题目 5. 上证 50 股票指数的构成

问题背景

上证 50 指数是通过上证 50 股票构建的。金融专家通过分析这些股票在经济体中的重要性，赋予权重 $\mathbf{w}$, 指数的风险低于单个股票。稳健型投资者可以依据股票指数的成分购买股票，在经济发展的阶段，股票指数应保持总体上升的趋势。按照指数购买股票，风险是最低的吗？从投资组合风险最小化的角度，成分股票的权重仍然是最优的吗？请对此展开研究。

数据集和资料

上证 50 股票 2016 ~ 2019 年的行情数据 `q_5_qttn.csv`

上证 50 指数 2016 ~ 2019 年的行情数据 `q_5_idx.csv`

在任何预测与计算之前，请将股票价格转换成对数收益率，直接对于价格的预测和模型评价是无效的。

关于 CAPM 等自定义的预测模型, 可以使用 `sklearn.base.RegressorMixin` 改写成标准的 scikit-learn 机器学习模型，从而能够套用超参数优化等工具。具体方法请参考 `https://github.com/Kensuke-Hinata/statistic/tree/master/AI/books` 中的《特征工程入门与实践》的有关小节。

求风险的有效前沿涉及到求解函数的最小值，可参考 `https://www.tensorflow.org/guide/autodiff#gradient_tapes` 使用梯度下降法自动求解最小值，也可使用 Lagrange 乘子法。

研究任务

问题 5-1. 根据 2016, 2017, 2018 年的数据，分别用 CAPM 模型估计 50 个股票的 β 值，无风险收益按 1.5% 计算。用所求的 β 估计 2017, 2018, 2019 年的收益率，与真实值比较并报告误差，阐述 CAPM 模型是否有效。将 CAPM 模型改写成 `scikit-learn` 回归模型，任选 `skopt.BayesSearchCV`, `sklearn.model_selection.GridSearchCV`, `sklearn.model_selection.RandomizedSearchCV` 中的一种方法，搜索以下超参数：无风险收益率（注意：年化收益率），计算 β 的时长，预测时长（原问题中的预测时长是 1 年）。计算最优超参数和最优结果。

问题 5-2. 计算所有股票的协方差矩阵；对于给定的权重 $\forall \mathbf{w}, w \in [0,1], \sum w = 1$, 求解期望收益率 $\mathbf{E}$ 和期望金融风险（用标准差表示）σ, 求解使 σ 尽可能取得最小值的权重 $\mathbf{w}$ 和对应的期望收益率。与上证 50 成分股票权重比较，阐述风险和收益水平的区别。

问题 5-3. 根据每日的上证 50 成分股收益率，训练随机森林算法预测**次日**的股票指数收益率，不需要使用交叉验证，但要分为 80% 的训练集，10% 的验证集，20% 的测试集。优化超参数并报告测试集上的准确率，阐述是否能通过成分股票的收益率预测股票指数的收益率。如不能，尝试使用过去 $m, m > 1$ 实现预测；如 $\exists\, m$ 能够实现预测，找出每个股票的特征重要性，比较与成分股权重的差异。