NJUAI 南京大学 人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY
南京大學 NANJING UNIVERSITY 1902

02 时间序列的预处理和评价

时间序列分析的预备工作

本章概要

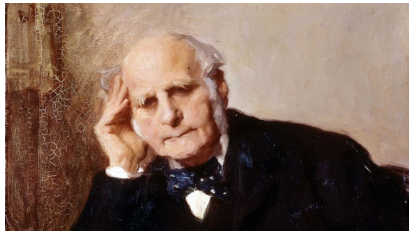
1. 时间序列的常用数据集

2. 时间序列预测的评价方式和指标

3. 时间序列的预处理方法

时间序列预测竞赛

古希腊，占卜竞争



【1707】【英】Francis Galton发现融合多个预测模型将提升预测性能

【1974】Paul Newbold和Clive Granger将不同方法在106个时序数据集上进行对比，讨论最精确的方法

【英】Maurice Priestley

时间序列分析本身是为了找到最能够反应时序数据特性的单一模型，而预测任务的目标为找到这一模型。如果时间序列满足Box-Jenkins模型（ARIMA）的假设，则使用ARMIA就是最优的选择

Rob J.Hyndman. A brief history of forecasting competitions. International Journal of Forecasting. 36(1), page 7-14. 2020.

时间序列数据集M4

2018年，Spyros Makridakis等人开展M4竞赛，其中包含100,000个涉及更多领域的时间序列预测任务

Time interval between successive observations	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	6,538	3,716	3,903	6,519	1,088	1,236	**23,000**
Quarterly	6,020	4,637	5,315	5,305	1,858	865	**24,000**
Monthly	10,975	10,017	10,016	10,987	5,728	277	**48,000**
Weekly	112	6	41	164	24	12	**359**
Daily	1,476	422	127	1,559	10	633	**4,227**
Hourly	0	0	0	0	0	414	**414**
Total	**25,121**	**18,798**	**19,402**	**24,534**	**8,708**	**3,437**	**100,000**

Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, The M4 Competition: **100,000 time series** and 61 forecasting methods. International Journal of Forecasting. 2020,

时间序列数据集M4

Methods			Description
Point Forecasts (PFs)	**Statistical benchmarks**	Naïve 1	A random walk model, assuming that future values will be the same as that of the last known observation.
		Naïve S	Forecasts are equal to the last known observation of the same period.
		Naïve 2	Like Naïve 1 but the data are seasonally adjusted, if needed, by applying a classical multiplicative decomposition. A 90% autocorrelation test is performed to decide whether the data are seasonal.
		SES	Exponentially smoothing the data and extrapolating assuming no trend. Seasonal adjustments are considered as per Naïve 2.
		Holt	Exponentially smoothing the data and extrapolating assuming a linear trend. Seasonal adjustments are considered as per Naïve 2.
		Damped	Exponentially smoothing the data and extrapolating assuming a damped trend. Seasonal adjustments are considered as per Naïve 2.
		Theta	As applied to the M3 Competition using two Theta lines, $\vartheta_1 = 0$ and $\vartheta_2 = 2$, with the first one being extrapolated using linear regression and the second one using SES. The forecasts are then combined using equal weights. Seasonal adjustments are considered as per Naïve 2.
		Comb	The simple arithmetic average of SES, Holt and Damped exponential smoothing (*used as the single benchmark for evaluating all other methods*).
	ML benchmarks	MLP	A perceptron of a very basic architecture and parameterization. Some preprocessing like detrending and deseasonalization is applied beforehand to facilitate extrapolation.
		RNN	A recurrent network of a very basic architecture and parameterization. Some preprocessing like detrending and deseasonalization is applied beforehand to facilitate extrapolation.
	Standards for comparison	ETS	Automatically provides the best exponential smoothing model, indicated through information criteria.
		ARIMA	An automatic selection of possible ARIMA models is performed and the best one is chosen using appropriate selection criteria.

M4竞赛的（部分）重要发现

集成（combination）的统计方法或机器学习方法能够获得更好的时间序列预测性能（这一点和之前的一些观测一样）；

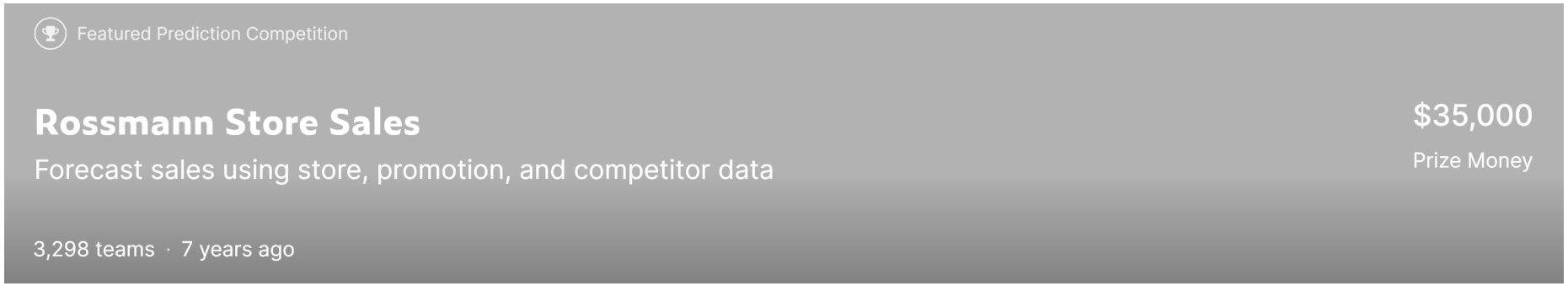
混合（hybrid）统计和机器学习方法具有优势。如Uber的Slawek Smyl结合统计方法指数平滑以及RNN，获得了极好的预测效果；

更复杂的方法可能可以获得更高的预测性能（和以往M竞赛结论不一致）；

通过综合多个时间序列的信息（information from aggregates of series）提升目标时间序列的预测能力；

单纯的机器学习方法无法取得较好的预测性能。

其他竞赛



Kaggle：
Rossmann store sales

Kaggle：
Wikipedia traffic forecast

KDD CUP 2022

Baidu KDD CUP 2022 Closed

Spatial Dynamic Wind Power Forecasting. This task has practical importance for the utilization of wind energy. Participants are expected to accurately estimate the wind power supply of a wind farm.

