 选后验概率最大的类别
- 由贝叶斯公式，等价于选 $P(\mathbf{x} | c)P(c)$ 最大的类别

例 纹理清晰的瓜： $P(\text{好} | \text{清晰}) = 0.78 > P(\text{坏} | \text{清晰}) = 0.22$ ，判为**好瓜**。

代价敏感的判定阈值

二分类：记 $p = P(\text{好} | x)$ ，比较两种判定的条件风险

判定	条件风险
判为 好瓜	$\mathcal{L}(\text{好}, \text{坏}) (1 - p)$
判为 坏瓜	$\mathcal{L}(\text{坏}, \text{好}) p$

判为好瓜，当且仅当

$$p > \frac{\mathcal{L}(\text{好}, \text{坏})}{\mathcal{L}(\text{好}, \text{坏}) + \mathcal{L}(\text{坏}, \text{好})}$$

例 设 $\mathcal{L}(\text{好}, \text{坏}) = 4$, $\mathcal{L}(\text{坏}, \text{好}) = 1$

阈值 $= 4/(4 + 1) = 0.8$

纹理清晰的瓜： $p = 0.78 < 0.8$ ，判为坏瓜

两种损失相等时阈值为 0.5，回到最大后验判定。

要点 损失只移动阈值，不改变后验。

记号 例中的损失为假设值：把坏瓜判为好瓜（买回坏瓜）的损失是把好瓜判为坏瓜（错过好瓜）的 4 倍。补充示例（非教材例题）：损失矩阵与代价敏感判定见 Bishop (2006) 第 1.5.2 节 p.41 的癌症诊断例。

类别不平衡问题

类别不平衡 (class-imbalance): 不同类别的训练样例数目差别很大

- 998 个反例、2 个正例：永远预测反例即可达到 99.8% 的精度，但毫无价值
- 分类器以 $\frac{y}{1-y} > 1$ 判为正例；若希望正、反例同等对待，而训练集正、反例数目不同，应以观测几率 $\frac{m^+}{m^-}$ 为阈值

$$\text{再缩放 (rescaling): } \frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$$

与开头“类别比例变化后的后验修正”是同一个思想

方法	基本思路
欠采样	去除一些反例，使正、反例数目接近；代表 EasyEnsemble
过采样	增加一些正例；代表 SMOTE（插值产生额外正例，而非简单重复）
阈值移动	用原始训练集学习，预测时嵌入再缩放

要点 再缩放也是代价敏感学习的基础：把 m^-/m^+ 换成误分类代价之比 $cost^+/cost^-$ 。

记号 y : 分类器预测为正例的可能性； m^+ 、 m^- : 正、反例数目；欠采样 (undersampling)、过采样 (oversampling)、阈值移动 (threshold-moving)。周志华《机器学习》第 3 章 3.6 节，式 (3.46)–(3.48)，pp.66–67。

类别不平衡的处理方法

两类途径：改变训练数据，或保持训练数据、调整决策

方法	做法	出处
改变训练数据		
欠采样	去除部分反例；随机丢弃会损失信息， EasyEnsemble 把反例分给多个学习器	西瓜书 p.67
过采样	增加正例；简单重复导致严重过拟合， SMOTE 插值产生新正例	西瓜书 p.67
均衡采样 + 先验补偿	用各类等量的数据学习后验，再除以训练集类别比例、乘以总体类别比例，并归一化	PRML p.45
保持训练数据，调整决策		
阈值移动	用原始训练集学习，预测时按式 (3.48) 再缩放	西瓜书 pp.66–67
代价敏感学习	以误分类代价之比 $cost^+/cost^-$ 代替 m^-/m^+ ；或用损失矩阵最小化期望损失	西瓜书 pp.35–36, 67; PRML pp.41–42

要点 评估不用精度：永远预测多数类即可达到 99.8%（西瓜书）或 99.9%（PRML）。

记号 先验补偿需要模型输出后验概率（PRML p.46）。一对其余（OvR）拆分也会产生不平衡（西瓜书 p.66；PRML p.338）。

后验从何而来

上面的判定都需要后验 $P(c | \mathbf{x})$ ；从训练样本得到它，有三种途径

途径	学习的对象	后验	代表
生成式模型	先验 $P(c)$ 与似然 $P(\mathbf{x} c)$	由贝叶斯公式计算	朴素贝叶斯
判别式模型	直接学习 $P(c \mathbf{x})$	模型直接输出	逻辑回归、神经网络
判别函数	$\mathbf{x} \mapsto$ 类别	不提供	支持向量机

- 前几讲的模型属于后两种；本讲的贝叶斯分类器属于生成式模型
- 生成式模型描述“每一类的样本如何产生”，可以由 $P(\mathbf{x} | c)$ 采样生成样本

记号 生成式模型 (generative model)；判别式模型 (discriminative model)。三种途径的比较见 Bishop (2006) 第 1.5.4 节；决策树叶结点的类别频率也可作后验的粗略估计。

估计似然的困难

生成式模型的难点： $P(\mathbf{x} | c)$ 是 $\mathbf{x}$ 全部属性的联合分布

- 西瓜数据集的 6 个离散属性：5 个各 3 种取值，1 个 2 种取值
- 属性组合共 $3^5 \times 2 = 486$ 种
- 训练样本只有 17 个，绝大多数组合从未出现
- 未出现 $\neq$ 概率为 0：直接数频率不可行

两个假设

1. 分布形式已知（伯努利、高斯等），只需估计少数参数：§2
2. 给定类别时各属性相互独立，联合分布拆成单个属性的分布之积：§3

问题 假设分布形式之后， $P(\mathbf{x} | c)$ 由若干参数确定。如何由训练样本估计这些参数？

记号 周志华《机器学习》第 7 章 7.1 节 p.148、7.3 节 p.150。离散属性：色泽、根蒂、敲声、纹理、脐部各 3 种取值，触感 2 种取值；§3 的朴素贝叶斯分类器同时使用这两个假设。

02

极大似然估计

Maximum Likelihood Estimation

似然与对数似然 · 抛硬币与西瓜的例子 · 与损失函数、正则项的关系

参数估计

假设 $P(\boldsymbol{x} | c)$ 的分布形式已知，由参数 θ_c 唯一确定，记为 $P(\boldsymbol{x} | \theta_c)$



- 各类别分别估计，互不影响
- 极大似然估计（Maximum Likelihood Estimation, MLE）：选使 D_c 出现概率最大的 θ_c

要点 估计一个分布的问题，转化为估计若干个参数的问题。

记号 θ_c ：第 c 类的分布参数，如伯努利分布取 1 的概率、高斯分布的均值与方差； $\hat{\theta}_c$ ：其估计值。周志华《机器学习》第 7 章 7.2 节，pp.149–150。

似然函数

训练样本已经观察到、固定不变；把它们出现的概率看作参数的函数

$$P(D_c | \theta_c) = \prod_{x \in D_c} P(x | \theta_c), \quad \hat{\theta}_c = \arg \max_{\theta_c} P(D_c | \theta_c)$$

- 样本独立同分布，联合概率为连乘
- 自变量是参数 θ_c ，数据 D_c 是常量

直接最大化连乘的两个困难

数值下溢：200 个约为 0.01 的因子相乘得 10^{-400} ，小于双精度浮点数能表示的最小正数（约 4.9×10^{-324} ），计算结果为 0

求导： $|D_c|$ 个因子之积的导数按乘积法则展开为 $|D_c|$ 项

问题 能否把连乘换成更易计算的形式，同时不改变最大值点？

记号 独立同分布：样本相互独立，且来自同一分布；此时 $P(D_c | \theta_c)$ 称为参数 θ_c 的似然。周志华《机器学习》第 7 章 7.2 节，式 (7.9)，p.149。

对数似然

$\ln$ 单调递增：取对数不改变最大值点，连乘变为连加

$$LL(\theta_c) = \ln P(D_c | \theta_c) = \sum_{x \in D_c} \ln P(x | \theta_c), \quad \hat{\theta}_c = \arg \max_{\theta_c} LL(\theta_c)$$

- 上页的例子： $200 \ln 0.01 \approx -921$ ，可正常表示
- 求导逐项进行： $\nabla LL = \sum \nabla \ln P(x | \theta_c)$

计算后验时同样取对数

各类的对数得分 $s_c = \ln P(c) + \ln P(x | c)$

$$P(c | x) = \frac{e^{s_c - s_{\max}}}{\sum_k e^{s_k - s_{\max}}}$$

减去 $s_{\max} = \max_k s_k$ 不改变比值，指数不会全部下溢

记号 LL ：对数似然 (log-likelihood)； $\ln$ 为自然对数；后验的计算式与 softmax 相同。只需判定类别时，直接比较 s_c 。周志华《机器学习》第 7 章 7.2 节，式 (7.10)–(7.11)，pp.149–150。

