量化交易工作坊 · 任务五：AI 交易引擎——机器学习算法与场景应用

基础分类型机器学习算法：逻辑回归 / 决策树 / 随机森林 | 评价指标：混淆矩阵 · AUC · ROC

摘要

本任务聚焦量化交易中日益重要的机器学习方法，以最基础的分类型算法为入口，系统讲解逻辑回归、决策树、随机森林的思想与特点，以及混淆矩阵、AUC、ROC 等模型评价指标的含义；并以 Python 完成端到端实践：分别加载 scikit-learn 自带的乳腺癌二分类数据集与由股票日线工程化得到的「次日涨跌方向」数据集，划分训练/测试集，训练上述三类模型，计算 AUC 并绘制 ROC 曲线与混淆矩阵。实验显示：在结构化医学数据上三类模型 AUC 均超过 0.94（逻辑回归与随机森林高达 0.99），判别力极强；而在股票次日涨跌任务上 AUC 全部接近或低于 0.5，揭示「用简单量价特征预测涨跌」几乎不具备可学习信息——这一对比本身正是机器学习在量化领域最值得记取的经验：数据中的信息量，决定了模型能力的上限。

一、分类型机器学习算法

分类是监督学习的核心任务：给定带标签样本 (x, y)（y 取自有限离散类别，如 0/1），训练模型 f 使其能对未见样本预测类别。在量化中，涨跌判断、违约识别、风格聚类等均可建模为分类问题。下面介绍三种最常被用作基线的分类算法。

（1）逻辑回归（Logistic Regression）：虽名含「回归」，实为线性分类器。它对特征做线性组合 $z=w_0+w_1x_1+\cdots$ 后，经 Sigmoid 函数 $\sigma(z)=1/(1+e^{-z})$ 压缩到 (0,1)，解释为「属于正类的概率」，以 0.5 为阈值判类。其优点是可解释性强（系数直接反映特征方向）、训练快、是极好的基线；局限是只能刻画线性关系。

（2）决策树（Decision Tree）：通过递归的 if-then 规则切分特征空间，每个叶子给出类别。构建时贪心选择最能「提纯」类别的特征与切分点（基尼不纯度或信息增益）。它天然处理非线性、无需特征缩放、结果直观；但单棵树易过拟合、对数据扰动敏感。

（3）随机森林（Random Forest）：集成学习代表。它训练大量（如 100 棵）决策树，每棵仅在随机样本子集与随机特征子集上训练（双重随机），最后多数投票。这显著降低单棵树的方差与过拟合，通常比单棵树更稳健，还能输出特征重要性。三者关系：逻辑回归是线性基线，决策树引入非线性但易过拟合，随机森林用集成放大优点、抑制缺点——在表格数据上常是最先尝试的强力模型。

二、机器学习模型评价指标

仅看准确率远远不够，尤其类别不平衡时。以下三个指标更全面：

（1）混淆矩阵（Confusion Matrix）：以二分类为例，交叉统计预测与真实标签得 2×2 矩阵——TP（真正例）、FN（假负例/漏报）、FP（假正例/误报）、TN（真负例）。由其派生：准确率=(TP+TN)/总数；精确率 Precision=TP/(TP+FP)（预测为阳性者中真实阳性比例，衡量误报）；召回率 Recall=TP/(TP+FN)（真实阳性中被找出比例，衡量漏报）；F1 为二者调和平均。

（2）ROC 曲线与 AUC：ROC 以假正率 FPR=FP/(FP+TN) 为横轴、真正率 TPR=TP/(TP+FN) 为纵轴；改变判定阈值得到一系列 (FPR,TPR) 点连成曲线。曲线越向左上凸出，判别力越强；左下的对角线为随机猜测基准（AUC=0.5）。AUC 为曲线下面积，$\in[0,1]$：0.5 表示无预测力，0.7~0.8 可接受，>0.9 很强。AUC 对类别不平衡不敏感、综合所有阈值表现，是排序/概率型分类器（金融风控、涨跌预测）的首选度量。