Tag：KDD　Competition Time：2022/03/16 - 2022/07/19

Sponsor：Baidu 大脑　国家能源集团 CHN ENERGY

KDD Cup：
Spatial Dynamic Wind Power Forecasting

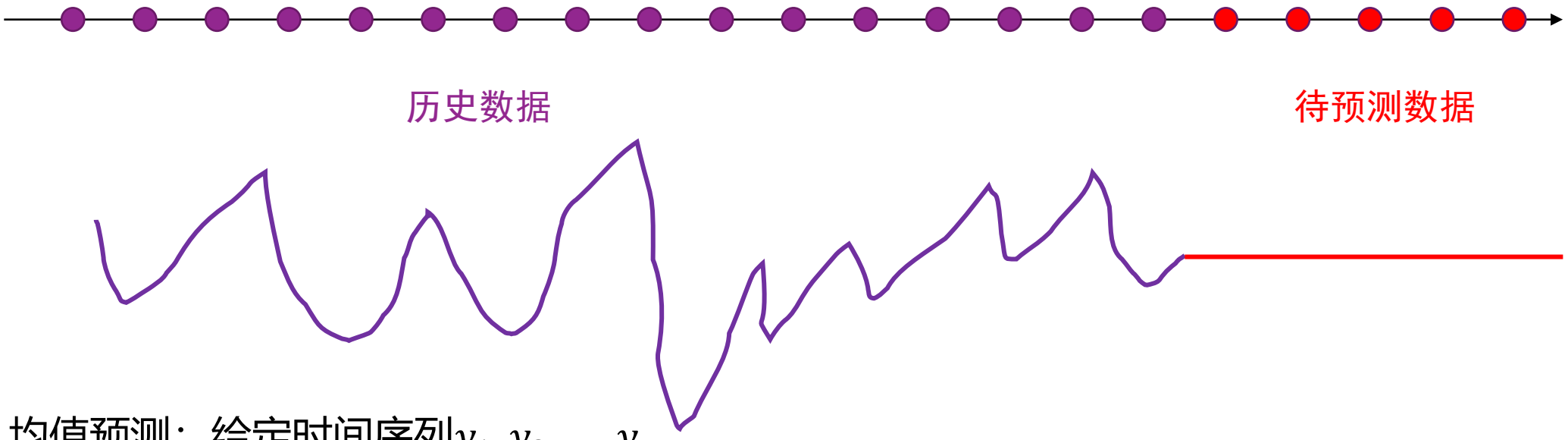
其他相关数据集

- **ETT** (Electricity Transformer Temperature)：包含两个子集，分别含有2个小时级别（ETTh）记录的和15分钟级（ETTm）别记录的数据集。 每一个数据集包含2016-2018年的7个油和附在特征。https://github.com/zhouhaoyi/ETDataset
- **Traffic**：包含小时级别记录的旧金山2015-2016交通数据。http://pems.dot.ca.gov/
- **Electricity**：包含2012-2014年321个客户的小时级别电力消耗数据。
  https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
- **Exchange-Rate:** 包含1990-2016年8个国家的 collects the 汇率。https://github.com/laiguokun/multivariate-time-series-data
- **Weather**：包含2020年德国的每10分钟记录的天气数据，共21个特征，如温度、湿度等。https://www.bgc-jena.mpg.de/wetter/
- **ILI7**: 记录2002-2021美国疾控中心每周流感流感症状疾病患者以及患者总数的比例。
  https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Datasets	ETTh1&ETTh2	ETTm1 &ETTm2	Traffic	Electricity	Exchange-Rate	Weather	ILI
Variates	7	7	862	321	8	21	7
Timesteps	17,420	69,680	17,544	26,304	7,588	52,696	966
Granularity	1hour	5min	1hour	1hour	1day	10min	1week

Baseline方法

- Naïve1：常数预测模型。使用最后一个观察值对后续样本进行预测；相当于给时间序列构建一个**随机游走**模型。



- 均值预测：给定时间序列$y_1, y_2, \ldots, y_n$

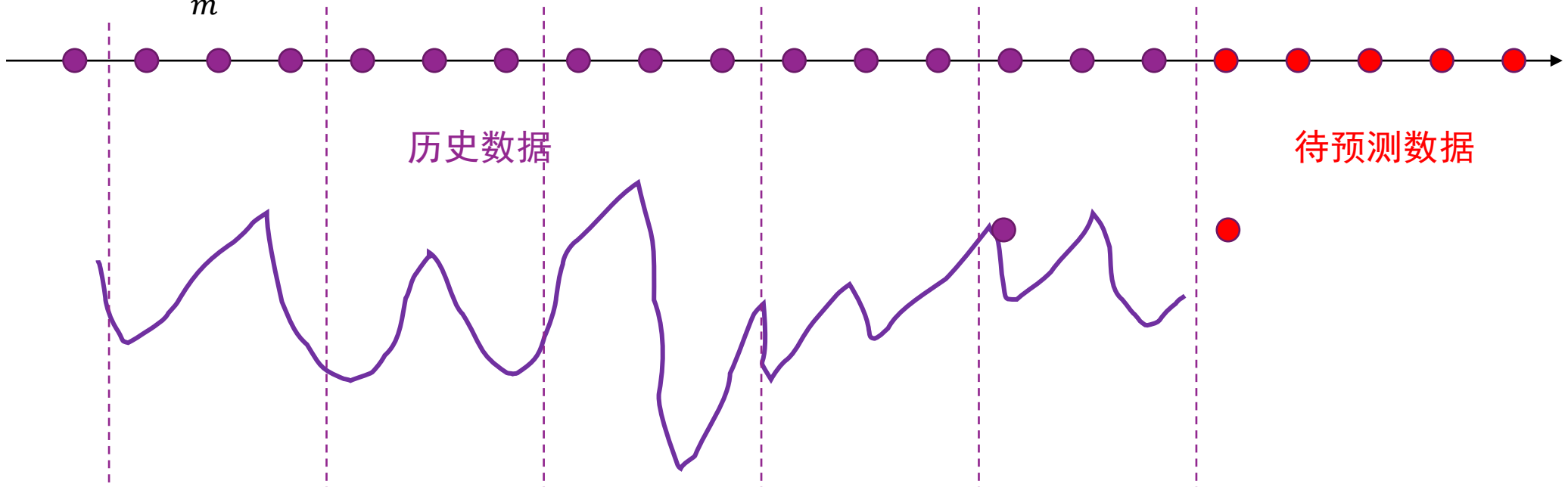
$$y_{n+h} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

