金融机器学习算法

第七讲

金融机器学习的交叉验证

本讲主要内容

- 学习本章的动机
- 标准 CV 法复习
- 标准 CV 法失效原因
- 净化-隔离 CV 法

学习本章的动机

- 交叉验证 (Cross-Validation: CV) 的目的是确定模型的泛化误差
- CV 法能够评估和控制过拟合 (Overfitting)
- 标准 CV 法不适用于金融机器学习，为什么?
- 通过什么方法能够弥补传统 CV 法的缺陷?

标准 CV 法复习

- CV 法的基本思想：将数据分割成训练集 (train set) 和测试集 (test set)，训练集和测试集不相交。使用训练集来训练模型，使用测试集来评估模型的性能。通过优化模型超参数来选择最优模型
- CV 的关键假设：预测数据满足 $i.i.d$，因此测试集与训练集仅仅的差异进在于噪声大小，具有相同的结构规律
- 在 CV 中调超参，导致进行多次 CV，但引入了多重测试偏误。很难反应泛化误差。因此标准 CV 采用重抽样 CV 法：[训练集 + 验证集]$\times$ 多次 + 测试集 ([train+validation]$\times$M+test)

标准 CV 法复习

K 折交叉验证 (K-fold CV)：建模样本分为 K 个尺度相当的子集，每次选出一个子集做验证集，其余作为测试集，然后在选另个子集做验证集，其余做测试集，....，重复 K 次

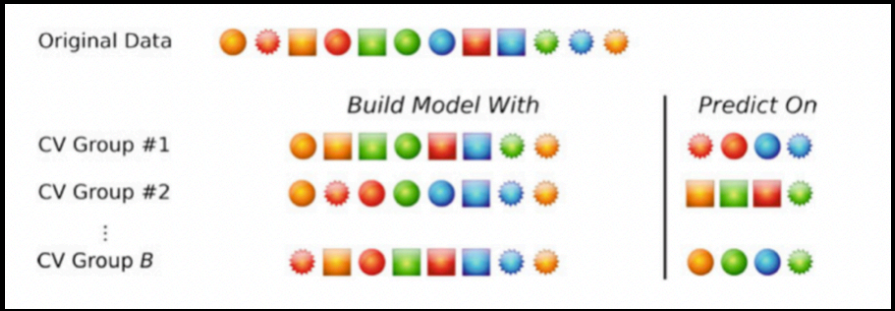
标准 CV 法复习

K 折 CV 的变体

- 留一 K 折交叉验证 (Leave only one K-fold CV: LOOCV) : K 折交叉验证的特例，每次放一个样本作为验证集)。
- 【思考】LOOCV 一共进行多少次 CV?
- 重复 K 折交叉验证 (repeated K-fold CV): 重复的进行 N 次 K 折交叉验证。每次将建模样本都分为 K 个尺度相当的子集。
- 【思考】重复 K 折交叉验证一共进行多少次 CV?

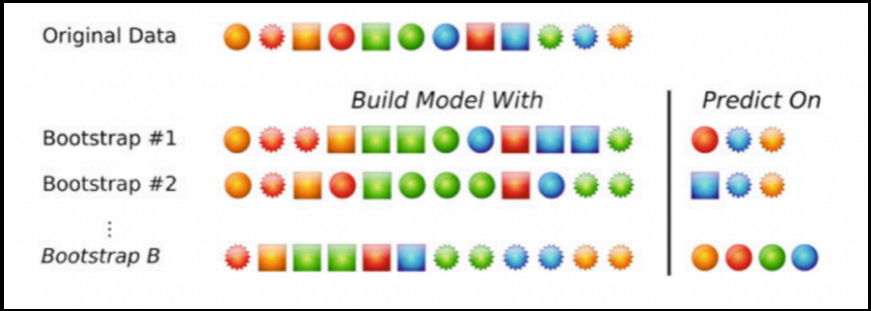
标准 CV 法复习

蒙特卡洛交叉验证 (Monte-Calo CV)：重复的对建模样本进行 B 次训练集/验证集劈分，劈分时采用固定的比例划分建模样本



标准 CV 法复习

自助交叉验证 (Bootstrap CV: BCV)：有放回抽取 B 个与建模数据等长的样本作为训练集，未抽到的样本进入验证集（袋外）。BCV 衡量的平均模型方差要低于 K 折 CV，但平均有 63.2% 的样本会出现一次，模型偏差与 2 折交叉验证相似。



