核心课③ ·2020.11.07.

作业回顾：XGBoost 适合海量大、特征工程比较完善的情况下，且敏感性低。

LightGBM ←

CatBoost：如果类别特别多，适合。

☆ 模型融合：Stacking？ bagging. boosting.

Voting / Averaging（平均）

~~比较~~ 调参 → 调优、调试.

LightGbm 过拟合，需要调参数： max_depth = -1, num-leaves = 30, learning_rate = 0.2

学习率 大 → ~~小~~ 速度 ✓

model.fit(- - - early_stopping_rounds = 10) 多次振荡，加速时长

怎么 stacking？
K-Fold.
上一层模型 — 多 oof → 训 → model → test.

多 oof 用 "Averaging" 的方法，总模型 每个权重等分数！

=> $w_1 oof_1 + w_2 oof_2 + w_n oof_n$ ← 线性回归方法，训练 $w_1\ w_2 \cdots w_n$ ↑ 先归一化标准化 z.B. StandardScaler

用固定的框架解决问题：

读取数据 → 数据预处理 →

特征工程 → 筛选特征 → 固定特征 → feats ~~[illegible]~~ cat_col

~~固定特征~~ 调参 → 跑多模型 → 模型融合

细化：机器学习有了步骤——搭建一个机器学习的框架.

类别特征 → label Encode ← (xgboost)

直接指定列的类别特征

lgb/catboost → data[i].astype('category')

训练回3个模型

样本量 → subsample=0.95

确定! → learning_rate=0.1

- xgb_model: objective='binary:logistic'
  - tree_method='gpu_hist'
  - max_depth=6 (这3个是最常用)
- lgb_model: objective='binary' metric='auc' subsample
  - learning rate = — max_length = -1
- cat_model:
  - task-type="GPU" eval-metric='auc' (评价指标)
  - learning rate = — max-length=7 (5~7)
  - iterations = 1000 subsample=0.95

回归之评价指标:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

值↓ 误差↓

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = 1 - \frac{SS_{residual}}{SS_{total}}$$

0: 效果差 ~ 1: 完美情况

时间序列的概述：根据以前的时间序列数据预测未来变化。

- 趋势的延续性。
- 未来的不确定性。(数据不完整)
- 不考虑因果关系。

时间序列
- 平稳序列 (随机性序列) ?
- 非平稳序列
  - 趋势性
  - 季节性
  - 随机性

预测步骤：

1. 数据平稳性检测。
2. 若不平稳，则差分化 (diff) 处理。
3. 确定最佳模型。
4. 模型预测与检验。

严平稳：统计性质不变。(正态分布，白噪声) $E(x)=0$，$MSE=1$

(弱) 宽平稳：Mean MSE 与时间无关；协方差 (cov) 与时间间隔有关。P75

非平稳：Mean, MSE, cov 与时间有关

ARIMA 模型：描述当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测。

AR I MA

- AR：自动回归模型 autoregression（满足平稳性）
- I：差分法 integrated
- MA：移动平均 Moving Average，是误差项的累加。(当前值与多少个值有关系) 消除随机波动！

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

(q：误差项系数)

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-1} + \varepsilon_t$$

- p：阶数，预定的
- μ：常数C
- r_i：自相关系数
- ε_t：误差

ARIMA：

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

- P, Q 自定
- r_i, θ_i 求解
- μ, ε_t 已知

必须自相关性：统计特征 (MSE, Mean) 确定 $\varphi_i \geq 0.5$

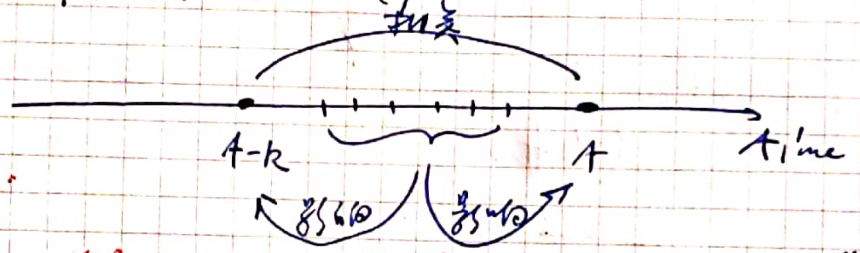
与自身相关

ACF：自相关系数 (Autocorrelation function)

$$ACF(k) = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)}$$

$\rho_k \in [-1 \sim 1]$

PACF：偏自相关系数 (partial autocorrelation function)



剔除掉P 得出真正的"ACF"

确定

	ACF	PACF
AR(p)	拖尾	p阶后截尾
MA(q)	q阶后截尾	拖尾
ARMA(q,p)	q阶后拖尾	p阶后拖尾

→ 确定 I (0,1,2)

模型残差检测：① ARIMA模型 均0或者常数的正态分布.

② QQ图：线性 (服从正态分布)

什么是差分?

统计学模型

```
from statsmodels.tsa.arima_model import ARIMA.
from    "    .graphics.tsaplots import plot_acf, plot_pacf.
```