例：抛硬币

模型：一次抛掷正面朝上的概率为 θ 。观察：抛 100 次，正面朝上 90 次

参数（待估计） $\theta \in [0, 1]$

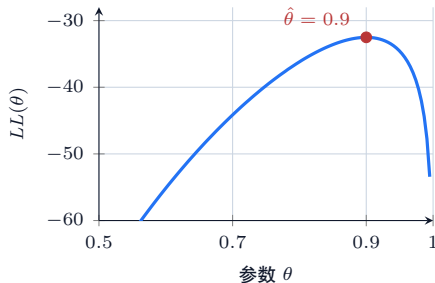
数据（固定） 90 次正面、10 次反面

$$P(D | \theta) = \theta^{90} (1 - \theta)^{10}$$

$$LL(\theta) = 90 \ln \theta + 10 \ln(1 - \theta)$$

对 θ 求导并令其为 0：

$$\frac{90}{\theta} - \frac{10}{1 - \theta} = 0 \implies \hat{\theta} = \frac{90}{100} = 0.9$$



要点 优化的变量是参数 θ ；数据固定不变。

记号 补充示例（非教材例题）；同类例子见 Bishop (2006) 第 2.1 节，pp.68–69。 D 指有序的抛掷序列。一般地， n 次中 k 次正面： $\hat{\theta} = k/n$ ，即频率。二阶导数 $-\frac{90}{\theta^2} - \frac{10}{(1-\theta)^2} < 0$ ，驻点为最大值点。

似然与损失函数

判别式模型也可用极大似然估计：对 $P(y | x)$ 作分布假设，模型 $f_\theta(x)$ 给出分布的参数

$$\max_{\theta} \sum_{i=1}^m \ln P(y_i | \mathbf{x}_i; \theta) \iff \min_{\theta} \frac{1}{m} \sum_{i=1}^m \underbrace{-\ln P(y_i | \mathbf{x}_i; \theta)}_{\text{损失 } \mathcal{L}(f_\theta(\mathbf{x}_i), y_i)}$$

对 y 的分布假设	$-\ln P(y x)$	损失	模型
高斯： $y \sim \mathcal{N}(\hat{y}, \sigma^2)$	$\frac{1}{2\sigma^2}(y - \hat{y})^2 + \text{常数}$	平方损失	线性回归
伯努利： $P(y = 1) = \hat{y}$	$-y \ln \hat{y} - (1 - y) \ln(1 - \hat{y})$	对数损失	逻辑回归
类别分布： $P(y = c) = \hat{y}_c$	$-\ln \hat{y}_y$	交叉熵	神经网络分类器

要点 平方损失、对数损失、交叉熵是相应分布假设下的**负对数似然**；hinge loss 等则没有这种对应。

记号 $\hat{y} = f_\theta(x)$ ：模型输出； σ^2 ：噪声方差，与 θ 无关，不影响最优解； $\hat{y}_y$ ：模型给真实类别 y 的概率。抛硬币即伯努利情形。

先验与正则项

把参数 θ 也看作随机变量，赋予先验密度 $p(\theta)$ ，取其后验的最大值点

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \ln p(\theta | D) = \arg \min_{\theta} \left[\underbrace{\sum_i -\ln P(y_i | \mathbf{x}_i; \theta)}_{\text{损失}} - \underbrace{\ln p(\theta)}_{\text{正则项}} \right]$$

参数的先验（各分量独立）	$-\ln p(\theta)$	正则项	例
高斯： $\theta_j \sim \mathcal{N}(0, \tau^2)$	$\frac{1}{2\tau^2} \ \theta\ _2^2 + \text{常数}$	L_2 正则（权重衰减）	岭回归
拉普拉斯： $p(\theta_j) \propto e^{- \theta_j /b}$	$\frac{1}{b} \ \theta\ _1 + \text{常数}$	L_1 正则	稀疏线性回归

要点 损失为负对数似然、正则项为负对数先验时，带正则的经验风险最小化即**最大后验估计**。

记号 最大后验估计（Maximum A Posteriori, MAP）；第一个等号用到 $p(\theta | D) \propto P(D | \theta) p(\theta)$ ； τ^2 、 b 越小，先验越集中于 0，正则越强。参见 Bishop (2006) 第 1.2.5 节 p.30、第 3.1.4 节 pp.144–146。

例：由纹理判断西瓜好坏

模型：只用“纹理”一个属性的贝叶斯分类器（第 1 节的例子）

待估计的分布	参数	数据（计数）	极大似然估计
先验 $P(c)$	$P(\text{好})$	17 个瓜中好瓜 8 个	$8/17$
似然 $P(\text{纹理} \mid \text{好})$	$\theta_{\text{清晰}}, \theta_{\text{稍糊}}, \theta_{\text{模糊}}$	8 个好瓜：清晰 7、稍糊 1、模糊 0	$7/8, 1/8, 0$
似然 $P(\text{纹理} \mid \text{坏})$	同上，另一组	9 个坏瓜：清晰 2、稍糊 4、模糊 3	$2/9, 4/9, 3/9$

- 属性有多个取值时，极大似然估计仍是频率： $\hat{\theta}_v = k_v / |D_c|$ ，是抛硬币的推广
- 代入贝叶斯公式： $P(\text{好} \mid \text{清晰}) = 7/9$ ，即第 1 节的结果

注意 $\hat{P}(\text{模糊} \mid \text{好}) = 0$ ：没有观察到的取值被估计为不可能，§3 讨论其修正。

记号 θ_v ：取值 v 的概率，各取值之和为 1； k_v ：取值 v 的计数； $|D_c|$ ：该类样本数。数据：周志华《机器学习》表 4.3，p.84；估计式见第 7 章 7.3 节，式 (7.16)–(7.17)，p.151。

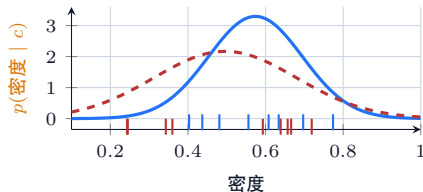
例：由密度判断西瓜好坏

连续属性“密度”：假设每一类的密度服从一维高斯分布 $\mathcal{N}(\mu_c, \sigma_c^2)$

$$LL(\mu, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

参数为每类的 μ, σ^2 ；求偏导并令其为 0：

$$\hat{\mu} = \frac{1}{N} \sum_i x_i, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2$$



	$\hat{\mu}$	$\hat{\sigma}$
好瓜 (8 个)	0.574	0.121
坏瓜 (9 个)	0.496	0.184

问题 以上每次只用一个属性。同时使用全部 8 个属性时， $P(\mathbf{x} | c)$ 是联合分布，如何估计？

记号 实线与上方刻度：好瓜；虚线与下方刻度：坏瓜。周志华《机器学习》第 7 章 7.2 节，式 (7.12)–(7.13)，p.150（分母 $|D_c|$ ；分母取 $|D_c| - 1$ 为无偏估计，§3 的算例与书一致采用后者）；数据见表 4.3，p.84。

03

朴素贝叶斯分类器

Naive Bayes Classifier

联合分布的参数量 · 属性条件独立性假设 · 完整算例 · 拉普拉斯修正

联合分布的参数量

使用全部属性时， $P(\mathbf{x} | c)$ 是 d 个属性的联合分布：每一种属性组合都是一个参数

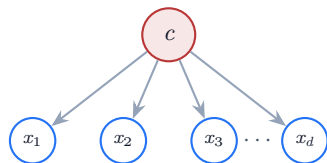
	每类的自由参数个数	可用的训练样本
d 个二值属性	$2^d - 1$	
西瓜数据集 6 个离散属性	$486 - 1 = 485$	17 个（好瓜 8、坏瓜 9）
$d = 100$ 个二值属性	$\approx 1.3 \times 10^{30}$	

问题 参数远多于样本，联合分布无法可靠估计。能否用一个假设，把参数量降到与 d 同阶？

记号 取 $|V_j|$ 种值的离散属性组合共 $\prod_j |V_j|$ 种，各组合概率之和为 1，故自由参数为 $\prod_j |V_j| - 1$ 。数据：周志华《机器学习》表 4.3，p.84。

属性条件独立性假设

给定类别后，各属性相互独立（第 1 节中的第二个假设）



属性之间没有连接

$$P(\mathbf{x} | c) = \prod_{j=1}^d P(x_j | c)$$

$$h_{nb}(\mathbf{x}) = \arg \max_c P(c) \prod_{j=1}^d P(x_j | c)$$

- 独立性只在给定类别的条件下成立，不要求各属性在全体样本中独立
- 联合分布化为 d 个一维分布，每个都用第 2 节的方法估计
- 西瓜数据集 6 个离散属性：每类参数由 485 个降为 $5 \times 2 + 1 = 11$ 个

记号 朴素贝叶斯分类器 (naive Bayes classifier); x_j : $\mathbf{x}$ 在第 j 个属性上的取值; h_{nb} : 朴素贝叶斯的判定准则 (0-1 损失)。周志华《机器学习》第 7 章 7.3 节, 式 (7.14)–(7.15), pp.150–151; Bishop (2006) 第 8.2.2 节。

朴素贝叶斯的参数估计

各项都是第 2 节的极大似然估计

待估计项	估计式	对应的极大似然估计
类先验	$\hat{P}(c) = \frac{ D_c }{ D }$	类别的频率
离散属性	$\hat{P}(x_j c) = \frac{ D_{c,x_j} }{ D_c }$	该类中取值的频率
连续属性	$p(x_j c) = \frac{1}{\sqrt{2\pi} \sigma_{c,j}} \exp\left(-\frac{(x_j - \mu_{c,j})^2}{2\sigma_{c,j}^2}\right)$	一维高斯: $\mu_{c,j}, \sigma_{c,j}^2$ 取样本均值与方差