标准 CV 法失效的原因

- 两个原因：（1）重抽样 CV 法仍然不能完全缓解多重测试偏误；（2）因为样本重叠，预测数据不再是 $i.i.d$
- 重抽样 CV 面对非 $i.i.d$ 的后果：
  - 信息泄露 (leakage)：训练集的信息泄露到了验证集中。
  - 假设 t 在训练集中，$t+1$ 在验证集中，t 和 $t+1$ 对应的因变量标签来此重叠数据，因此 $Y_{t+1} \approx Y_t$。
  - 由于序列相关性 $X_{t+1} \approx X_t$
  - 此时，及时 X 是预测能力很弱的变量，仍然有 $\mathbb{E}(Y_{t+1} \mid X_{t+1}) \approx \mathbb{E}(Y_t \mid X_t)$，CV 的评价被高估

弥补传统 CV 法的缺陷的办法

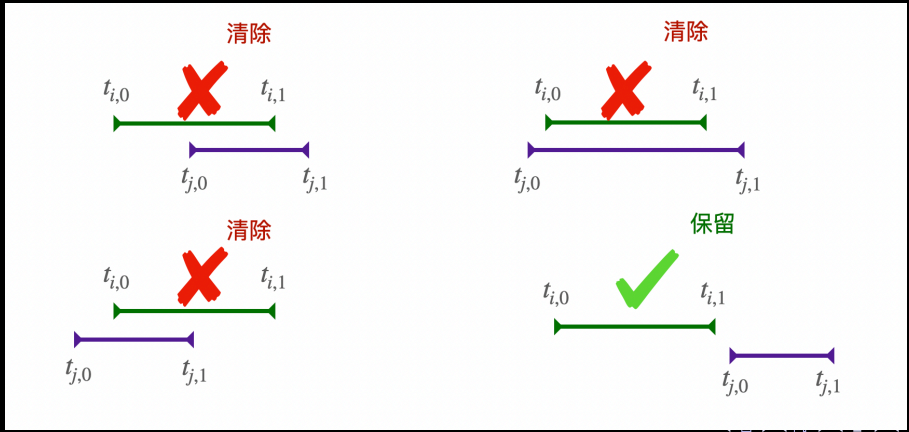
- 净化法（数据层面）：Y_j 属于验证集，Y_i 属于训练集，且 Y_i 与 Y_i 有时间重叠，则在训练时清除 Y_i
- 避免容易过拟合的分类器（算法层面）：从源头上直接避免模型过拟合，比如采用装袋式集成
  - 注意设置 Early stop 机制
  - 注意避免训练时的样本重叠，采用 DSB
  - 注意采用 AvgU×N 生成单个分类器，保证分类器的多样性
- 【思考】：对于特征而言，X_i 和 X_j 有时间重叠，是否会造成信息泄露?

净化-隔离 CV 法

- 净化 CV 法 (Purged-CV)：将会与验证集内的数据产生重叠的训练集内的那部分数据清除（仅清除训练集中的数据）
- 验证集中的标签 Y_j，$Y_j = f([t_{j,0}, t_{j,1}])$
- 训练集中的标签 Y_i，$Y_i = f([t_{i,0}, t_{i,1}])$
- 如果 $[t_{i,0}, t_{i,1}]$ 与 $[t_{j,0}, t_{j,1}]$ 产生重叠，则在训练时清除 $[t_{i,0}, t_{i,1}]$

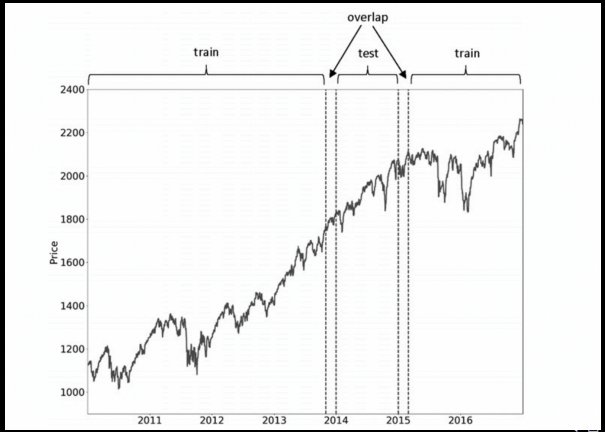
净化-隔离 CV 法

单次 Purged-CV 时应该清除哪些样本?



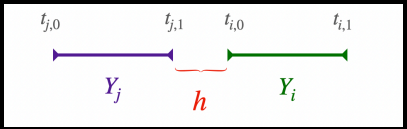
净化-隔离 CV 法

单次 Purged-CV 时训练集与验证集全貌



净化-隔离 CV 法

- 隔离机制 (Embargo)：在劈分数据集时，因为时间序列具有序列相关的性质，应该将验证集切分处之后一段时间内的数据也从训练集中删除。因为 h 不充分大，仍有可能 $Y_j \approx Y_i$



- h 可取 $0.01T$
- 附加了隔离机制的净化 CV 法称为净化-隔离 CV 法

净化-隔离 CV 法

单次 Purged-embargo CV 时训练集与验证集全貌