三、Python 实现与结果

以下代码片段展示从数据加载到评估的核心流程（完整代码见 Notebook）。关键点：用 train_test_split 按比例划分训练/测试集并以 stratify 保持类别比例；逻辑回归前对特征做标准化（StandardScaler，仅在训练集拟合并变换测试集，杜绝未来函数）；三类模型分别训练后在测试集上计算 AUC、混淆矩阵等指标。

```
from ml_lib import (load_dataset, prepare, train_one, evaluate_model, MODEL_REGISTRY)
models = ('logistic', 'tree', 'forest')
X, y, meta = load_dataset(kind='cancer')        # 乳腺癌：569样本/30特征/标签0-1
sp = prepare(X, y, test_size=0.3, random_state=42, scale=True)  # 训练/测试划分+标准化
for mk in models:
    model = train_one(mk, sp['X_train'], sp['y_train'])        # 训练
    ev = evaluate_model(model, sp['X_test'], sp['y_test'])      # 评估
    print(mk, 'AUC=%.3f' % ev['auc'], '准确率=%.3f' % ev['accuracy'])
```

3.1 乳腺癌诊断数据集（结构化医学特征）

数据集共 569 个样本、30 个特征，正类（良性）357 例、负类（恶性）212 例；按 7:3 划分后测试集 171 例。各模型测试集表现如下：

模型	AUC	准确率	精确率	召回率	F1	TP	FP	FN	TN
逻辑回归	0.998	0.988	0.991	0.991	0.991	106	1	1	63
决策树	0.944	0.930	0.944	0.944	0.944	101	6	6	58
随机森林	0.991	0.936	0.944	0.953	0.949	102	6	5	58