Baseline方法

- NaïveS：对于**周期性**时间序列，使用上一个周期同期的观察值作为当前时刻的预测值。这一方法充分考虑了周期性，因此需要给定算法的记录粒度（granularity）。
- 给定时间序列$y_1, y_2, \ldots, y_n$，其周期为m，则

$$\hat{y}_{n+h} = y_{n+h-m(k+1)}$$

其中，k是$\frac{h-1}{m}$的整数部分，索引到最近一个周期中同一位置处的取值。



Baseline方法

- Drift方法：充分考虑到时间序列前后的变化。每两个相邻的时间序列可以计算变化值的均值，即

$$\frac{h}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1})$$

- 因此，可用这一变化值指导后续的预测

$$\hat{y}_{n+h} = y_n + \frac{h}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1})$$

$$= y_T + h\left(\frac{y_n - y_1}{n-1}\right)$$

本章概要

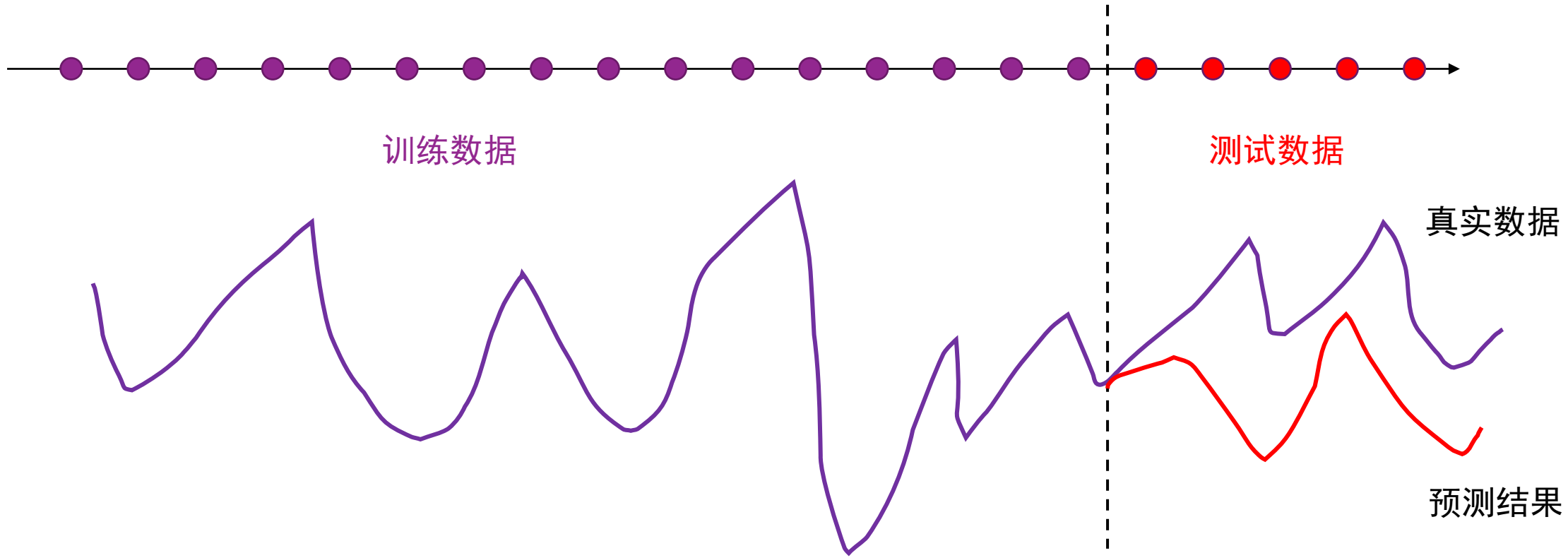
1. 时间序列的常用数据集

2. 时间序列预测的评价方式和指标

3. 时间序列的预处理方法

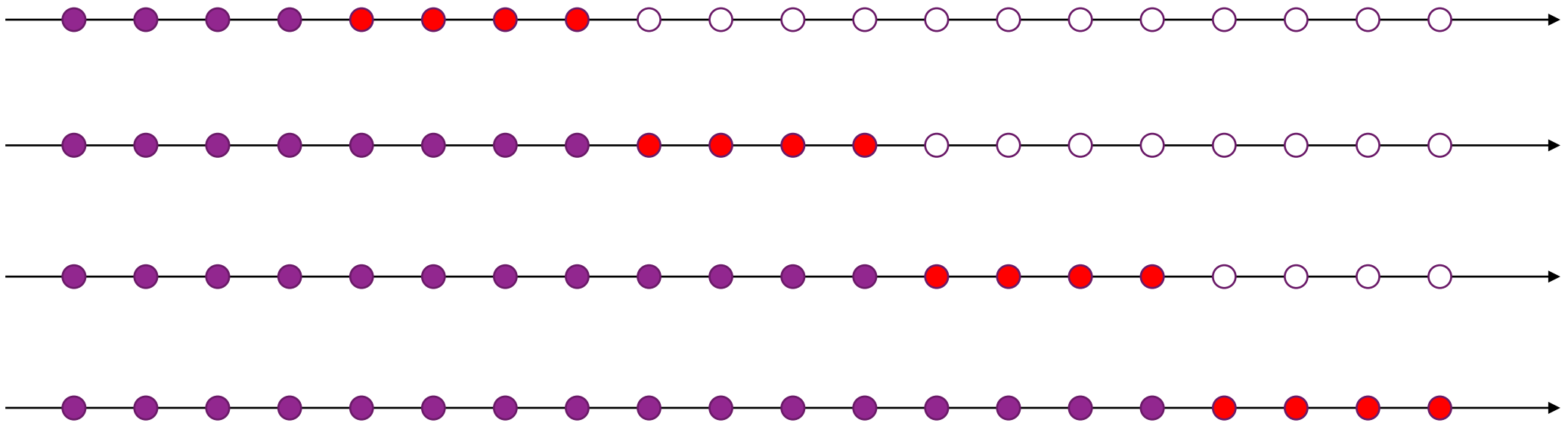
评估方式

- 划分出一段数据作为测试集（hold-out）



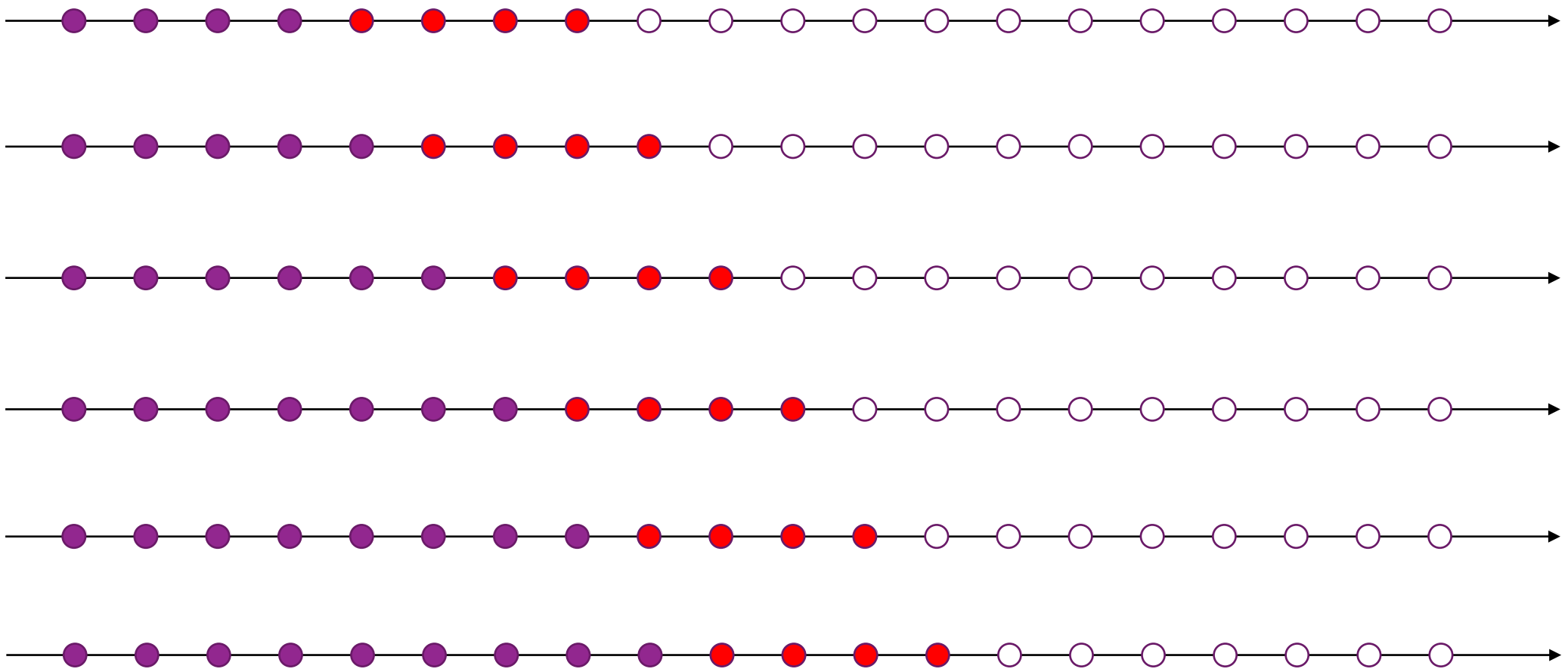
评估方式

- 交叉验证（Cross-Validation）



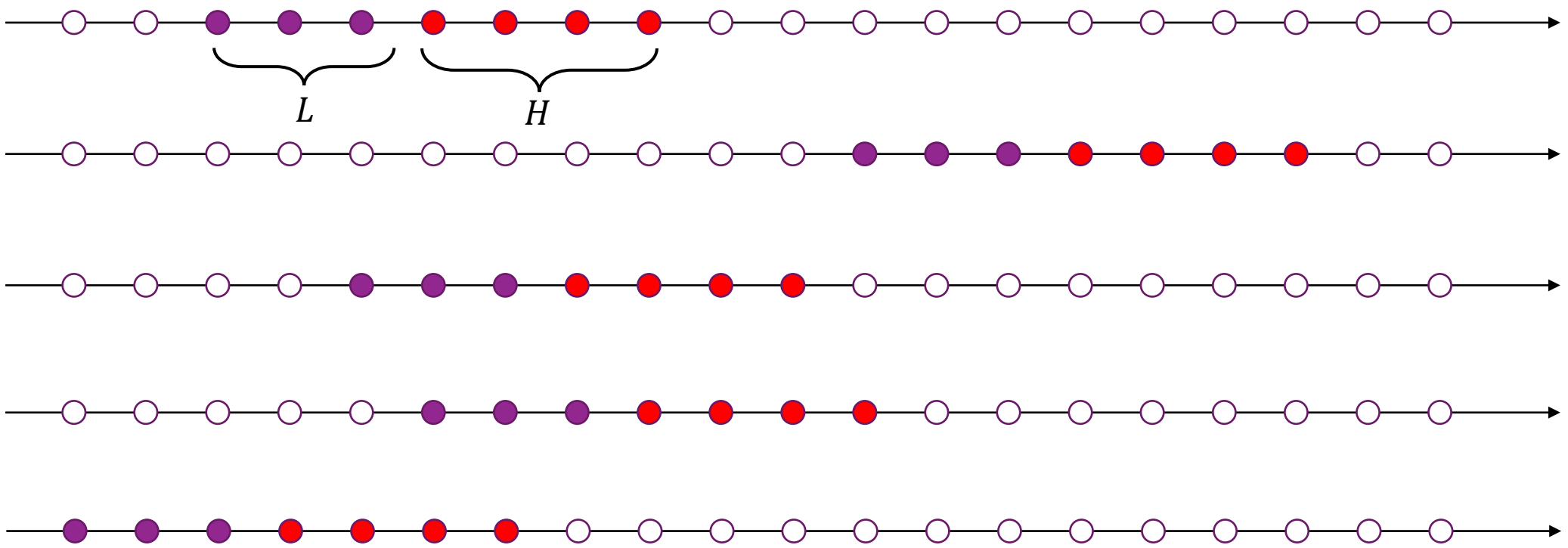
评估方式

- 循环验证（rolling），基于历史数据验证backtesting



评估方式

- $L-H$评测
  - 基于长度为L的历史数据，预测长度为H的后续数据
