

NIPT 第四问：基于可解释机器学习的个体化检测与复检策略

项目：NIPT 的时点选择与胎儿的异常判定

September 5, 2025

1 目标

在男胎样本上，利用可解释的概率模型给出首检孕周 t_1 与（可选）复检孕周 t_2 的最优策略，使

$$J = \underbrace{\text{risk}(T_{\text{res}})}_{\text{孕周风险}} + \lambda(1 - P_{\text{succ}}) + c_1 + \mathbf{1}\{\text{复检}\} c_r + \kappa \mathbb{E}[T_{\text{res}} - t_{\min}]$$

最小。这里 $\text{risk}(t) \in \{1, 2, 3\}$ 分别对应 $t \leq 12, 13 \leq t \leq 27, t \geq 28$ 周。

2 命中率模型 $P_{\text{hit}}(t \mid x)$

用可解释模型（GAM/EBM/LightGBM(单调约束)）预测在孕周 t 达到阈值（如 $Y \geq 4\%$ ）的概率。设特征 $x = (\text{GA}, \text{BMI}, \text{Age}, \text{Height}, \text{Weight})$ ，标签 $y = \mathbf{1}\{Y \geq 0.04\}$ 。

单调约束 为符合生理先验，要求 P_{hit} 对 GA 单调不减：

$$\frac{\partial}{\partial \text{GA}} P_{\text{hit}}(t \mid x) \geq 0.$$

实际实现中在 LightGBM 里对 GA 特征施加正单调约束；GAM/EBM 则靠样条/分段函数与等概率校准实现近似单调。

概率校准 先得原始分数 $s(x)$ ，再用单调的校准函数 $g(\cdot)$ （如保序回归）得到

$$\hat{P}_{\text{hit}}(x) = g(s(x)),$$

以提升概率可解释性。

3 策略优化

3.1 单次检测

在网格 $t \in [t_{\min}, t_{\max}]$ 上求

$$J_1(t) = \text{risk}(t) + \lambda(1 - \hat{P}_{\text{hit}}(t \mid \tilde{x})) + c_1 + \kappa(t - t_{\min}),$$

取最小点 t_1^* 。 \tilde{x} 为分组情景（如 BMI 组中位数）。

3.2 一次复检

设 $t_2 = t_1 + \Delta$ ， $\Delta \in \mathcal{D}$ （如 $\{1, 1.5, 2, 3\}$ 周）。一次成功概率 $P_1 = \hat{P}(t_1)$ ；二次命中概率 $P'_2 = \alpha \hat{P}(t_2) + (1 - \alpha) \hat{P}(t_1)$ （ $\alpha \in [0, 1]$ 控制独立性）。成功概率 $P_{\text{succ}} = P_1 + (1 - P_1)P'_2$ ，期望结果孕周 $\mathbb{E}[T_{\text{res}}] = P_1 t_1 + (1 - P_1)t_2$ 。目标

$$J_2(t_1, \Delta) = c_1 + (1 - P_1)c_r + \mathbb{E}[\text{risk}(T_{\text{res}})] + \lambda(1 - P_{\text{succ}}) + \kappa(\mathbb{E}[T_{\text{res}}] - t_{\min}).$$

在网格 (t_1, Δ) 上暴力搜索最优值。

4 解释性

输出特征贡献曲线（GAM/EBM 形状函数）、LightGBM 的单调 PDP、或 SHAP（如可用）。并提供策略表（各 BMI 组的 $t_1^{t_2}$ 、成功率与置信带）。

5 实现

附带的 `nipt_q4_ml_policy.py` 脚本实现上述流程，默认：

- 模型优先 LightGBM(单调) \rightarrow EBM \rightarrow GAM(样条 GLM)；
- 校准优先 Isotonic \rightarrow Platt \rightarrow 无；
- 输出策略表与可视化（热力图、曲线）。