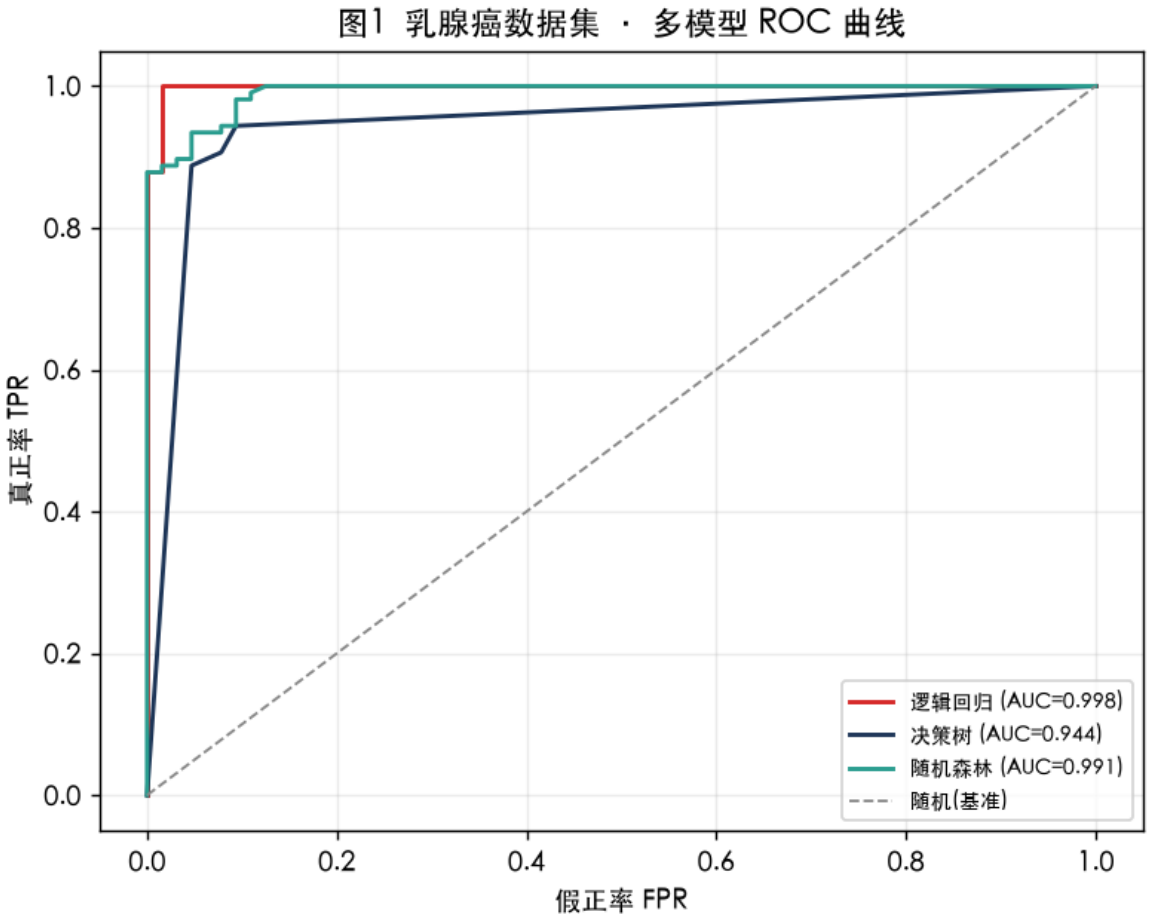

图1 乳腺癌数据集 · 多模型 ROC 曲线（逻辑回归/决策树/随机森林）

三条 ROC 曲线均显著位于随机基准（对角线）左上方，说明三类模型都能有效区分良/恶性肿瘤。其中逻辑回归 AUC=0.998、随机森林 AUC=0.991，均接近 0.99；决策树略低（=0.944，单棵树略有过拟合）。结构化医学数据中特征与标签存在强可学习结构，故线性与树模型皆表现优异。

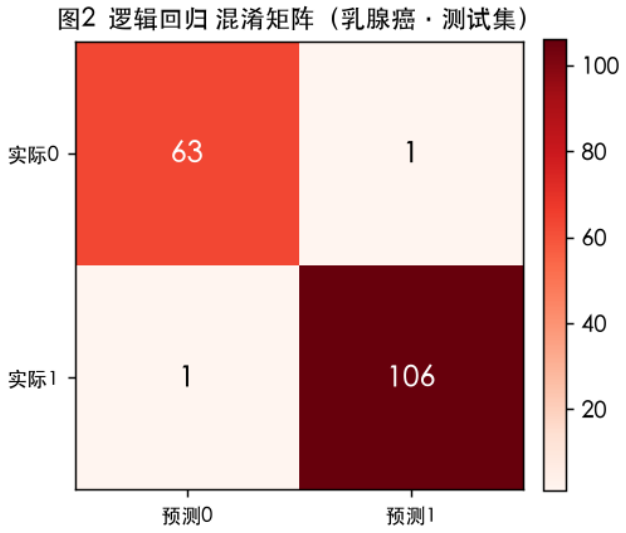

图2 逻辑回归 混淆矩阵（乳腺癌 · 测试集，171例）

以表现最好的 逻辑回归 为例，测试集 171 例中真正例 TP=106、真负例 TN=63，仅假负例 FN=1、假正例 FP=1。在医疗诊断中
FN（漏诊恶性肿瘤）代价极高，实践常调低判定阈值以抬高召回率、压低 FN——这正是 ROC 曲线协助「按代价选阈值」的价值。