  - 在时间序列中，抽取出类似的$L-H$对，将时间序列预测问题进行变化为一般的“机器学习”子任务



预测评价指标

- **MAE** (Mean Absolute Error)

$$\mathrm{MAE} = \frac{1}{H}\sum_{i=1}^{H} |y_{n+i} - \hat{y}_{n+i}|$$

- **MSE** (Mean Square Error)

$$\mathrm{MSE} = \frac{1}{H}\sum_{t=1}^{H} (y_{n+t} - \hat{y}_{n+t})^2$$

预测评价指标

- $\hat{y}$ 为y的预测值，m为数据周期（如按月统计的数据有$m = 12$）

- **MAPE** (Mean Absolute Percentage Error)

$$\text{MAPE} = \frac{100}{H}\sum_{i=1}^{H}\frac{|y_{T+i} - \hat{y}_{T+i}|}{|y_{T+i}|}$$

- MASE (Mean Absolute Scaled Error)

$$\text{MASE} = \frac{1}{H}\sum_{i=1}^{H}\frac{|y_{T+i} - \hat{y}_{T+i}|}{\frac{1}{T+H-m}\sum_{j=m+1}^{T+H}|y_j - y_{j-m}|}$$

- sMAPE (symmetric MAPE)

$$\text{SMAPE} = \frac{200}{H}\sum_{i=1}^{H}\frac{|y_{T+i} - \hat{y}_{T+i}|}{|y_{T+i}| + |\hat{y}_{T+i}|}$$

- OWA (overall weighted average)

$$\text{OWA} = \frac{1}{2}\left[\frac{\text{SMAPE}}{\text{SMAPE}_{\text{Naïve2}}} + \frac{\text{MASE}}{\text{MASE}_{\text{Naïve 2}}}\right]$$

