第 32 章 收官

32.1 特征利用技术

命题 32.1.1 (利用核方法挖掘海量特征)

- 核函数族

  通过 Mercer 核 $K(x,x')=\langle\Phi(x),\Phi(x')\rangle$ 隐式定义海量特征空间：
  - 多项式核：$K(x,x')=(1+x^\top x')^d$（有限维多项式变换）；
  - 高斯核：$K(x,x')=\exp\!\big(-\gamma\|x-x'\|^2\big)$（无限维特征）；
  - Stump 核：以决策桩为基变换的核。
- 核运算规则

  Mercer 核的线性组合（和）与乘积仍是 Mercer 核，对应特征空间的并集与组合。
- 典型算法
  - 核岭回归（Kernel Ridge Regression）
  - 核逻辑回归（Kernel Logistic Regression）
  - 支持向量机（SVM）、支持向量回归（SVR）、概率 SVM
  - 核 PCA、核 k-Means 等无监督方法。

结论　核方法以“隐式特征 + 线性模型”范式，在无需显式计算高维 $\Phi(x)$ 的情况下，高效利用海量特征并避免维度灾难。

命题 32.1.2 (通过聚合利用预测特征)

设隐藏特征由若干基预测器给出：$\Phi_t(x)=g_t(x)$。借助集成策略将简单特征提升为强大预测力。

1. 基预测器示例
   - 决策桩（Decision Stump）：单层阈值函数；
   - 决策树（Decision Tree）：递归划分 + 叶节点预测；
   - 高斯 RBF：原型中心影响函数。
2. 聚合策略
   - 均匀投票：Bagging（Bootstrap Aggregating）；
   - 非均匀加权：AdaBoost、Gradient Boosting；
   - 条件聚合：随机森林（Random Forest）；
   - 最近邻：距离加权投票；
   - 无限集成：决策树 +SVM 组合等。

结论　通过恰当的聚合机制，简单预测特征被组合成高容量、低偏差且可控方差的强预测器，构成集成学习的核心思想。

命题 32.1.3 (通过“提取”利用隐藏特征)

隐藏特征被视作待与常规权重联合优化的隐藏变量，并可借助无监督学习进行初始化或正则化。

1. 隐藏变量示例

- 神经网络：神经元权重；
- RBF 网络：径向基中心；
- 矩阵分解：用户/电影隐因子；
- 决策树：划分阈值与叶值；
- AdaBoost：弱学习器参数；
- k-Means：聚类中心；
- PCA：主方向；
- 自编码器：编码层权重。

2. 联合优化范式

$$\min_{\theta,\beta} \mathcal{L}\big(\beta, \Phi(x;\theta)\big) + \mathcal{R}(\theta),$$

其中

- θ：隐藏特征参数；
- β：线性组合权重；
- $\mathcal{R}(\theta)$：可选的无监督正则项（预训练、聚类、降维）。

3. 进阶组合

- Gradient-Boosted Neurons：逐步构造并优化神经元；
- 在分解后的特征上再训练深度网络；
- 多阶段预训练（如自编码器 $\rightarrow$ 深层网络）。

结论　通过将隐藏特征作为可学习变量并与下游任务联合优化，模型既能自动发现数据内在结构，又保留端到端的监督优势，构成现代机器学习的主流范式。

命题 32.1.4 (通过压缩利用低维特征)

将原始高维输入压缩为低维表示，再送入后续模型，兼顾表达力与计算效率。

1. 信息保持型压缩
   - 自编码器（Autoencoder）：非线性降维并重建输入；
   - PCA：线性正交投影至最大方差子空间。
2. 任务导向型压缩
   - 决策桩（Decision Stump）：沿单一最优轴投影；
   - 决策树 / 随机森林：通过递归划分实现“分段”低维嵌入；
   - 矩阵分解：将用户-电影高维稀疏向量压缩为低维隐因子。
3. 随机或选择性压缩
   - 随机投影：Johnson-Lindenstrauss 保证近似距离保持；
   - 特征选择：仅保留“最有帮助”的少数特征。

统一结论　无论信息保持、任务导向还是随机/选择压缩，低维特征既降低过拟合风险，又提升计算效率，是连接高维原始数据与高效学习器的桥梁。

例题 32.1AdaBoost-Stump 最终假设形式（PCA 预处理场景）

在经 PCA 预处理的数据集上运行 AdaBoost-Stump，从原始特征 $\boldsymbol{x}$ 的角度看，最终假设 $G(\boldsymbol{x})$ 呈现何种形式？

1) 隐藏神经元含 $\tanh(\cdot)$ 的神经网络
2) 隐藏神经元含 $\text{sign}(\cdot)$ 的神经网络
3) 决策树
4) 随机森林