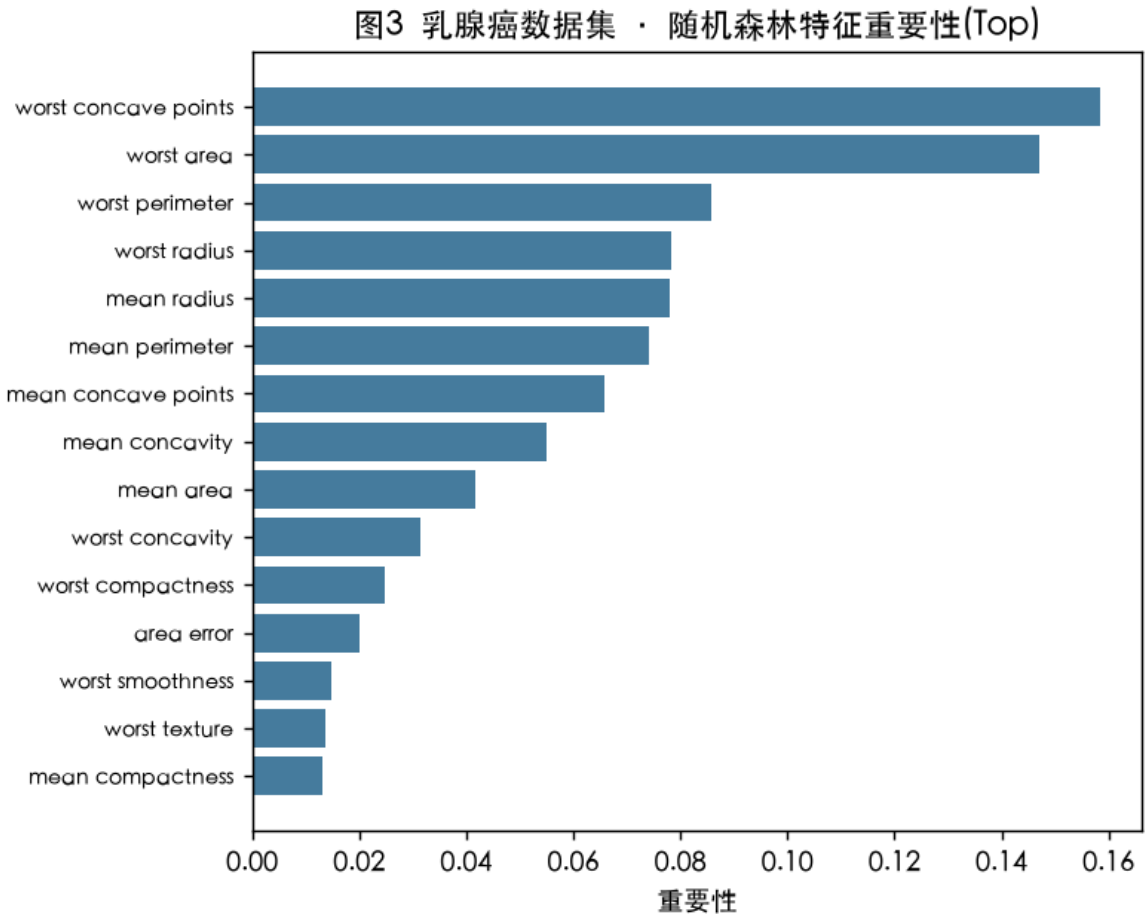

图3 乳腺癌数据集 · 随机森林特征重要性（Top 特征）

随机森林给出的特征重要性清晰指向 worst concave points、worst area、worst perimeter 等「worst 类」几何特征——这些描述细胞核极端形态的指标，是判别良/恶性肿瘤最关键的线索。特征重要性

既能解释模型，也能反过来指导特征工程与降维。

3.2 股票次日涨跌数据集（量价特征）

以芯动联科（688582.SH）日线工程化：特征为 9 个量价技术指标（各周期收益率、滚动波动率、量比、均线价差），标签为「次日收益率是否为正」（1=涨，0=跌）。区间 20230630 ~ 20260710，共 714 个样本，其中上涨 347 例、下跌 367 例。各模型表现如下：

模型	AUC	准确率	精确率	召回率	F1
逻辑回归	0.445	0.470	0.439	0.346	0.387
决策树	0.484	0.470	0.464	0.615	0.529
随机森林	0.389	0.414	0.402	0.433	0.417