本章概要

1. 时间序列的常用数据集

2. 时间序列预测的评价方式和指标

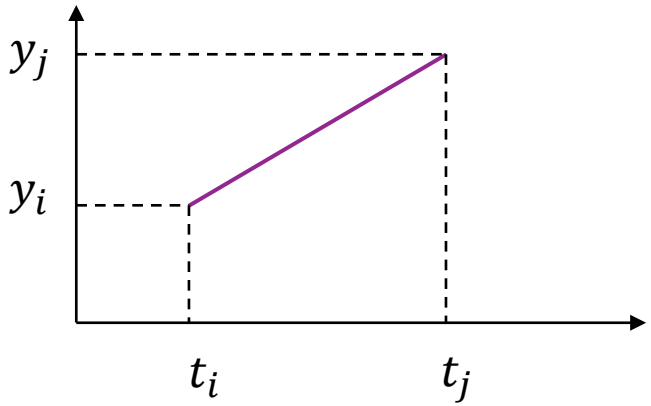
3. 时间序列的预处理方法

应对缺失值和噪声

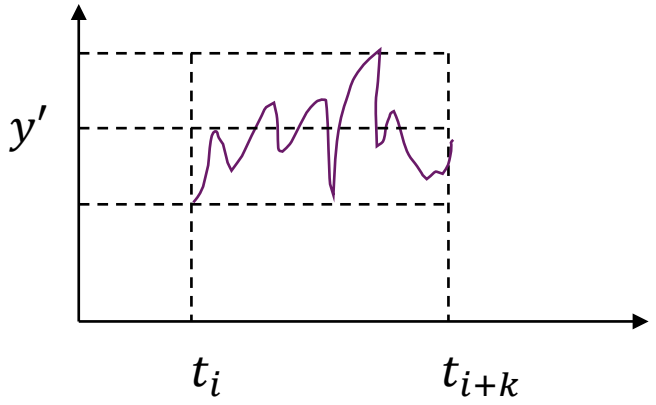
- 处理缺失数据：基于上下文进行补全
- 线性（内）插值（linear interpolation）：假设$i < j$，$t \in (t_i, t_j)$，则

$$y = y_i + \left(\frac{t - t_i}{t_j - t_i}\right) \cdot (y_j - y_i)$$

插值过程中也可以使用更复杂的方法进行拟合



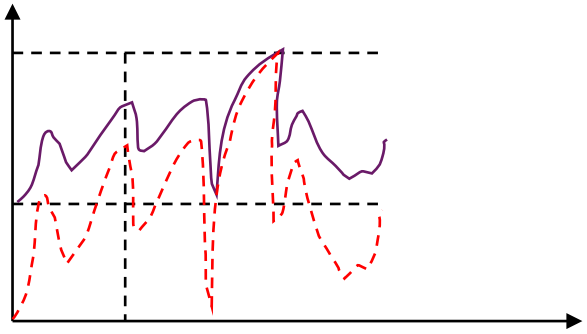
- 处理时序数据噪声
  - 一般用于去除短时间内的数据扰动（short-term fluctuations）
  - 装箱（Binning）
    - 将时间序列按照一定间隔分组（如间隔k），使用**均值**代替原始的k个值



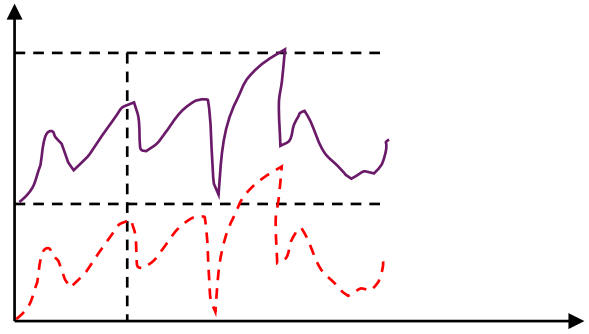
  - 移动平均（Moving-Average）

时间序列的归一化

- 时序数据尺度（scale）变化



- 时序数据平移（translation）变化



- 目标：
  - 对于任意尺度常数a和平移常数b，时间序列的线性变换$ax + b$不影响其相似度的计算
- 归一化（normalization）：将时序数据取值限制在[0,1]

$$y_t' = \frac{y_t - y_{min}}{y_{max} - y_{min}}$$

- 标准化（Standardization）/ Z-Score：将时序数据变换为0均值以及标准方差

$$z_t = \frac{y_t - \mu}{\sigma}$$

- 平均归一化（Mean Normalization）

$$y_t' = \frac{y_t - \mu}{y_{max} - y_{min}}$$

其他的时间序列预处理方法

- 时间序列的分解
  - 相加分解（additive decomposition）

$$y_t = S_t + T_t + R_t$$

  - 相乘分解（multiplicative decomposition）

$$y_t = S_t \times T_t \times R_t$$

  - S_t：季节项（seasonal component），刻画时间序列的周期性变换
  - T_t：趋势项（trend-cycle component），刻画序列的整体变化趋势
  - R_t：剩余项（remainder component）

针对序列的变换

- 对数变换
  - 针对变化程度建模
  - Make highly skewed distributions less skewed

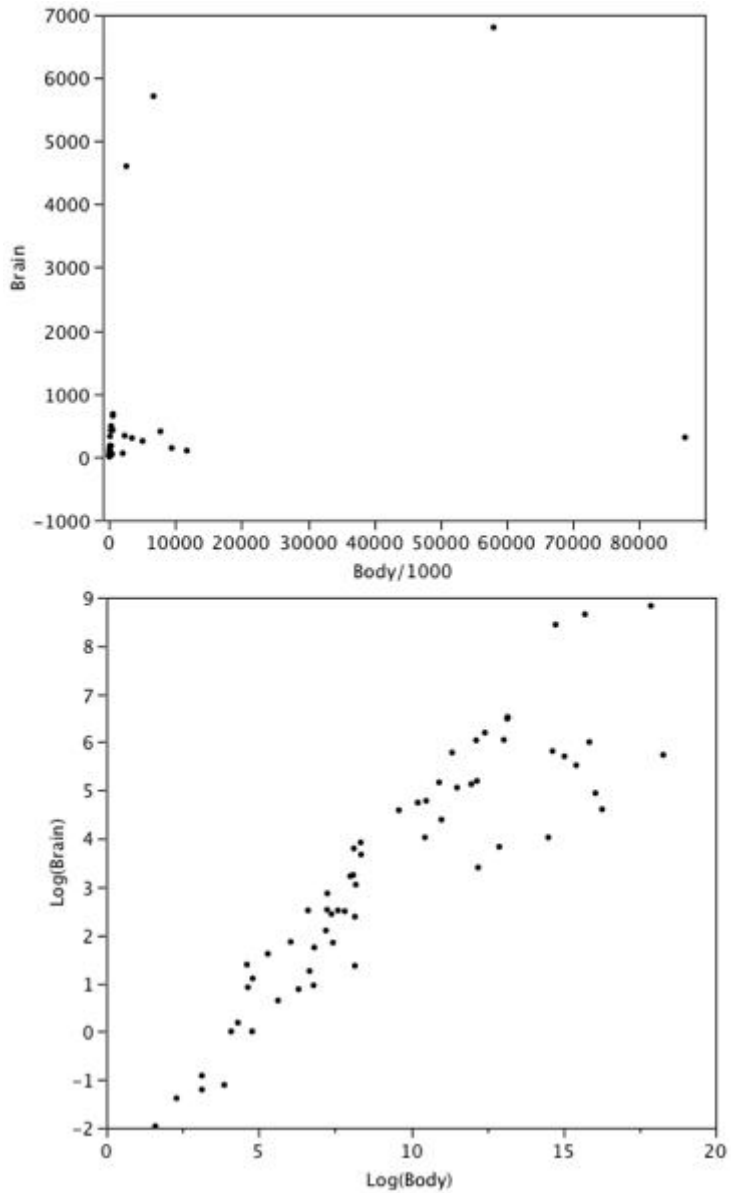
$$\nabla \log(y_t) = \log\left(\frac{y_t}{y_{t-1}}\right) = \log\left(1 + \frac{\nabla y_t}{y_{t-1}}\right) \simeq \frac{\nabla y_t}{y_{t-1}}$$