解答 正确选项为 $\boxed{2}$。PCA 对原始特征 $\boldsymbol{x}$ 进行线性变换，不改变数据的线性可分性本质。在变换后的数据上使用决策树桩（decision stump），等效于在原始特征空间中使用感知机（perceptron），而感知机的激活函数通常为符号函数 $\text{sign}(\cdot)$。

AdaBoost-Stump 通过集成多个决策树桩构建强学习器，其最终假设 $G(\boldsymbol{x})$ 本质是多个感知机的线性组合，对应" 隐藏神经元含 $\text{sign}(\cdot)$ 的神经网络" 结构（感知机可视为最简单的神经网络形式）。 ■

32.2 误差优化技术

命题 32.2.1 (数值优化：梯度下降的统一视角)

当目标函数 E 对参数 θ 可微时，可构造一阶近似更新

$$\theta^{\text{new}} \leftarrow \theta^{\text{old}} - \eta \nabla_\theta E,$$

其中 η 为学习率。根据批量大小与策略差异，演化出多种具体算法。

1. 批量梯度下降（**GD**） 每次迭代使用全部样本计算梯度，收敛稳定但计算量大。
2. 随机梯度下降（**SGD**） 每次仅随机抽取单个样本，更新快、内存省，适用于大规模数据。
3. 小批量梯度下降（**Mini-batch GD**） 折中方案：每步使用 B 个样本，兼顾效率与收敛稳定性。
4. 最速下降（**Steepest Descent**） 沿线搜索最优步长 η 而非固定值，加速收敛。
5. 函数梯度下降（**Functional GD**） 在函数空间执行梯度步，AdaBoost、GradientBoost 均为此例。
6. 典型模型对应
   - 神经网络：反向传播（backprop）；
   - 矩阵分解：交替 SGD/ALS；
   - 线性 SVM：次梯度下降；
   - 核逻辑回归：核空间梯度下降。
7. 进阶扩展
   - 二阶方法：牛顿法、拟牛顿（L-BFGS）；
   - 约束优化：投影梯度、Frank-Wolfe。

统一结论 梯度下降及其变体为各类模型提供通用且高效的数值优化框架，依据数据规模、模型结构与计算资源灵活选用。 ♠

命题 32.2.2 (间接优化：通过等价解求解)

当原问题难以直接优化时，可构造与原问题等价的易解形式，从而在等价域内获得最优解。

1. 对偶 **SVM**（**Dual SVM**）
   原问题：带约束的二次规划（QP）

等价解：对偶 QP，仅需支持向量，核技巧自然嵌入。

2. 核逻辑回归（**Kernel Logistic Regression**）

   原问题：非线性高维优化

   等价解：表示定理（representer theorem）给出有限维参数化，转化为核矩阵上的凸优化。

3. 主成分分析（**PCA**）

   原问题：最大化方差的高维投影

   等价解：协方差矩阵的特征问题，直接通过特征分解获得全局最优。

4. 核岭回归（**Kernel Ridge Regression**）

   原问题：无限维正则化最小二乘

   等价解：表示定理 + 核矩阵求逆，转化为有限维线性系统。

5. 推广意义

   现代核方法与诸多 Boosting 模型均重度依赖“等价解”技术，将复杂或无限维问题转化为有限维、凸或特征问题，实现高效且稳定的求解。

统一结论　“间接优化”通过数学等价性把难解问题映射到易解空间，是机器学习算法设计与工程实现的核心技巧之一。

命题 32.2.3 (复杂优化：多步骤分解策略)

当原始问题难以一次性求解时，可将其拆解为若干更易处理的子问题，通过多阶段、交替或分治策略逐步逼近全局最优。

1. 多阶段（**Multi-Stage**）优化
   - 概率支持向量机（probabilistic SVM）：先求硬间隔 SVM，再校准概率输出；
   - 深度网络预训练（DeepNet pre-training）：逐层无监督预训练 + 端到端微调。
2. 交替优化（**Alternating Optimization**）
   - k-Means：交替更新簇划分与聚类中心；
   - 交替最小二乘（Alternating Least Squares）：矩阵分解中固定用户因子/电影因子轮流求解；
   - RBF 网络：先确定中心（k-Means），再求线性权重。
3. 分治（**Divide & Conquer**）
   - 决策树：递归划分特征空间，在每个叶节点独立拟合；
   - 线性混合（Linear Blending）：先训练若干基模型，再线性组合其输出；
   - 堆叠（Stacking）：分层模型，逐级提炼预测。