要点 训练只需一遍计数；新样本到来时，更新计数即可。

记号 D_{c,x_j} : D_c 中第 j 个属性取值为 x_j 的样本集合； $\mu_{c,j}, \sigma_{c,j}^2$: 第 c 类样本在第 j 个属性上的均值与方差。周志华《机器学习》第 7 章 7.3 节，式 (7.16)–(7.18)，p.151。

朴素贝叶斯算例：西瓜分类

Step 1 先验 → Step 2 离散属性 → Step 3 连续属性（如有） → Step 4 连乘判定：测试样本是否为好瓜

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否
测试	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

记号 西瓜数据集 3.0：周志华《机器学习》第 4 章 4.4.1 节，表 4.3，p.84；测试样本即第 7 章 7.3 节 p.151 的“测 1”，与 1 号瓜相同。

Step 1: 从数据中估计先验 $\hat{P}(c)$

数一数每一类有多少个瓜，除以总数

$$\hat{P}(c) = \frac{|D_c|}{|D|}$$

类别	训练样本	$\hat{P}(c)$
好瓜	1-8 号，共 8 个	$8/17 = 0.471$
坏瓜	9-17 号，共 9 个	$9/17 = 0.529$

要点 类先验只由类别比例决定，与测试样本无关。

记号 周志华《机器学习》第 7 章 7.3 节，式 (7.16)，p.151。

Step 2: 从数据中估计离散属性的条件概率 $\hat{P}(x_j | c)$

在每一类中，数出与测试样本取值相同的瓜，除以该类的瓜数

$$\hat{P}(x_j | c) = \frac{|D_{c,x_j}|}{|D_c|} \quad \text{例：8 个好瓜中青绿 3 个（1、4、6 号），} \hat{P}(\text{青绿} | \text{好}) = 3/8$$

	青绿	蜷缩	浊响	清晰	凹陷	硬滑
好瓜	$\frac{3}{8} = 0.375$	$\frac{5}{8} = 0.625$	$\frac{6}{8} = 0.750$	$\frac{7}{8} = 0.875$	$\frac{5}{8} = 0.625$	$\frac{6}{8} = 0.750$
坏瓜	$\frac{3}{9} = 0.333$	$\frac{3}{9} = 0.333$	$\frac{4}{9} = 0.444$	$\frac{2}{9} = 0.222$	$\frac{2}{9} = 0.222$	$\frac{6}{9} = 0.667$

要点 差异最大的是纹理与脐部：好瓜多为清晰、凹陷。

记号 周志华《机器学习》第 7 章 7.3 节，式 (7.17)，pp.151–152。书中两处印刷有误： $P_{\text{蜷缩}|\text{是}}$ 印作 $5/8 = 0.375$ （应为 0.625）， $P_{\text{凹陷}|\text{是}}$ 印作 $6/8 = 0.750$ （按表 4.3 应为 $5/8 = 0.625$ ）；此处按数据取正确值。

Step 3: 从数据中估计连续属性的概率密度 $\hat{p}(x_j | c)$ (如有)

密度、含糖率各用一维高斯：先求该类样本的均值 $\hat{\mu}$ 、标准差 $\hat{\sigma}$ ，再代入测试样本的取值

	密度 = 0.697		含糖率 = 0.460	
	$\hat{\mu}, \hat{\sigma}$	$p(x_j c)$	$\hat{\mu}, \hat{\sigma}$	$p(x_j c)$
好瓜	0.574, 0.129	1.959	0.279, 0.101	0.788
坏瓜	0.496, 0.195	1.203	0.154, 0.108	0.066

以好瓜的密度为例：

$$p(\text{密度} = 0.697 | \text{好}) = \frac{1}{\sqrt{2\pi} \times 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \times 0.129^2}\right) \approx 1.959$$

记号 密度函数值可以大于 1，它不是概率。 $\hat{\sigma}^2$ 取无偏估计（分母 $|D_c| - 1$ ）；第 2 节的极大似然估计分母为 $|D_c|$ 。例题：周志华《机器学习》第 7 章 7.3 节，式 (7.18)，p.152。

Step 4: 连乘各项，比较两类的得分

每一类把 9 个因子连乘：Step 1 的先验 $\times$ Step 2 的 6 个概率 $\times$ Step 3 的 2 个密度

好瓜： $0.471 \times 0.375 \times 0.625 \times 0.750 \times 0.875 \times 0.625 \times 0.750 \times 1.959 \times 0.788 \approx 5.24 \times 10^{-2}$

坏瓜： $0.529 \times 0.333 \times 0.333 \times 0.444 \times 0.222 \times 0.222 \times 0.667 \times 1.203 \times 0.066 \approx 6.80 \times 10^{-5}$

归一化得后验：

$$\frac{P(\text{好} | \mathbf{x}) = 5.24 \times 10^{-2}}{5.24 \times 10^{-2} + 6.80 \times 10^{-5}} \approx 0.999$$

对数得分（第 2 节）：

$$s_{\text{好}} \approx -2.95, \quad s_{\text{坏}} \approx -9.60$$

要点 好瓜的得分约为坏瓜的 770 倍，判为**好瓜**。

记号 得分为 $\hat{P}(c) \prod_j \hat{P}(x_j | c)$ ，与后验只差公共的证据 $P(\mathbf{x})$ ； s_c 为其自然对数。例题：周志华《机器学习》第 7 章 7.3 节，p.153；书中好瓜的得分印作 0.038，是沿用了上页两处印刷错误的结果，坏瓜 6.80×10^{-5} 与书一致。

零概率问题

某个属性值在某类中从未出现：频率估计为 0，连乘结果为 0

- 8 个好瓜中，敲声为“清脆”的 0 个
- $\hat{P}(\text{清脆} | \text{好}) = 0/8 = 0$
- 敲声清脆的瓜，无论其他属性多像好瓜：

$$\hat{P}(\text{好}) \prod_j \hat{P}(x_j | \text{好}) = 0$$

好瓜中计数为 0 的取值

根蒂	硬挺
敲声	清脆
纹理	模糊
脐部	平坦

问题 17 个样本中没有观察到，不代表真实概率为 0；一个零因子会抹掉其余全部属性的信息。如何修正？

拉普拉斯修正

给每个可能的取值预先加一个计数

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + K}, \quad \hat{P}(x_j | c) = \frac{|D_{c,x_j}| + 1}{|D_c| + |V_j|}$$

	修正前	修正后
$\hat{P}(\text{好})$	$8/17 = 0.471$	$9/19 = 0.474$
$\hat{P}(\text{清脆} \text{好})$	$0/8 = 0$	$1/11 = 0.091$
$\hat{P}(\text{青绿} \text{好})$	$3/8 = 0.375$	$4/11 = 0.364$
$\hat{P}(\text{青绿} \text{坏})$	$3/9 = 0.333$	$4/12 = 0.333$

- 修正后的估计不再为 0，且各取值概率之和仍为 1
- 样本越多，修正的影响越小
- 算例的测试样本： 2.56×10^{-2} 对 7.72×10^{-5} ，仍判为好瓜

记号 拉普拉斯修正 (Laplacian correction)，亦称拉普拉斯平滑； K ：类别数（书中记 N ）； $|V_j|$ ：第 j 个属性可能的取值数（书中记 N_i ，敲声为 3）。例题：周志华《机器学习》第 7 章 7.3 节，式 (7.19)–(7.20)，pp.153–154；最后一条为按修正后各项重算的结果。

应用：文本分类

属性数量大、每类样本相对少，独立性假设的优势明显

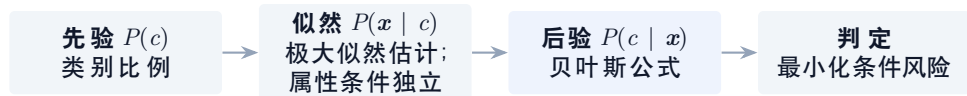


事件模型	属性	类条件分布
伯努利模型	每个词是否出现（未出现也计入）	每个词一个伯努利分布
多项式模型	每个词出现的次数	词袋上的多项式分布

- Sahami 等（1998）用朴素贝叶斯过滤垃圾邮件
- McCallum & Nigam（1998）：词表较大时，多项式模型的准确率通常更高

记号 Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. AAAI Workshop on Learning for Text Categorization; McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. 同一研讨会。

贝叶斯分类器：小结



- 贝叶斯最优分类器是**理论上限**；实际性能取决于后验的估计质量
- 损失矩阵决定判定阈值
- 朴素贝叶斯以独立性假设换取**低方差**与极低的训练开销
- 零计数用拉普拉斯修正；连乘用对数域

问题 朴素贝叶斯依赖分布形式与独立性假设；属性相关时（如图像的相邻像素）假设不成立。能否不作这些假设，直接估计后验？

04

k 近邻学习

k-Nearest Neighbor

后验的局部估计 · 算法 · 错误率分析 · 维数灾难

估计后验的两种假设

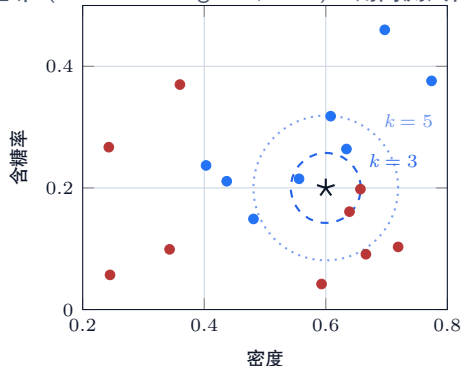
不假设分布形式，只假设“相近的样本类别相近”：用 x 附近的样本估计后验

	朴素贝叶斯	k 近邻
假设	给定类别时，各属性条件独立	后验在 x 附近变化平缓
估计方式	逐个属性估计分布，再连乘	统计 x 邻域内各类样本的比例
模型	先验与各属性分布的参数	非参数：保存全部训练样本
失效情形	属性强相关	高维时邻域内样本稀少