$\frac{\nabla y_t}{y_{t-1}}$反应了相对变化（上述近似要求相对变化不能过大）

- 变换之后的均值相当于原始的几何平均

X	$Log_{10}(X)$
1	0
10	1
100	2

几何平均$(1 \times 10 \times 100)^{1/3} = 10$　　均值1

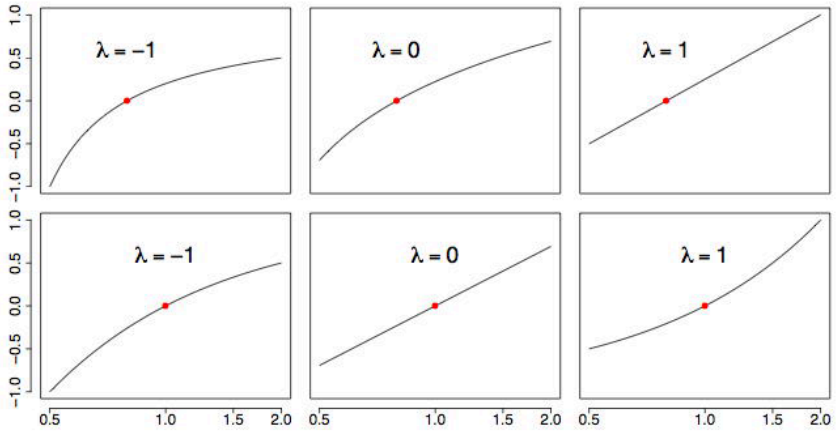


Box-Cox变换

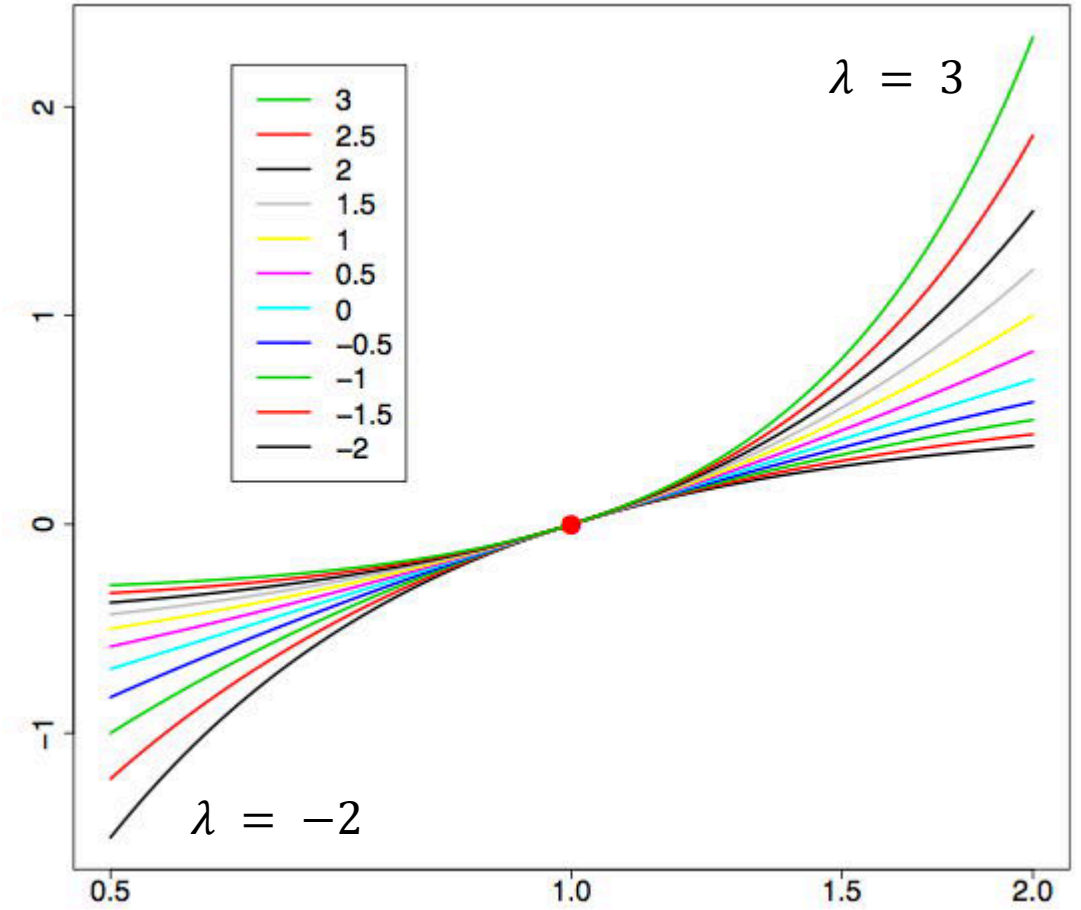
- Box-Cox变换（1964）
  - 数据值域较大
  - 数据呈现季节性变换

$$y_t^{(\lambda)} = \begin{cases} \dfrac{(y_t^{\lambda} - 1)}{\lambda} & \lambda \neq 0 \\ \log y_t & \lambda = 0 \end{cases}$$

- $y_t = 1$有$y_t^{(\lambda)} = 0$



$y_t^{(\lambda)}$
$-x$



$y_t^{(\lambda)}$
$-\log(x)$

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, *26*, 211-252.