统一结论　通过将复杂优化拆解为易解子问题，并以多阶段、交替或分治方式协同求解，现代机器学习得以训练高容量模型并保持计算与统计可行性。

例题 32.2DeepNet 算法优化技术判定

在经 PCA 预处理的数据集上运行 Lecture 213 介绍的 DeepNet 算法时，使用了哪种优化技术？

1) 梯度下降变种
2) 定位等效解
3) 多阶段优化

4) 以上所有

解答 正确选项为 4。深度网络（DeepNet）训练涉及多环节优化技术：

- 训练阶段：采用小批量梯度下降（minibatch GD ），属于梯度下降变种，对应选项 1 ；
- PCA 预处理：求解特征问题（eigenproblem ）时，涉及等效解定位，对应选项 2 ；
- 预训练阶段：常通过多阶段优化（如逐层预训练）提升模型，对应选项 3 。

因涵盖选项 1、2、3 的优化逻辑，故最终为以上所有。 ■

32.3 过拟合消除技术

命题 32.3.1 (通过正则化消除过拟合)

当模型过于强大时，可在不同环节“加刹车”抑制过拟合，核心手段如下。

1. 大间隔 / 边界正则化
   - SVM、SVR：最大化几何间隔；
   - AdaBoost：通过加权投票间接实现大间隔。
2. 权重惩罚
   - L2 正则：岭回归、核模型、神经网络权重衰减；
   - L1/L0 正则：权重消除、特征选择。
3. 投票 / 平均化
   - Bagging：Bootstrap 重采样后均匀投票；
   - 随机森林：Bagging + 随机子特征；
   - 均匀混合（uniform blending）：线性组合多模型输出。
4. 去噪 / 鲁棒化
   - 去噪自编码器：对输入加噪声训练；
   - 随机失活（dropout）：神经网络内部去噪。
5. 结构约束
   - 决策树剪枝（pruning）；
   - RBF 网络限制中心数量；
   - 神经网络早停（early stopping）。

结论 正则化技术——从权重惩罚到结构约束——是抑制过拟合、保证模型泛化能力的最重要手段。

命题 32.3.2 (通过验证消除过拟合)

当模型过于强大时，需诚实且严格地监控性能，核心验证手段如下。

- 内部验证 **(Internal Validation)**
  - SVM / SVR：利用支持向量数量 (#SV) 作为复杂度指标；
  - 决策树：剪枝时采用内部交叉验证或代价复杂度剪枝；
  - 随机森林：计算袋外误差 (OOB) 估计泛化性能；
  - 模型混合 (blending)：在验证集上估计组合权重。
- 简单但必要

- 留出法、k 折交叉验证；
- 早停 (early stopping)：监控验证集损失。

结论　内部或外部验证是防止过拟合的最后防线，确保模型在未见数据上表现可靠。

例题 32.3 随机森林过拟合消除技术判定

随机森林（Random Forest）中消除过拟合的主要技术是什么？

1) 投票/平均（voting/averaging）
2) 剪枝（pruning）
3) 早停（early stopping）
4) 权重消除（weight - elimination）

解答　正确选项为 $\boxed{1}$。随机森林是多决策树集成算法，核心通过以下方式消除过拟合：

- 构建多棵决策树（引入随机性，降低单棵树方差）；
- 最终结果由多棵树“投票（分类）”或“平均（回归）”得到，通过集成效应缓解过拟合。

选项分析：

- 选项 1：投票/平均是集成学习的核心操作，有效减少过拟合，正确；
- 选项 2：剪枝是单决策树的优化，非随机森林主要技术，错误；
- 选项 3：早停是神经网络技术，与随机森林无关，错误；
- 选项 4：权重消除是正则化手段（如神经网络 L1/L2 ），不适用随机森林，错误。

■

32.4 机器学习实践

命题 32.4.1 (NTU KDDCup 2010 冠军方案)

Yu 等人在 KDDCup 2010 竞赛中夺冠的核心策略为“特征工程 + 分类器集成”。

- 模型架构

  采用线性混合（linear blending）融合
  - 逻辑回归（Logistic Regression）；
  - 随机森林（Random Forest）。
- 特征工程
  - 大量原始特征的直接编码（rawly encoded features）；
  - 人工设计的高阶特征（human-designed features）。

结论　通过丰富特征与多模型线性融合，NTU 方案在 KDDCup 2010 取得世界第一。

命题 32.4.2 (NTU KDDCup 2011 Track 1 冠军方案)

Chen 等人在 KDDCup 2011 Track 1（音乐评分预测）中夺冠，采用“多层模型 + 线性集成”的策略。

1. 基模型
   - 神经网络（NNet）；
   - 类决策树模型（DecTree-like）。

2. 矩阵分解及其扩展
   - 多种矩阵分解变体（含概率 PCA）；
   - 受限玻尔兹曼机（RBM）：视为“扩展自编码器”。
3. 其他模型
   - k 近邻（k Nearest Neighbors）；
   - 概率潜在语义分析（PLSA）：以“软聚类”作为隐藏变量的提取模型。
4. 集成方式
   - 对上述模型输出进行线性回归融合；
   - 辅助使用神经网络与梯度提升决策树（GBDT）进一步优化预测。