要点 两者都在近似贝叶斯最优分类器：前者是全局的参数化假设，后者是局部的非参数估计。

k 近邻算法

k 近邻 (k-Nearest Neighbor, kNN): 用离测试样本最近的 k 个训练样本做预测



1. 计算测试样本到所有训练样本的距离
2. 取距离最近的 k 个样本
3. 分类取多数投票，回归取平均

要点 多数投票，就是用邻域内各类的比例估计后验 $P(c | x)$ ，再取最大者：本质上是贝叶斯分类（下页推导）。

记号 没有显式的训练过程，计算推迟到预测时，称懒惰学习 (lazy learning)。图中为西瓜数据集 3.0（周志华《机器学习》第 4 章 4.5 节，表 4.5，p.89）；测试样本为补充示例（教材图 10.1 为示意，p.225）：● 好瓜，● 坏瓜，★ 为测试样本 (0.6, 0.2)；虚线圆内为最近的 3 个样本，点线圆内为最近的 5 个样本。

k 近邻对后验的估计

以 $\mathbf{x}$ 为中心取一个邻域，恰好包含 k 个训练样本，体积为 V ，其中第 c 类有 k_c 个

$$p(\mathbf{x}) \approx \frac{k}{m V}, \quad p(\mathbf{x} | c) \approx \frac{k_c}{|D_c| V}, \quad P(c) \approx \frac{|D_c|}{m}$$

$$\implies P(c | \mathbf{x}) = \frac{p(\mathbf{x} | c) P(c)}{p(\mathbf{x})} \approx \frac{k_c}{k}$$

- 第一式：邻域足够小时密度近似不变，邻域的概率 $\approx p(\mathbf{x}) V$ ，又 $\approx$ 样本比例 k/m ；各类同理
- 多数投票 = $\arg \max_c k_c/k$ ：最大后验判定的近似；代价敏感时比较 k_c/k 与判定阈值
- 估计准确要求邻域既小又有样本： $k \rightarrow \infty$ 且 $k/m \rightarrow 0$

要点 k 近邻用邻域内的类别比例估计后验，再按贝叶斯判定准则分类。

记号 推导见 Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, 第 2.5.2 节，该书把 k 近邻作为非参数密度估计放在第 2 章；周志华《机器学习》放在第 10 章降维与度量学习的开头。 $|D_c|$ ：第 c 类训练样本数。

k 的取值与距离度量

同一测试样本， k 不同，预测可能不同

k	近邻（编号：类别）	预测
1	5: 好	好瓜
3	5: 好, 13: 坏, 14: 坏	坏瓜
5	再加 3: 好, 4: 好	好瓜

k	1	3	5	7
留一法错误数	1	3	4	5

- k 小：决策边界曲折，对噪声敏感
- k 大：边界平滑，可能淹没少数类
- k 由交叉验证选择
- 距离依赖属性的尺度：先标准化；更一般地，学习距离度量 (§7)

记号 补充示例（非教材例题），数据为西瓜数据集 3.0，p.89。测试样本 (0.6, 0.2)；欧氏距离；留一法：每次取出 1 个样本作测试，其余 16 个作近邻，共 17 次。

最近邻分类器的错误率

Cover & Hart (1967): 样本数趋于无穷时, 最近邻分类器的错误率不超过贝叶斯错误率的两倍

给定测试样本 x 与其最近邻 x' , 二者类别不同即出错:

$$\begin{aligned} P(\text{err}) &= 1 - \sum_c P(c | x)P(c | x') \\ &\approx 1 - \sum_c P^2(c | x) \\ &\leq 1 - P^2(c^* | x) \\ &= (1 + P(c^* | x))(1 - P(c^* | x)) \\ &\leq 2 \times (1 - P(c^* | x)) \end{aligned}$$

- $c^* = h^*(x)$: 贝叶斯最优分类器的预测
- $1 - P(c^* | x)$: 贝叶斯最优分类器在 x 处的错误率
- 第二步要求 x' 离 x 足够近:
 $m \rightarrow \infty$ 时成立; 有限样本下只是近似

记号 推导: 周志华《机器学习》第 10 章 10.1 节, 式 (10.1), pp.225–226; 对 x 求期望即得总体错误率的界。Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

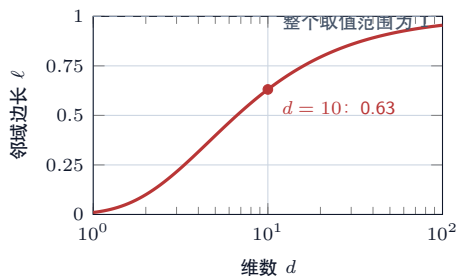
维数灾难：近邻不再“近”

样本均匀分布在单位立方体 $[0, 1]^d$ 中；取一个小立方体作为邻域，使它包含 1% 的样本

立方体邻域的边长记为 ℓ ，体积 ℓ^d 等于它包含的样本比例：

$$\ell^d = 0.01 \implies \ell = 0.01^{1/d}$$

维数 d	1	2	3	10	100
边长 ℓ	0.01	0.10	0.22	0.63	0.95



要点 $d = 10$ 时，只为找 1% 的近邻，邻域就要覆盖每个属性 63% 的取值范围：所谓“近邻”已经很远。

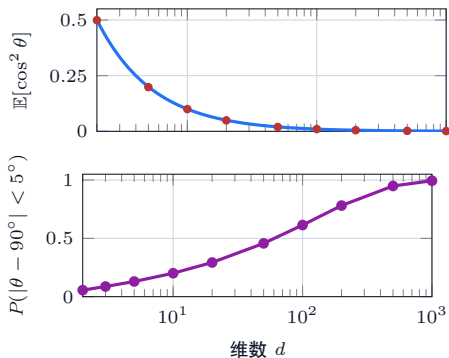
记号 维数灾难 (curse of dimensionality)；右图曲线为 $\ell = 0.01^{1/d}$ ， d 增大时趋于 1。补充示例，参见 Hastie, Tibshirani & Friedman (2009). *The Elements of Statistical Learning*, 第 2.5 节；Bishop (2006) 第 1.4 节，pp.33–38。周志华《机器学习》第 10 章 10.1 节 p.226 的例子：属性归一化后要求每个属性上样本间距不超过 0.001，单个属性需 1000 个样本，20 个属性需 $(10^3)^{20} = 10^{60}$ 个。

维数灾难：随机方向近似正交

u, v 为 $\mathbb{R}^d$ 中独立、均匀的随机单位向量，夹角为 θ ， $\cos \theta = u^\top v$

- 不妨取 $v = (1, 0, \dots, 0)$ ，则 $\cos \theta = u_1$
- u 是单位向量： $u_1^2 + u_2^2 + \dots + u_d^2 = 1$
- 两边取期望： $\mathbb{E}[u_1^2] + \dots + \mathbb{E}[u_d^2] = 1$
- 各坐标地位相同， d 个期望相等

$$\mathbb{E}[\cos^2 \theta] = \mathbb{E}[u_1^2] = \frac{1}{d}$$



要点 $d = 1000$ 时，99.4% 的夹角在 $90^\circ \pm 5^\circ$ 之内：高维随机向量几乎两两垂直。

记号 上图曲线为 $1/d$ ，点为 5000 对随机向量的模拟（随机种子 0）；下图由 $\cos \theta$ 的精确分布积分得到。

维数灾难：距离趋于集中

x, z 独立、均匀地取自 $[0, 1]^d$, 记 $\xi_j = (x_j - z_j)^2$

$$\|x - z\|^2 = \sum_{j=1}^d \xi_j, \quad \xi_j \text{ 独立同分布}$$

$$\mathbb{E}[\xi_j] = \frac{1}{6}, \quad \text{Var}[\xi_j] = \frac{1}{15} - \frac{1}{36} = \frac{7}{180}$$

$$\frac{\text{标准差}(\|x - z\|^2)}{\mathbb{E}\|x - z\|^2} = \frac{\sqrt{7d/180}}{d/6} \approx \frac{1.18}{\sqrt{d}}$$

500 个均匀随机点到一个随机点的距离：

d	2	10	100	1000
$\frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}}$	40.7	4.65	0.36	0.11

距离的均值按 $\sqrt{d/6}$ 增长，相对波动按 $1/\sqrt{d}$ 缩小。

要点 $d = 1000$ 时，最远的点只比最近的点远约 11%：“最近”几乎失去区分度。

记号 右表为一次随机模拟（随机种子 0）。一般条件下的结论见 Beyer, K. et al. (1999). When is “nearest neighbor” meaningful? ICDT.