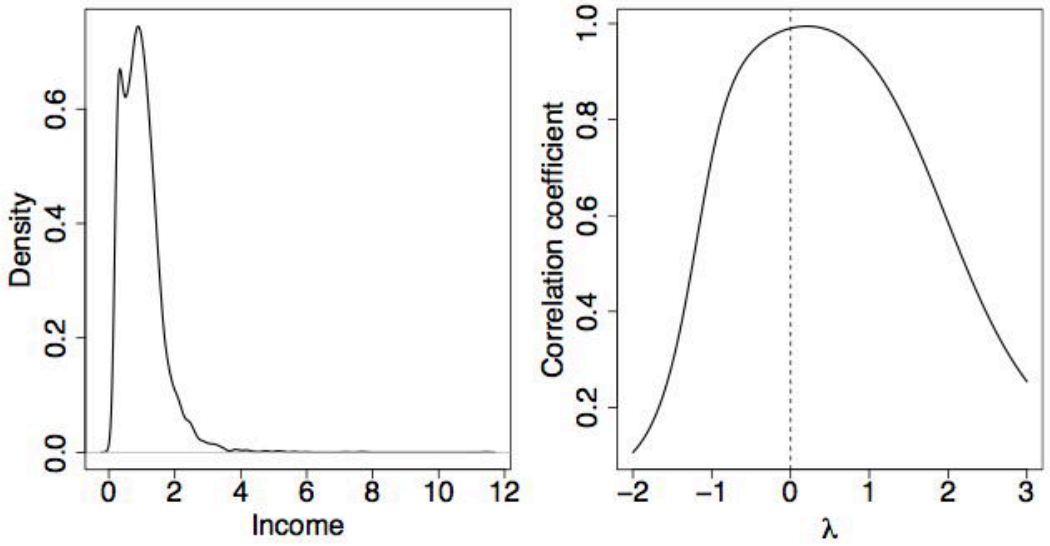
Box-Cox变换

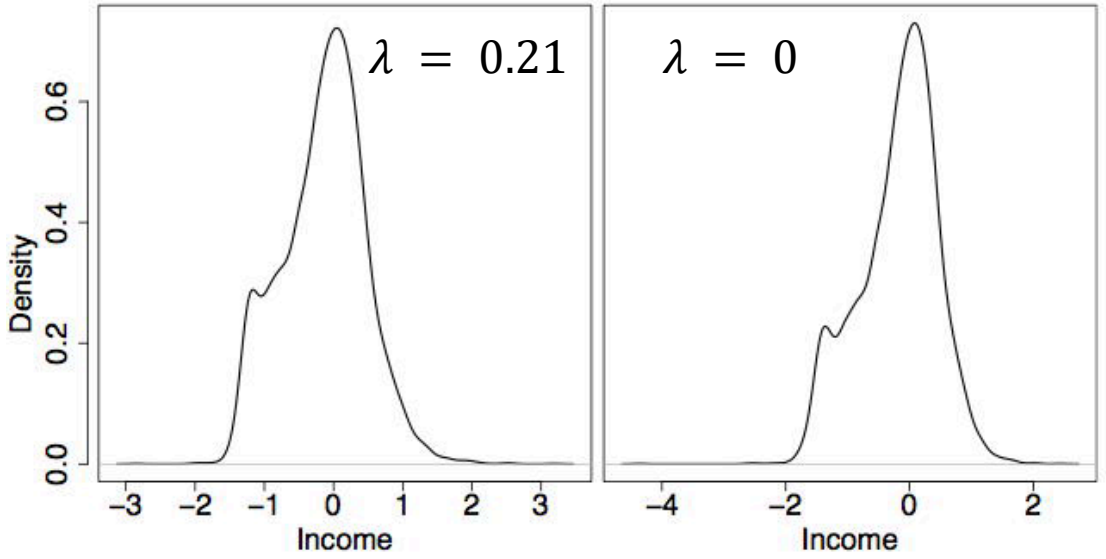
- Box-Cox变换用于分布"正态"程度矫正

$$y_t^{(\lambda)} = \begin{cases} \dfrac{(y_t^{\lambda} - 1)}{\lambda} & \lambda \neq 0 \\ \log y_t & \lambda = 0 \end{cases}$$

$$\left(\Phi^{-1}\left(\frac{i - 0.5}{n} \right), y_t \right)$$

1973 British income data





Box-Cox变换

- 在回归/时序任务中

$$\min_{\theta} \sum_{i}^{N} (y_i - \hat{y}_i)^2$$

- 变换之后尺度有影响如何处理?

$\dot{y}$为y_t序列的几何平均

$$y_t^{(\lambda)} = \begin{cases} \dfrac{\left(y_t^{\lambda} - 1\right)}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} \log y_t & \lambda = 0 \end{cases}$$

λ	S_λ	λ	S_λ	λ	S_λ
−0.4	13,825.5	−0.1	11,627.2	0.2	11,784.3
−0.3	12,794.6	0.0	11,458.1	0.3	12,180.0
−0.2	12,046.0	0.1	11,554.3	0.4	12,633.2

如何处理负值序列?

- Box-Cox变换只针对取值非负的序列，如何处理取值可能为负值的序列?

- 二参数Box-Cox变换
- (two-parameter Box-Cox)

$$y_t^{(\lambda)} = \begin{cases} \dfrac{(y_t + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \ln(y_t + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

  - 允许$y_t > -\lambda_2$

- 修正的Box-Cox变换（Bickel & Doksum）

$$y_t^{(\lambda)} = \begin{cases} \log(y_t) & \text{if } \lambda = 0 \\ \left(\text{sign}(y_t)|y_t|^{\lambda} - 1\right)/\lambda & \textit{otherwise.} \end{cases}$$

$\lambda = 1$时，数据形状不变，但向下平移

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. Journal of the American Statistical Association, 76(374), 296–311

如何处理负值序列?

- Box-Cox变换只针对取值非负的序列，如何处理取值可能为负值的序列?

- Yeo–Johnson变换

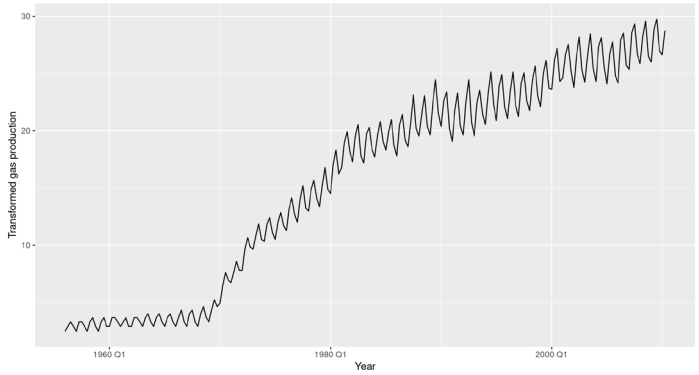
$$y_t^{(\lambda)} = \begin{cases} \left((y_t+1)^{\lambda}-1\right)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y_t+1) & \text{if } \lambda = 0, y \geq 0 \\ -\left((-y_t+1)^{(2-\lambda)}-1\right)/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y_t+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

  - $\lambda = 1$时为恒等变换

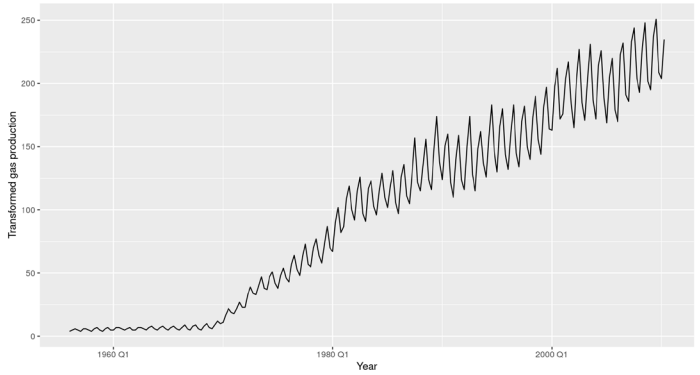
Yeo, In-Kwon; Johnson, Richard A. (2000). "A New Family of Power Transformations to Improve Normality or Symmetry". Biometrika. 87 (4): 954–959.

