人工智能与机器学习基础 2025-HW3

TA: 王珏

November 2025
Deadline : 2025/12/31 23:59

1 支持向量机（Support Vector Machine）

支持向量机（SVM）通过最大化分类间隔（Margin）来获得泛化能力强的决策边界。在高维空间中，通过核函数（Kernel Function）可实现非线性分类。

(a)(15 分) 硬间隔 SVM 给定线性可分数据集，SVM 的目标为：

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t. \quad y_i(w^\top x_i + b) \geq 1$$

请完整复现上课推导步骤，说明如何得到以上目标。

(b)(10 分) 软间隔 SVM 硬间隔的定义要求要把正负类点完全分开，这个要求在数据集含有噪音点时可能过于严格，为了解决该问题，我们引入松弛因子 ζ_i，有

$$y_i(w \cdot x_i + b) \geq 1 - \zeta_i$$

相应的，新的目标函数变为：

$$\min_{w,b,\zeta} \quad \frac{1}{2}\|w\|^2 \; + \; C\sum_{i=1}^{N} \zeta_i^2$$

$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 - \zeta_i, \quad \forall i = 1, 2, \ldots, N,$$

$$\zeta_i \geq 0, \quad \forall i = 1, 2, \ldots, N.$$

试写出其对偶形式。

(c)(15 分) 支持向量机求解 考虑如下 3 个二维样本，其中两正一负：

$$X_1 = (3, 3), \quad y_1 = +1; \qquad X_2 = (4, 3), \quad y_2 = +1; \qquad X_3 = (1, 1), \quad y_3 = -1.$$

我们希望通过硬间隔 SVM 求得最大间隔分离超平面。其对偶优化问题为：

$$\max_{\alpha} \sum_{i=1}^{3} \alpha_i - \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \alpha_i \alpha_j y_i y_j (x_i^\top x_j),$$

满足约束：

$$\alpha_i \geq 0, \qquad i = 1, 2, 3,$$

$$\alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 = 0.$$

(1) 写出该对偶问题的显式形式（不消元）。

(2) 利用约束消去变量，写出化简后的二元对偶目标函数。

(3) 对该对偶问题进行求解，给出最优对偶变量 $(\alpha_1, \alpha_2, \alpha_3)$.

(4) 根据最优对偶变量，利用

$$w = \sum_i \alpha_i y_i x_i$$

求出原始空间中的法向量 w。

(5) 选取任意一个支持向量，求出偏置 b。

(6) 写出最终分类超平面方程，形如

$$w^\top x + b = 0.$$

(c)(5 分) 思考

(1) 由此题思考**解释支持向量在模型中的作用**：为什么只有支持向量会影响最优超平面？（尽量用自己的语言表述，提示：拉格朗日对偶条件的互补松弛性）

2 决策树（Decision Tree）

决策树是一种可解释性强的监督学习算法，它通过递归划分特征空间，将复杂的决策问题分解为一系列“如果–那么”的规则。核心问题是如何选择最优划分特征，常见准则包括信息增益（Information Gain）和基尼指数（Gini Index）。

(a)(10 分) 决策树分裂特征选择与分类错误率计算 我们希望利用决策树来预测贷款申请人是否违约（1 表示违约，0 表示未违约）。给定如下训练样本（与课堂案例一致）：

编号	收入水平	是否有房	是否违约
1	高	是	0
2	高	否	0
3	中	是	0
4	低	否	1
5	低	否	1

(1) 课堂上我们使用“分类错误率（classification error）”来评估一次分裂是否有效。请分别以收入水平和是否有房为分裂特征，在根节点上构建一个决策树桩（decision stump），并计算每种分裂方式的分类错误率。请写出你的计算过程，并根据错误率选择根节点的最优分裂特征。

(2) 课堂中我们提到：由于分类错误率可能对部分“看似有效但局部不稳定”的分裂不敏感，因此需要额外的停止条件或剪枝（pruning）策略以避免过拟合。请结合课堂内容简述：

- 为什么仅用分类错误率作为分裂标准可能导致过拟合?
- 提前停止（early stopping）与剪枝（pruning）分别如何缓解这一问题?

(3) 假设将“收入水平”替换为一个连续特征“年收入（万元）”。根据课堂讲述的“阈值分裂（threshold split）”，回答：

- 如何确定候选阈值?
- 为什么只需要考虑相邻样本之间的中点?

- 这个年收入特征能否反复使用?

(4) 为了更好地区分不同分裂的“纯度改善”，ID3 使用信息熵（entropy）与信息增益（information gain）：

$$\mathrm{Gain}(D, A) = H(D) - H(D|A)$$

请计算上述两个特征（收入水平、是否有房）的信息增益，并比较它们与分类错误率的选择结果是否一致。

(5) 信息增益容易偏好“取值数较多”的特征（over-favor multi-valued attributes）。C4.5 通过引入“增益率（gain ratio）”解决此问题：

$$\mathrm{GainRatio}(A) = \frac{\mathrm{Gain}(D, A)}{\mathrm{IV}(A)}$$

其中 $\mathrm{IV}(A)$ 是“固有值”（intrinsic value）。请解释：

- 为什么信息增益会偏好取值较多的特征?
- 增益率是如何缓解这一问题的?

由于课堂上未提及，此处补充两个算法的介绍供大家参考：C4.5 算法介绍，ID3 算法介绍

3 集成学习（Ensemble Learning）

集成学习通过组合多个弱学习器（如决策树）以提升性能。常见方法包括 Bagging、随机森林（Random Forest）与 Boosting 系列（如 AdaBoost、XGBoost）。

(a)(15 分) Bagging 与 Boosting 的比较与推导

(1) 试解释 Bagging 与 Boosting 在“样本采样方式”和“模型训练目标”上的根本区别。

(2) 假设我们使用 AdaBoost 算法，在第 t 轮弱分类器的加权错误率为 ϵ_t，请证明其权重系数：

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}$$

并计算该权重系数的范围，解释该公式反映的直觉含义。

(b)(10 分) 随机森林与偏差-方差权衡 假设你训练了一个包含 100 棵树的随机森林，每棵树的误差率为 0.2，且不同树之间错误独立。

(1) 请计算整体模型错误率的期望值（可用二项分布近似）。

(2) 若树与树之间相关性增大，会对整体性能产生何种影响？请结合“偏差–方差分解”进行分析。

(c)(10 分) 问答题 判断以下说法是否正确并给出理由：

(1) 有效的集成学习需要集合中的模型具有单一性，最好将同一类型的预测模型结合起来。

(2) 训练集成模型时，单个模型的参数不会随之更新。

(3) AdaBoost 算法中，需要按照之前学习器的结果对训练数据进行加权采样。

(4) 尝试解释下课堂上所讲的：**Random forest does not perform as well in general as boosting.**（随机森林在一般情况下性能不如 boosting）。

注意：如果完全使用人工智能来完成这次基础作业，你将会得到一个不太高的分数。我们并不需要一份完全正确的解答，而是希望你可以真正掌握这几类经典机器学习算法，希望以上习题一定能给你一些小的帮助！谢谢大家！如果有任何建议欢迎私聊我！——TA：王珏