维数灾难对 k 近邻的影响

k 近邻的两个前提：邻域内有样本；邻域足够小

高维现象	对 k 近邻的后果
邻域边长 $\ell \rightarrow 1$	邻域内的后验与 x 处的后验不同, $P(c x) \approx k_c/k$ 失效
方向近似正交、距离集中	最近与最远几乎一样远, 近邻的排序易受噪声干扰

降低维数

高维数据常分布在低维结构附近；在低维空间中度量距离 (§5-6)

学习距离

由标注信息学习距离，使同类近、异类远 (§7)

05

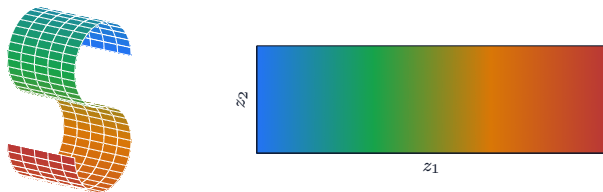
主成分分析

Principal Component Analysis

数据的低维结构 · 投影的两个要求 · 目标函数与求解 · 方差贡献率 · 算例 · 特征脸

数据的低维结构

维数灾难的出路：高维样本常分布在某个低维结构附近



- 样本由 d 个属性描述，有效的自由度往往远小于 d ；如人脸图像主要随姿态、光照等少数因素变化
- 降维：为每个样本找低维坐标 $z_i \in \mathbb{R}^{d'}$ ， $d' \ll d$ ，并保留样本间的关系

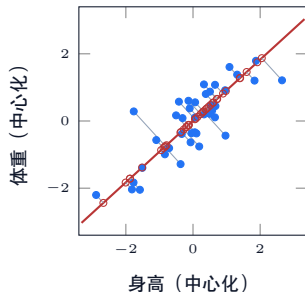
问题 最简单的低维结构是过原点的线性子空间。如何找到最合适的子空间？

记号 颜色表示样本在曲线上的位置，展开后保持不变；左图为 $\mathbb{R}^3$ 中的 S 形曲面（ $d = 3$ ），右图为展开后的二维坐标（ $d' = 2$ ）。对应周志华《机器学习》第 10 章 10.2 节图 10.2（p.227）与 10.5 节图 10.6（p.232）的 S 形曲面，此处为按同一参数化生成的网格。

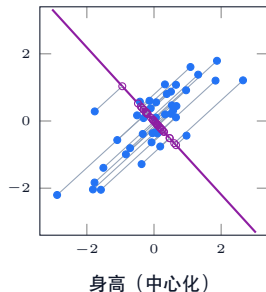
线性降维的直觉：身高与体重

二维样本降为一维：投影到哪条直线上？

方向 w_1 ：重构误差 0.21，投影方差 2.00



方向 w_2 ：重构误差 2.00，投影方差 0.21



两个要求 样本离直线近（重构误差小）； 投影分得开（投影方差大）。左图同时满足。

记号 补充示例（非教材例题），对应周志华《机器学习》第 10 章 10.3 节图 10.4（p.231）的两个投影方向。40 个模拟样本（沿 45° 方向拉长的二维高斯）；灰线：样本到投影点的距离；数值为样本平均。

投影矩阵

d' 维子空间由 d' 个方向向量 $w_1, \dots, w_{d'} \in \mathbb{R}^d$ 张成

$$W = (w_1, \dots, w_{d'}) \in \mathbb{R}^{d \times d'}, \quad W^\top W = I_{d'}$$

各方向单位长度、两两正交

坐标（编码） $z = W^\top x$, $z_j = w_j^\top x$ 为 x 在 w_j 上的投影长度

重构（解码） $\hat{x} = Wz = W(W^\top x)$

重构误差 $\|x - WW^\top x\|^2$

为什么要求 $W^\top W = I$

$z_j = w_j^\top x$ 是投影长度，需要 $\|w_j\| = 1$ ；
方向两两正交，坐标互不重复。

子空间内的点 Wz 再取坐标不变：
 $W^\top(Wz) = z$

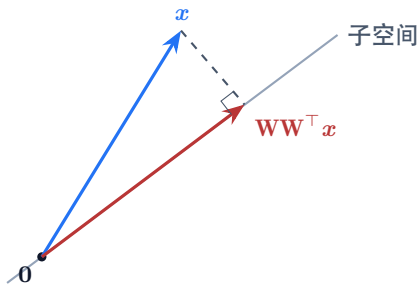
误差从哪里来

$WW^\top$ 是 $d \times d$ 矩阵，秩为 d' 。 $d' < d$ 时 $WW^\top \neq I_d$ ： x 在子空间之外的分量被丢弃。 $d' = d$ 时 $WW^\top = I$ ，误差为 0，但没有降维。

记号 本节假定样本已中心化： $\sum_i x_i = 0$ 。上页样本 1： $x = (2.65, 1.21)$, $z = 2.77$, $\hat{x} = (2.05, 1.87)$, 重构误差 0.80。

两个要求的数学形式

由勾股定理，两个要求是同一个目标



虚线：重构误差

$$\|x_i\|^2 = \underbrace{\|\mathbf{W}\mathbf{W}^\top x_i\|^2}_{=\|\mathbf{W}^\top x_i\|^2} + \underbrace{\|x_i - \mathbf{W}\mathbf{W}^\top x_i\|^2}_{\text{重构误差}}$$

对全部样本求和，左边 $\sum_i \|x_i\|^2$ 与 $\mathbf{W}$ 无关：

$$\min \sum_i \|x_i - \mathbf{W}\mathbf{W}^\top x_i\|^2 \iff \max \sum_i \|\mathbf{W}^\top x_i\|^2$$

右边是投影后的方差（乘以 m ）：中心化后 z_i 的均值为 0

$$\max_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W}) \quad \text{s.t. } \mathbf{W}^\top \mathbf{W} = \mathbf{I} \quad \text{其中 } \sum_i \|\mathbf{W}^\top x_i\|^2 = \text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W})$$

记号 $\|\mathbf{W}\mathbf{W}^\top x\|^2 = x^\top \mathbf{W}(\mathbf{W}^\top \mathbf{W})\mathbf{W}^\top x = \|\mathbf{W}^\top x\|^2$ 用到 $\mathbf{W}^\top \mathbf{W} = \mathbf{I}$; $\mathbf{X} = (x_1, \dots, x_m) \in \mathbb{R}^{d \times m}$; $\frac{1}{m} \mathbf{X} \mathbf{X}^\top$ 为样本协方差矩阵。

用拉格朗日乘子法求解

先看 $d' = 1$: $\max_w w^\top \mathbf{X}\mathbf{X}^\top w$, s.t. $w^\top w = 1$

$$L(w, \lambda) = w^\top \mathbf{X}\mathbf{X}^\top w - \lambda(w^\top w - 1)$$

$$\nabla_w L = 0 \implies \mathbf{X}\mathbf{X}^\top w = \lambda w$$

$$\text{目标值 } w^\top \mathbf{X}\mathbf{X}^\top w = \lambda w^\top w = \lambda$$

- 驻点是 $\mathbf{X}\mathbf{X}^\top$ 的特征向量
- 目标值等于对应的特征值：取最大特征值
- 一般的 d' ：取最大的 d' 个特征值对应的特征向量

要点 $\mathbf{X}\mathbf{X}^\top$ 的特征值 $\lambda_1 \geq \dots \geq \lambda_d \geq 0$ ：保留的投影平方和为 $\sum_{j \leq d'} \lambda_j$ ，最小重构误差之和为舍去的特征值之和 $\sum_{j > d'} \lambda_j$ ；除以 m 即为方差。

记号 L ：拉格朗日函数（第 3 次课）； λ ：拉格朗日乘子； $\mathbf{X}\mathbf{X}^\top$ 对称半正定，特征值非负、特征向量可取为两两正交。

拉格朗日乘子与奇异值

对样本矩阵做奇异值分解，拉格朗日条件的解直接可读出

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top \implies \mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{S}^2\mathbf{U}^\top \implies \mathbf{X}\mathbf{X}^\top \mathbf{u}_j = s_j^2 \mathbf{u}_j$$

拉格朗日条件 $\mathbf{X}\mathbf{X}^\top \mathbf{w} = \lambda \mathbf{w}$ 中的量 恰好等于

主成分方向 $\mathbf{w}_j$	$\mathbf{X}$ 的左奇异向量 $\mathbf{u}_j$	$\mathbf{W} = \mathbf{U}$ 的前 d' 列
拉格朗日乘子 λ_j	奇异值的平方 s_j^2	第 j 个方向上的投影方差 $\times m$
低维坐标 z_i	$\mathbf{S}\mathbf{V}^\top$ 的前 d' 行的第 i 列	$\mathbf{W}^\top \mathbf{X} = \mathbf{S}_{d'} \mathbf{V}_{d'}^\top$

要点 主成分分析 (Principal Component Analysis, PCA) 就是对中心化的样本矩阵做奇异值分解；数值上也比先算 $\mathbf{X}\mathbf{X}^\top$ 再分解更稳定。

记号 奇异值分解 (Singular Value Decomposition, SVD); $\mathbf{U} \in \mathbb{R}^{d \times d}$ 、 $\mathbf{V} \in \mathbb{R}^{m \times m}$ 为正交矩阵; $\mathbf{S}$ 的对角元素 $s_1 \geq s_2 \geq \dots \geq 0$ 为奇异值; $\mathbf{S}^2$ 简记 $\mathbf{S}\mathbf{S}^\top$ 。

