题目 2. 金融时间序列的风险预测

问题背景

有时依据商业秘密保护的需要，金融数据分析师得到的指标是匿名的，这使得分析师无法依据金融知识划定金融风险阈值。在课程中，我们学习了 VaR (在险价值) 评价是否发生金融风险。同时，将无风险时的样本视为负样本，将风险发生时的样本视为正样本，这就变成了一个极度不平衡的分类问题。不平衡分类模型是否比 VaR 有优势？请对此展开研究。

数据集和资料

数据集：训练集 `q_2_train.csv` 测试集 `q_2_test.csv`

从 `https://github.com/NetManAIOps/KPI-Anomaly-Detection/tree/master/Preliminary_dataset` 获得。

在数据集中，正样本的比例约为 3%，即金融风险发生的概率。每个 KPI ID 代表一条匿名金融时间序列，timestamp 代表时间戳（以秒为单位），采样间距是 1 分钟，value 代表时间序列的取值，label=0 代表正常，label=1 代表异常。数据集包括训练集和测试集，在 VaR 模型中，时间序列的均值和标准差也必须只通过训练集计算，否则会出现优化偏差 (Optimization bias).

VaR 方法假设时间序列是正态分布的，在整个时间序列上求得均值 μ，标准差 σ，估计取值在正态分布 $N(\mu, \sigma^2)$ 上的分位数。当分位数超过双侧置信水平时，则认为 “超限” 也就是异常。有些分类模型的评价指标需要模型输出 “样本属于某类的概率”，在 `scikit-learn` 中通常用 `predict_proba` 函数计算。在 VaR 方法中的计算方法是，当双侧置信水平 $0 < \alpha < 50\%$ 时，有越来越多的正常样本变成异常，记录变成异常时对应的 $1 - \alpha$，作为 “样本属于某类的概率”。

移动平均法改进的 VaR 方法认为时间序列的标准差是关于时间 t 的函数，所以不是用全局的标准差 σ，而是采用一小段时间长度 m 计算得到的 $N(\mu_m(t), \sigma_m(t)^2)$ 作为 VaR 的估计。

有时，时间序列中某一时刻的值相对于上一时刻的变化，不符合正态分布。此时应使用核密度估计 (Kernel Density Estimation), 使用不同核函数的。参考 `https://scikit-learn.org/stable/modules/density.html`

拓展阅读：时间序列发生异常时，通常连续的一小段数据都为异常，这和自然语言处理领域的命名实体识别 (Named-Entity Recognition) 类似。把时间序列看作一段文本，异常值对应实体，LSTM-CRT 神经网络模型可以预测异常值发生的时机。该模型可以从如下地址下载：

- TensorFlow 版本 `https://github.com/guillaumegenthial/sequence_tagging`
- PyTorch 版本 `https://github.com/jidasheng/bi-lstm-crf`

拓展阅读的模型不是必须应用到本题中的，也不要在 Notebook 中运行或提交，但如果你成功应用了这个模型并取得了更好的效果，请在研究报告中阐述之。使用了拓展阅读中的模型可获得加分，但研究报告的总分不超过上限。

研究任务

问题 2-1. 使用 PACF, ACF 或同类时间序列的显著性检验方法确定移动窗口的长度，确定 ARIMA 模型的 p, q 参数（差分固定为 1 阶），实现时间序列的预测。根据真实值在回归预测值分布

上的分位数，作为 “样本属于某类的概率”，评价是否发生金融风险，绘制 ROC 曲线并计算 AUC 得分。使用 VaR 模型，比较与 ARIMA 在测试集上的 AUC 分数。

问题 2-2. 将 ARIMA 模型推广为 Logistic 回归（广义线性回归），比较与 ARIMA 在测试集上的 AUC 分数。

问题 2-3. 对 ARIMA-Logistic 回归模型，任选 `skopt.BayesSearchCV`，`sklearn.model_selection.GridSearchCV`，`sklearn.model_selection.RandomizedSearchCV` 中的一种方法，搜索 ARIMA 模型的项数 p, q 的最优解。对 VaR 模型，搜索核函数和移动窗口长度 m 的最优解。