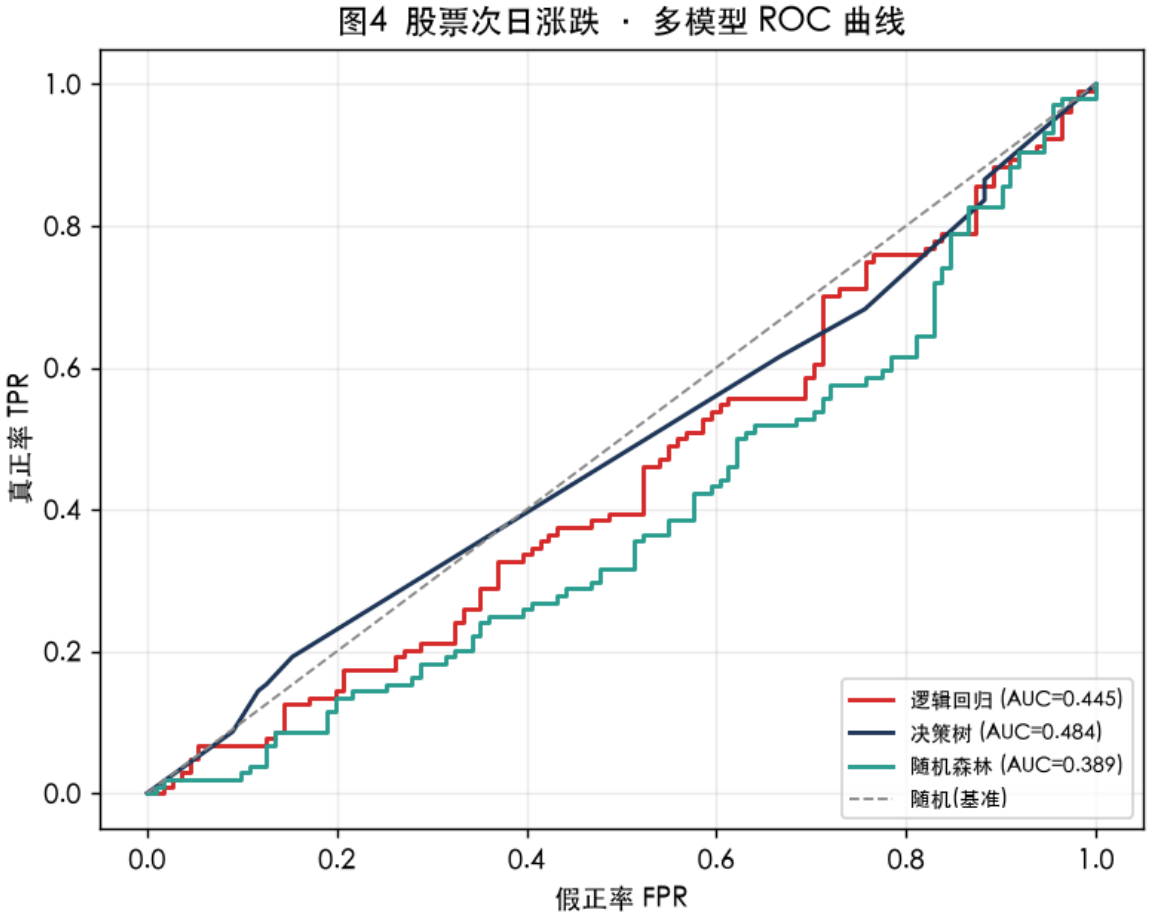

图4 股票次日涨跌 · 多模型 ROC 曲线

在股票次日涨跌任务上，三类模型 AUC 均接近甚至低于 0.5（随机森林仅约 0.389），ROC 曲线几乎贴着对角线甚至下凹。这揭示重要事实：用简单量价特征预测「明天涨跌」几乎不具备可学习预测力。金融时间序列噪声大、有效信息被快速套利，使「方向预测」成为业界公认的难题——这也正是机器学习在量化中更常见的用法：不是直接押注涨跌，而是做概率化风控、截面排序、特征降维、异常检测等辅助决策。

四、总结与心得

1）算法选型：逻辑回归是线性基线、可解释；决策树引入非线性但易过拟合；随机森林以集成换取稳健与特征重要性，是表格数据上的强力默认选择。2）评价指标：准确率在不平衡场景会失真，必须结合混淆矩阵、精确率/召回率、ROC 与 AUC 综合判断；AUC

因对阈值与不平衡不敏感，是概率型分类器首选。3）数据决定上限：同样算法在乳腺癌 AUC≈0.99、在股票涨跌 AUC≈0.4——模型能力受限于特征蕴含的信息量，提示「特征工程 > 调模型」，也点明机器学习在金融的现实边界。4）量化应用方向：与其执着预测涨跌，不如把分类/排序模型用于风险评分、违约预警、风格聚类、因子有效性筛选等更稳健场景，并始终以 AUC、回测与样本外检验严格把关，避免过拟合与幸存者偏差。

声明：本报告所有模型在本地实时训练与评估（scikit-learn），仅用于量化学习与机器学习原理演示，不构成任何投资建议。