主成分分析算法

Require: 样本集 $\{x_1, \dots, x_m\}$, 低维空间维数 d'

- 1: 中心化: $x_i \leftarrow x_i - \frac{1}{m} \sum_l x_l$
- 2: 计算协方差矩阵 $XX^\top$
- 3: 对 $XX^\top$ 做特征值分解 (或对 X 做奇异值分解)
- 4: 取最大的 d' 个特征值对应的单位特征向量 $w_1, \dots, w_{d'}$

Ensure: 投影矩阵 $W = (w_1, \dots, w_{d'})$

- w_j 称为第 j 个主成分方向
- 新样本: 先减去训练集的均值, 再乘 $W^\top$
- 均值与 W 都只在训练集上计算

要点 舍弃的方向方差小, 常对应噪声; 舍弃后有一定的去噪作用。

方差贡献率

第 j 个主成分上的投影方差，等于协方差矩阵的第 j 大特征值 λ_j

$$\text{方差贡献率 } \eta_j = \frac{\lambda_j}{\sum_{i=1}^d \lambda_i}$$

$$\text{累计方差贡献率 } \frac{\sum_{j=1}^k \lambda_j}{\sum_{i=1}^d \lambda_i}$$

- 总方差 $\sum_i \lambda_i = \text{tr}(\mathbf{S}) =$ 各属性方差之和
- η_j : 第 j 个主成分解释的总方差比例
- 取累计方差贡献率 $\geq t$ 的最小 k 作为 d' ;
 t 常取 0.85

要点 舍去部分 $\sum_{j>d'} \lambda_j$ 就是重构误差：累计方差贡献率越高，丢失的信息越少。

记号 $\mathbf{S}$: 样本协方差矩阵。周志华《机器学习》式 (10.18), p.231, 称 t 为重构阈值; Bishop (2006) 式 (12.18), p.565, 最小失真 $J = \sum_{i>M} \lambda_i$ 。

算例：中心化与协方差矩阵

4 个样本、2 个指标（每行一个样本）；协方差分母取 $m - 1$

第 1 步 中心化：每列减去均值 $\mu = (6, 4)$

$$\mathbf{X} = \begin{pmatrix} 2 & 3 \\ 5 & 3 \\ 8 & 3 \\ 9 & 7 \end{pmatrix} \implies \tilde{\mathbf{X}} = \begin{pmatrix} -4 & -1 \\ -1 & -1 \\ 2 & -1 \\ 3 & 3 \end{pmatrix}$$

第 2 步 协方差矩阵

$$\begin{aligned} \mathbf{S} &= \frac{1}{m-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \\ &= \frac{1}{3} \begin{pmatrix} 30 & 12 \\ 12 & 12 \end{pmatrix} = \begin{pmatrix} 10 & 4 \\ 4 & 4 \end{pmatrix} \end{aligned}$$

记号 补充示例（非教材例题）。本例按行存放样本， $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times d}$ ，故写作 $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ ；前面推导按列存放，写作 $\mathbf{X}\mathbf{X}^\top$ 。只做中心化、不做标准化；指标单位不同时需先标准化，见 Bishop (2006) 第 12.1.3 节 p.567。

算例：特征值分解与主成分

第 3 步 对 S 做特征值分解

$$\lambda_1 = 12, \quad \mathbf{w}_1 = \frac{1}{\sqrt{5}}(2, 1)^\top$$

$$\lambda_2 = 2, \quad \mathbf{w}_2 = \frac{1}{\sqrt{5}}(1, -2)^\top$$

第一主成分 $y_1 = \frac{2}{\sqrt{5}}(x_1 - 6) + \frac{1}{\sqrt{5}}(x_2 - 4)$

第 4 步 按方差贡献率取前 d' 个

	贡献率	累计
第一主成分	$12/14 = 85.7\%$	85.7%
第二主成分	$2/14 = 14.3\%$	100%

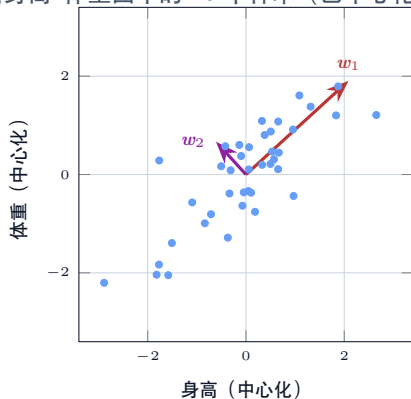
要求累计 $\geq 85\%$: 保留 1 个

要点 方差贡献率 = 该主成分的特征值占全部特征值之和的比率: $\eta_j = \lambda_j / \sum_i \lambda_i$ 。

记号 特征值由 $\det(S - \lambda I) = \lambda^2 - 14\lambda + 24 = 0$ 求得; 特征向量的符号可任取; 分母取 m 时特征值同比缩放, 特征向量与贡献率不变。步骤与周志华《机器学习》图 10.5 的 PCA 算法 (p.231)、Bishop (2006) 第 12.1.1 节 (p.563) 一致。

身高-体重数据的主成分

前面身高-体重图中的 40 个样本（已中心化）



$$\frac{1}{m}\mathbf{X}\mathbf{X}^\top = \begin{pmatrix} 1.185 & 0.894 \\ 0.894 & 1.022 \end{pmatrix}$$

$$\lambda_1 = 2.002, \quad \mathbf{w}_1 = (0.738, 0.674)^\top$$
$$\lambda_2 = 0.205, \quad \mathbf{w}_2 = (-0.674, 0.738)^\top$$

- $\mathbf{w}_1$ 与身高轴成 42° ：身高、体重同增同减的方向
- 第一主成分的方差贡献率
 $2.002/2.207 = 90.7\%$

记号 为便于解释，这里用协方差矩阵 $\frac{1}{m}\mathbf{X}\mathbf{X}^\top$ ，特征向量与 $\mathbf{X}\mathbf{X}^\top$ 相同，特征值相差 m 倍；箭头长度为 $2\sqrt{\lambda_j}$ 。

身高-体重数据的投影与重构

只保留第一主成分 $w_1 = (0.738, 0.674)^\top$

样本 1: $x_1 = (2.652, 1.211)^\top$

$$z_1 = w_1^\top x_1 = 0.738 \times 2.652 + 0.674 \times 1.211 \approx 2.77$$

$$\hat{x}_1 = w_1 z_1 = (2.049, 1.871)^\top$$

$$\|x_1 - \hat{x}_1\|^2 = 0.603^2 + 0.660^2 = 0.80$$

40 个样本的平均	数值	等于
投影方差 $\frac{1}{m} \sum z_i^2$	2.002	λ_1
重构误差	0.205	λ_2
总方差 $\frac{1}{m} \sum \ x_i\ ^2$	2.207	$\lambda_1 + \lambda_2$

要点 总方差 = 保留的方差 + 重构误差：最小重构误差恰为舍去的特征值。

记号 样本已中心化，重构不必再加均值；原始坐标下的重构为 $\mu + w_1 z_1$ 。补充示例（非教材例题），PCA 的教材内容见周志华《机器学习》第 10 章 10.3 节，pp.229–232；Bishop (2006) 第 12.1 节。

直接用图像作为输入的问题

Olivetti 研究实验室人脸数据集：40 人，每人 10 张，共 400 张 112×92 灰度图



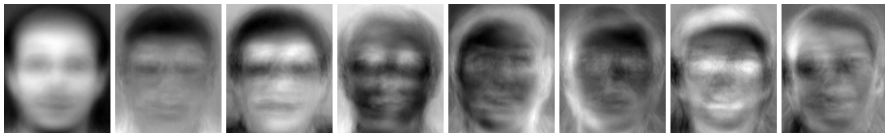
- 每张图展平为 $d = 10304$ 维向量，样本只有 400 个：维数远大于样本数
- 相邻像素高度相关：90% 的方差集中在 111 个方向上
- 1-近邻识别：每次比较需计算 10304 维的距离，且距离受光照、背景像素影响

记号 Samaria, F., & Harter, A. (1994). Parameterisation of a stochastic model for human face identification. IEEE Workshop on Applications of Computer Vision. 每人前 5 张训练、后 5 张测试的 1-近邻正确率：原始像素 90.0%，PCA 降至 50 维 88.5%。

特征脸

Turk & Pentland (1991): 主成分方向 $w_j \in \mathbb{R}^{10304}$ 本身可以显示为一张图像

平均脸与前 7 个主成分（特征脸，eigenface）



用前 d' 个主成分重构同一张脸：人脸 $\approx$ 平均脸 + $\sum_{j \leq d'} z_j w_j$



$d' = 5$

$d' = 20$

$d' = 50$

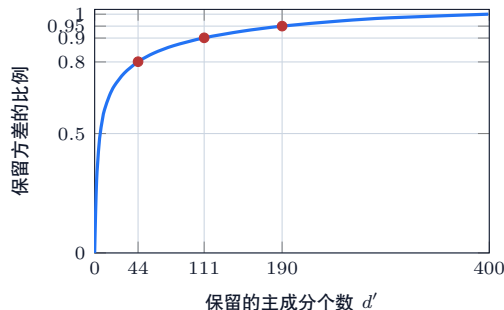
$d' = 150$

原图

记号 特征脸由 400 张图计算：Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 71–86.