结论　通过丰富的基模型与线性集成，NTU 团队在 KDDCup 2011 Track 1 中取得世界第一。

命题 32.4.3 (NTU KDDCup 2012 Track 2 冠军方案)

Wu 等人在 KDDCup 2012 Track 2（广告排序）中夺冠，采用“两阶段 + 线性混合”的集成框架。

1. 第一阶段：多样基模型
   - 神经网络（NNet）；
   - 类梯度提升决策树（GBDT-like）。
2. 第二阶段：线性混合对下列模型输出进行线性融合
   - 多种线性回归变体（含线性 SVR）；
   - 多种逻辑回归变体；
   - 多种矩阵分解变体。
3. 关键策略
   - 通过谨慎的线性权重学习实现“正确混合而不过拟合”。

结论　凭借两阶段集成与防过拟合的混合策略，NTU 团队在 KDDCup 2012 Track 2 中夺得世界第一。

命题 32.4.4 (NTU KDDCup 2013 Track 1 冠军方案)

Li 等人在 KDDCup 2013 Track 1（论文-作者匹配）中夺冠，核心策略为“巨量特征工程 + 排序模型线性集成”。

1. 模型集成
   - 超大随机森林（Random Forest，极多棵树）；
   - 多种梯度提升决策树（GBDT）变体。
2. 特征工程
   - 投入巨量精力人工设计特征；
   - 关键：结合领域知识构造高区分度特征。
3. 集成方式
   - 对上述模型的输出进行线性混合（linear blending）。

结论　凭借海量领域特征与强集成策略，NTU 团队在 KDDCup 2013 Track 1 中夺得世界第一。

命题 32.4.5 (ICDM 2006 十大经典数据挖掘算法)

2006 年 ICDM 会议评选出的最具影响力的十大算法如下（按字母顺序）：

1. C4.5：经典决策树扩展。
2. k-Means：经典的聚类算法。
3. 支持向量机 SVM：大间隔分类器。
4. AdaBoost：自适应提升集成方法。
5. Apriori：频繁项集挖掘。
6. 朴素贝叶斯 Naive Bayes：基于统计的简单线性概率模型。
7. EM：用于含隐变量模型的交替优化算法。
8. k 近邻 k Nearest Neighbor：惰性学习代表。
9. PageRank：链接分析，与矩阵分解思想相通。
10. CART：分类与回归树。

个人补充

在作者看来，现代机器学习竞赛中最具竞争力的五大算法并未全部入选，包括：

- 线性回归 LinReg
- 逻辑回归 LogReg
- 随机森林 Random Forest
- 梯度提升决策树 GBDT
- 神经网络 NNet

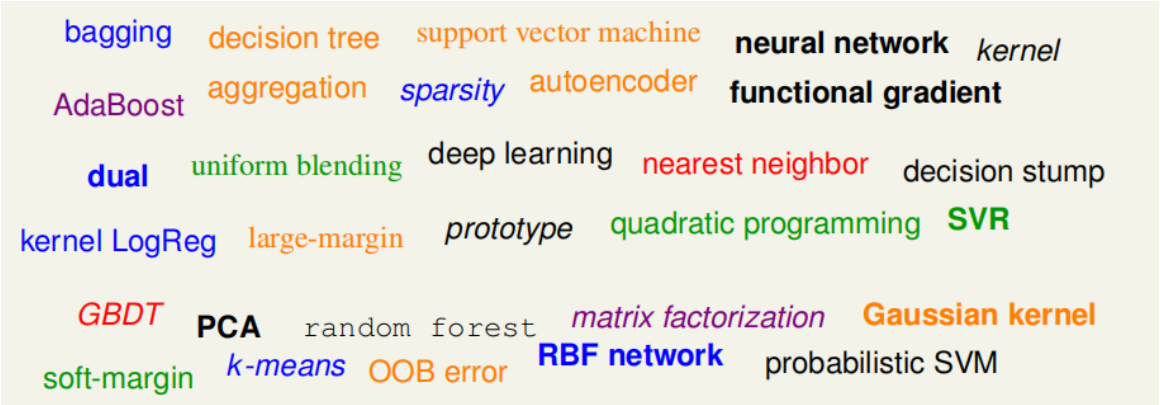

图 32.4.1: 机器学习丛林

32.5 总结

笔记 [收官]

- 特征利用技术：核方法、集成、特征提取、降维。
- 误差优化技术：梯度下降、等价转换、分阶段训练。
- 过拟合消除技术：大量正则化与验证。
- 机器学习实践：欢迎来到“丛林”!