维数 d' 的选择

按保留方差的比例，或按下游任务的验证性能



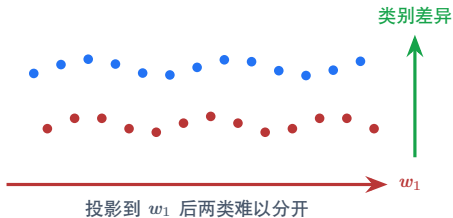
$$\frac{\sum_{j=1}^{d'} \lambda_j}{\sum_{j=1}^d \lambda_j} \geq t$$

- 左式为前 d' 个主成分的累计方差贡献率；取满足上式的最小 d' ， t 常取 0.85–0.95
- 人脸数据： $t = 0.9$ 时 $d' = 111$ ， $t = 0.95$ 时 $d' = 190$
- 降维作为预处理时，也可把 d' 当作超参数，用交叉验证选择

记号 曲线由上页的 400 张人脸图计算； t ：预先设定的阈值； λ_j ： $\mathbf{XX}^\top$ 的第 j 大特征值。

主成分分析的局限

方差大的方向，未必是对分类有用的方向



- PCA **不使用标签**：只保留方差，不保证保留类别信息
- 对属性**尺度敏感**：单位不同的属性需先标准化
- 只能找到线性子空间，且需要样本的**坐标**

问题 有时只知道样本两两之间的距离（或相异程度），没有坐标。能否只由距离求出低维坐标？

记号 左图两类样本沿水平方向分布很长，类别差异在方差很小的竖直方向；使用标签的线性降维有线性判别分析（Linear Discriminant Analysis, LDA）。

06

多维缩放

Multidimensional Scaling

由距离求内积 · 由内积求坐标 · 与主成分分析的关系

多维缩放的目标

多维缩放 (Multidimensional Scaling, MDS): 只给定样本间的距离, 求低维坐标

输入: 距离矩阵 $\mathbf{D} \in \mathbb{R}^{m \times m}$, 元素 dist_{ij}

输出: $\mathbf{Z} = (z_1, \dots, z_m) \in \mathbb{R}^{d' \times m}$, 使

$$\|z_i - z_j\| = \text{dist}_{ij}$$

中间量: 内积矩阵 $\mathbf{B} = \mathbf{Z}^\top \mathbf{Z} \in \mathbb{R}^{m \times m}$,
 $b_{ij} = z_i^\top z_j$

$$\begin{aligned}\text{dist}_{ij}^2 &= \|z_i\|^2 + \|z_j\|^2 - 2z_i^\top z_j \\ &= b_{ii} + b_{jj} - 2b_{ij}\end{aligned}$$

思路

1. 由距离 dist_{ij} 求出内积 b_{ij}
2. 对 $\mathbf{B}$ 做特征值分解, 得到 $\mathbf{Z}$

要点 距离对平移不变, 需先固定位置:
令低维坐标中心化, $\sum_i z_i = 0$ 。

记号 $\mathbf{D}$: 距离矩阵 (与数据集 D 区分); dist_{ij} : 样本 x_i 与 x_j 的距离; m : 样本数。周志华《机器学习》第 10 章 10.2 节, pp.227–229。

由距离求内积矩阵

中心化使 B 的每行、每列之和为 0: $\sum_i b_{ij} = \sum_j b_{ij} = 0$

对 $\text{dist}_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$ 分别按行、按列、全部求平均:

$$\text{dist}_{i\cdot}^2 = \frac{1}{m} \sum_j \text{dist}_{ij}^2 = b_{ii} + \frac{1}{m} \text{tr}(B)$$

$$\text{dist}_{\cdot j}^2 = \frac{1}{m} \sum_i \text{dist}_{ij}^2 = b_{jj} + \frac{1}{m} \text{tr}(B)$$

$$\text{dist}_{\cdot\cdot}^2 = \frac{1}{m^2} \sum_{i,j} \text{dist}_{ij}^2 = \frac{2}{m} \text{tr}(B)$$

消去 $b_{ii}, b_{jj}, \text{tr}(B)$:

$$b_{ij} = -\frac{1}{2} (\text{dist}_{ij}^2 - \text{dist}_{i\cdot}^2 - \text{dist}_{\cdot j}^2 + \text{dist}_{\cdot\cdot}^2)$$

矩阵形式:

$$\mathbf{B} = -\frac{1}{2} \mathbf{J} \mathbf{D}^{(2)} \mathbf{J}, \quad \mathbf{J} = \mathbf{I} - \frac{1}{m} \mathbf{1} \mathbf{1}^\top$$

记号 $\mathbf{D}^{(2)}$: 元素为 dist_{ij}^2 的矩阵; $\mathbf{J}$: 中心化矩阵, 左乘减去列均值、右乘减去行均值; $\mathbf{1}$: 全 1 列向量。

由内积矩阵求坐标

B 对称半正定（距离为欧氏距离时），做特征值分解

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_m), \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$$

$$\mathbf{B} = \mathbf{Z}^\top \mathbf{Z} \implies \mathbf{Z} = \mathbf{\Lambda}_{d'}^{1/2} \mathbf{V}_{d'}^\top \in \mathbb{R}^{d' \times m}$$

- 非零特征值个数 = 精确嵌入所需的维数
- 只保留最大的 d' 个特征值，得到近似的低维嵌入
- 距离不是欧氏距离时，B 可能有负特征值，舍去
- d' 不小于 B 的秩时精确嵌入，坐标与原坐标相差一个旋转或反射； d' 更小时为近似

记号 $\mathbf{V}$ ：特征向量按列排成的正交矩阵； $\mathbf{\Lambda}_{d'}$ ：最大的 d' 个特征值构成的对角矩阵； $\mathbf{V}_{d'}$ ：对应的 d' 个特征向量。

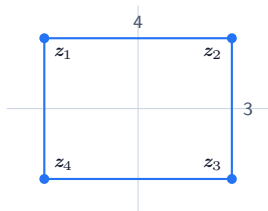
例：矩形的四个顶点

只给出 4 个点两两之间的距离，恢复它们的平面坐标

$$\mathbf{D}^{(2)} = \begin{pmatrix} 0 & 16 & 25 & 9 \\ 16 & 0 & 9 & 25 \\ 25 & 9 & 0 & 16 \\ 9 & 25 & 16 & 0 \end{pmatrix}$$

$$\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J} = \begin{pmatrix} 6.25 & -1.75 & -6.25 & 1.75 \\ -1.75 & 6.25 & 1.75 & -6.25 \\ -6.25 & 1.75 & 6.25 & -1.75 \\ 1.75 & -6.25 & -1.75 & 6.25 \end{pmatrix}$$

特征值：16, 9, 0, 0



$$z_1 = (-2, 1.5), z_2 = (2, 1.5) \\ z_3 = (2, -1.5), z_4 = (-2, -1.5)$$

- 两个非零特征值：二维即可精确嵌入
- 恢复出 4×3 的矩形，中心在原点

记号 4 个点依次为 4×3 矩形的顶点，相邻边长 4、3，对角线 5；特征值 16、9 对应的特征向量分别给出横、纵坐标。补充示例（非教材例题），方法见周志华《机器学习》第 10 章 10.2 节，式 (10.4)–(10.11)，图 10.3，pp.228–229。

多维缩放与主成分分析

距离为欧氏距离时，两者给出相同的低维坐标

$$\mathbf{B} = \mathbf{X}^\top \mathbf{X} = \mathbf{V} \mathbf{S}^\top \mathbf{S} \mathbf{V}^\top \implies Z_{\text{MDS}} = \mathbf{\Lambda}_{d'}^{1/2} \mathbf{V}_{d'}^\top = \mathbf{S}_{d'} \mathbf{V}_{d'}^\top = \mathbf{W}^\top \mathbf{X}$$

	主成分分析	多维缩放
输入	样本坐标 $\mathbf{X}$	两两距离 $\mathbf{D}$
分解的矩阵	$\mathbf{X}\mathbf{X}^\top$ ($d \times d$)	$\mathbf{B}$ ($m \times m$)
距离	欧氏距离	任意的相异度（路程、评分差异等）
新样本	直接计算 $\mathbf{W}^\top \mathbf{x}$	需另行处理

问题 多维缩放与 k 近邻的结果都由距离决定。距离从何而来？能否从数据中学习一个合适的距离？

记号 $\mathbf{X}$ 已中心化， $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ 为奇异值分解； $\mathbf{S}_{d'}$ 、 $\mathbf{V}_{d'}$ ：取前 d' 个奇异值与右奇异向量； $\mathbf{W} = \mathbf{U}$ 的前 d' 列。

07

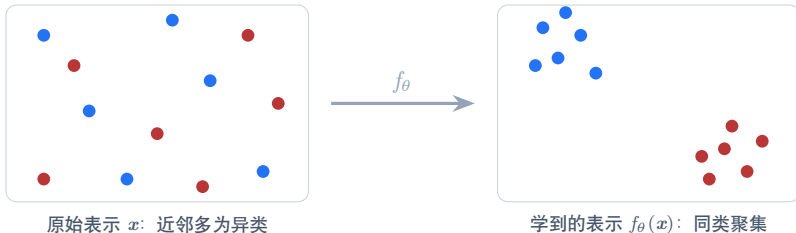
度量学习

Metric Learning

度量学习的对象 · 线性度量学习 · 孪生网络与三元组损失 · 对比学习

度量学习学的是什么

同样是欧氏距离，在不同的表示上计算，近邻关系完全不同



- 主成分分析、多维缩放保留数据自身的结构，不考虑任务
- k 近邻、检索、人脸验证的效果，取决于距离是否反映任务所需的相似性

要点 度量学习 (metric learning) 学习一个映射 f_{θ} ，以 $d_{\theta}(x, x') = \|f_{\theta}(x) - f_{\theta}(x')\|$ 作为距离，目标是**同类距离小、异类距离大**。 f_{θ} 为线性映射 $P^{\top}x$ 时是马氏距离，为神经网络时是深度度量学习。

线性度量学习

$f_{\theta}(x) = \mathbf{P}^{\top} x$: 学习一个线性映射，与 PCA 的 $\mathbf{W}^{\top} x$ 形式相同

$$\text{dist}_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^{\top} \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) = \|\mathbf{P}^{\top} \mathbf{x}_i - \mathbf{P}^{\top} \mathbf{x}_j\|^2, \quad \mathbf{M} = \mathbf{P} \mathbf{P}^{\top} \succeq 0$$

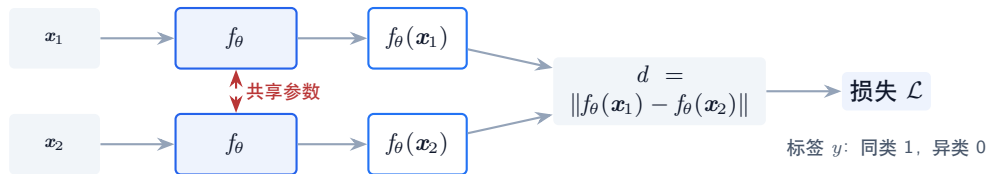
方法	监督信息	目标
近邻成分分析	类别标签	最大化随机 1 近邻的期望留一正确率
成对约束 (Xing 等)	相似 / 不相似的样本对	相似对平方距离之和最小，不相似对距离之和不小于 1；凸问题

问题 $\mathbf{P} \in \mathbb{R}^{d \times d'}$ 时同时完成降维；但线性映射表达能力有限，图像、语音的相似性需要非线性映射。

记号 马氏距离 (Mahalanobis distance), $\mathbf{M} = \Sigma^{-1}$ 时为狭义马氏距离；近邻成分分析 (Neighbourhood Component Analysis, NCA): Goldberger 等 (2004, NIPS); Xing, Ng, Jordan & Russell (2002, NIPS)。周志华《机器学习》第 10 章 10.6 节，式 (10.34)–(10.39), pp.237–240。

孪生网络与对比损失

两个样本输入同一个网络（共享参数），在输出的特征上计算距离



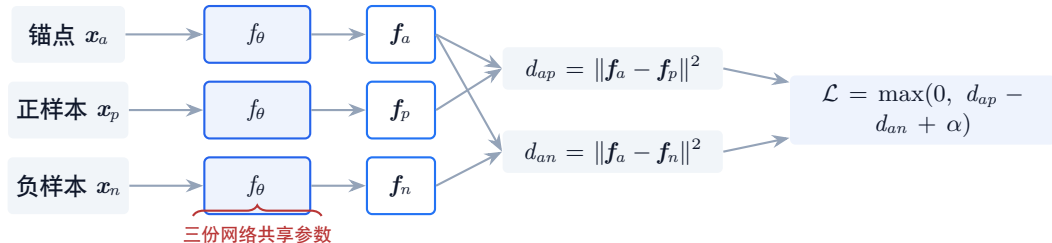
$$\mathcal{L} = y d^2 + (1 - y) \max(0, \alpha - d)^2$$

- 同类对 ($y = 1$): 损失为 d^2 , 训练使两者靠近
- 异类对 ($y = 0$): 距离小于间隔 α 时产生损失, 使两者推远; 超过 α 不再计损失

记号 孪生网络 (siamese network): Bromley 等 (1993, NIPS) 用于签名验证; 对比损失 (contrastive loss): Hadsell, Chopra & LeCun (2006, CVPR); $\alpha > 0$: 间隔。

三元组损失

一次输入三个样本：锚点、同类样本（正样本）、异类样本（负样本）



$d_{ap} + \alpha \leq d_{an}$ 损失为 0：负样本已足够远

$d_{ap} + \alpha > d_{an}$ 梯度使 d_{ap} 减小、 d_{an} 增大

要求同类距离比异类距离至少小 α ，与支持向量机的间隔思想相同。

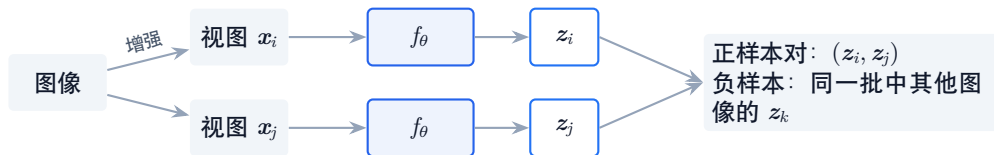
记号 $f_k = f_\theta(x_k)$ ； $\alpha > 0$ ：间隔；三元组损失 (triplet loss)；LFW: Labeled Faces in the Wild 数据集；Schroff, Kalenichenko & Philbin (2015, CVPR)。

FaceNet (Schroff 等, 2015)

f_θ ：卷积神经网络，输出 128 维特征；
两张脸的特征距离小于阈值即判为同一人；LFW 人脸验证准确率 99.63%

对比学习

没有标签时，用数据增强构造“同类”：同一张图像的两个增强视图互为正样本



$$\mathcal{L}_i = -\ln \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}$$

- 形式与 softmax 交叉熵相同：在一批候选中**认出自己的另一个视图**
- 拉近正样本对、推远负样本，**不需要标签**；训练好的 f_θ 用作下游任务的特征

记号 sim ：余弦相似度； $\tau > 0$ ：温度；SimCLR: Chen, Kornblith, Norouzi & Hinton (2020, ICML), 在 ImageNet 上用其特征训练线性分类器，top-1 准确率 76.5%。

降维与度量学习方法的比较

方法	监督信息	映射	优化目标
主成分分析	无	线性 $W^T x$	最小重构误差 / 最大投影方差
多维缩放	无	只给出训练样本的坐标	保持给定的距离
线性度量学习	标签或成对约束	线性 $P^T x$	同类近、异类远
孪生网络、三元组	标签	神经网络 $f_\theta(x)$	对比损失、三元组损失
对比学习	无（数据增强）	神经网络 $f_\theta(x)$	认出同一图像的另一视图

要点 无监督方法保留数据自身的结构；度量学习让距离反映任务所需的相似性。

08

小结

Summary

本讲要点 · 中英文术语对照 · 课后作业

本讲要点

贝叶斯分类器

- 最小化条件风险 $\Rightarrow$ 贝叶斯最优分类器；0-1 损失下即**最大后验**
- 生成式模型：估计先验 $P(c)$ 与似然 $P(x | c)$
- 极大似然估计：频率、样本均值与方差；常用损失是负对数似然，正则项是负对数先验
- 朴素贝叶斯：属性条件独立；拉普拉斯修正

降维与度量学习

- k 近邻：以邻域内的类别比例估计后验；高维中受**维数灾难**限制
- PCA：最小重构误差 $\Leftrightarrow$ 最大投影方差；解为 X 的左奇异向量
- MDS：由距离求内积矩阵，再做特征值分解；欧氏距离下与 PCA 等价
- 度量学习：学习 f_θ ，使同类近、异类远

中英文术语对照

英文全称	缩写	中文
Maximum Likelihood Estimation	MLE	极大似然估计
Maximum A Posteriori	MAP	最大后验估计
k-Nearest Neighbor	kNN	k 近邻
Principal Component Analysis	PCA	主成分分析
Singular Value Decomposition	SVD	奇异值分解

英文全称	缩写	中文
Linear Discriminant Analysis	LDA	线性判别分析
Multidimensional Scaling	MDS	多维缩放
Neighbourhood Component Analysis	NCA	近邻成分分析
Labeled Faces in the Wild	LFW	自然场景人脸数据集
A Simple Framework for Contrastive Learning of Visual Representations	SimCLR	一种视觉对比学习方法

课后作业

PCA 推导

设样本 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^d$ 已中心化, $X = (\mathbf{x}_1, \dots, \mathbf{x}_m)$, $W \in \mathbb{R}^{d \times d'}$ 满足 $W^\top W = I$ 。

1. 证明 $\|\mathbf{x}_i\|^2 = \|W^\top \mathbf{x}_i\|^2 + \|\mathbf{x}_i - WW^\top \mathbf{x}_i\|^2$, 并由此说明“最小重构误差”与“最大投影方差”等价
2. 证明 $\sum_i \|W^\top \mathbf{x}_i\|^2 = \text{tr}(W^\top XX^\top W)$
3. $d' = 1$ 时, 用拉格朗日乘子法求解 $\max_w w^\top XX^\top w$, s.t. $w^\top w = 1$, 说明最优 w 是 $XX^\top$ 最大特征值对应的特征向量
4. 证明最小重构误差之和等于 $XX^\top$ 舍去的特征值之和; 并说明这些特征值恰为 X 的奇异值的平方

参考: 周志华《机器学习》第 7、10 章; Bishop《Pattern Recognition and Machine Learning》第 1.5、2.5、12.1 节。