Box-Cox变换对时序的影响

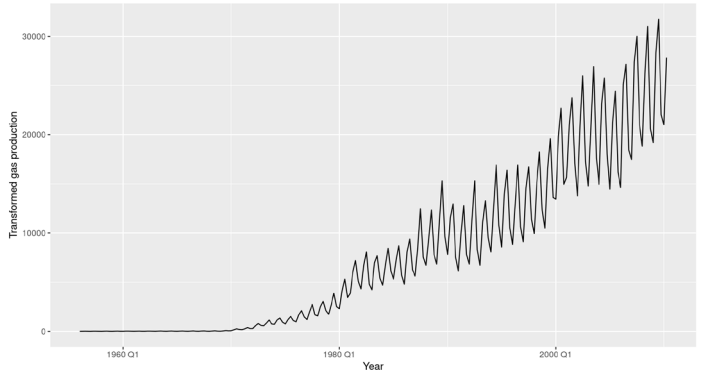
- 针对Australian Quarterly Gas Production数据，使用修正的Box-Cox变换



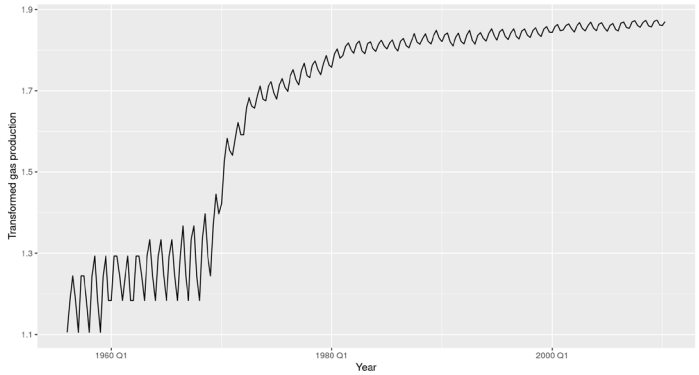
$\lambda = 0.5$



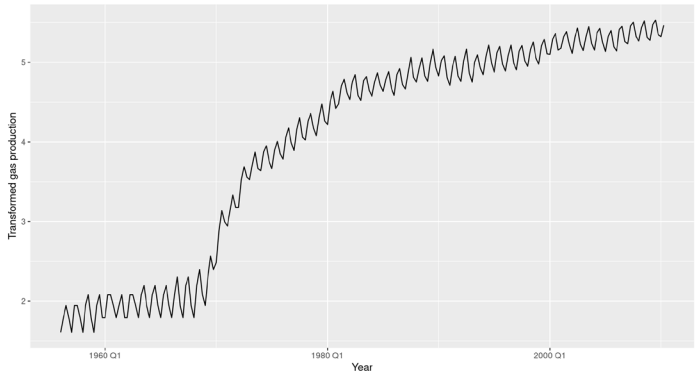
$\lambda = 1$



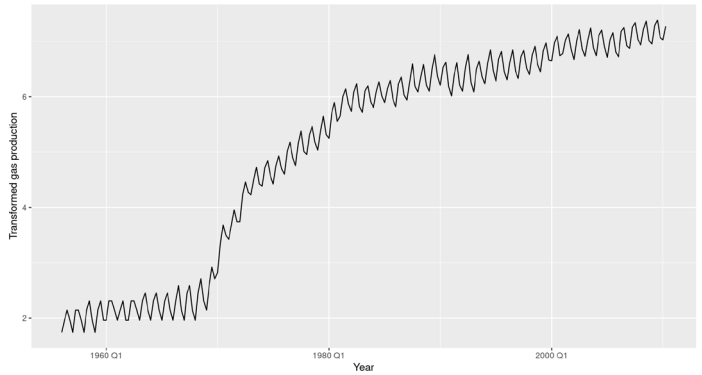
$\lambda = 2$



$\lambda = -0.5$



$\lambda = 0$



$\boldsymbol{\lambda = 0.1}$

Tukey Ladder of Powers

- Tukey Ladder of Powers用于将有偏的分布"矫正"，趋向于正态分布（辅助概率化建模）
  - 考虑如下变换，使变化后关系（x^λ或y^λ）趋于**线性**

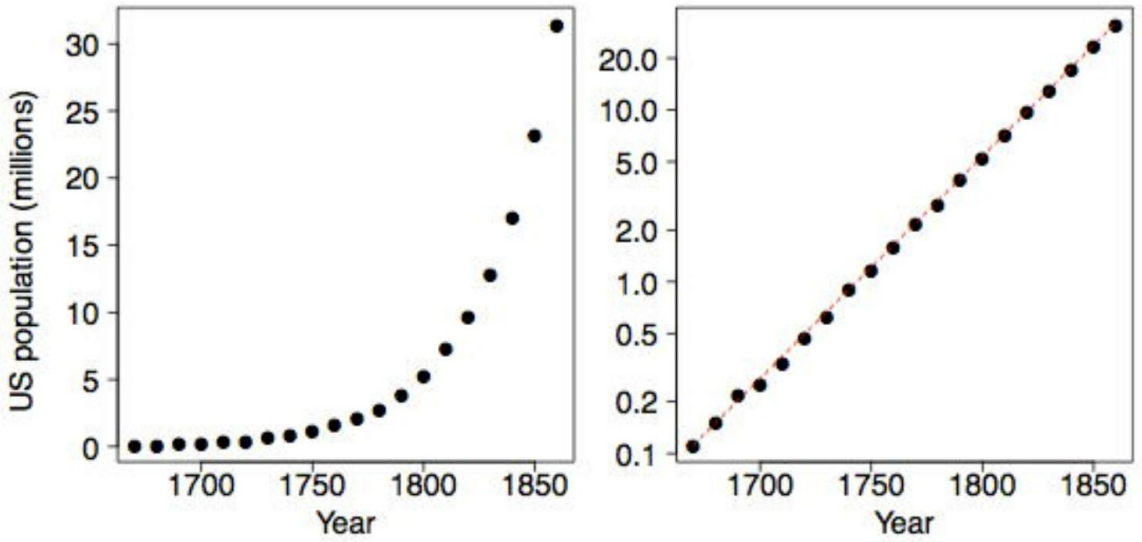
$$y = b_0 + b_1 x^\lambda$$

或

$$y^\lambda = b_0 + b_1 x$$

- Tukey定义当$\lambda = 0$时变换为log(·)而非一个常数

λ	-2	-1	$-1/2$	0	$1/2$	1	2
y	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	log$x$	$\sqrt{x}$	x	x^2



The US population from 1670 - 1860.
The Y axis on the right panel is on a log scale.

一般限制$x > 0$

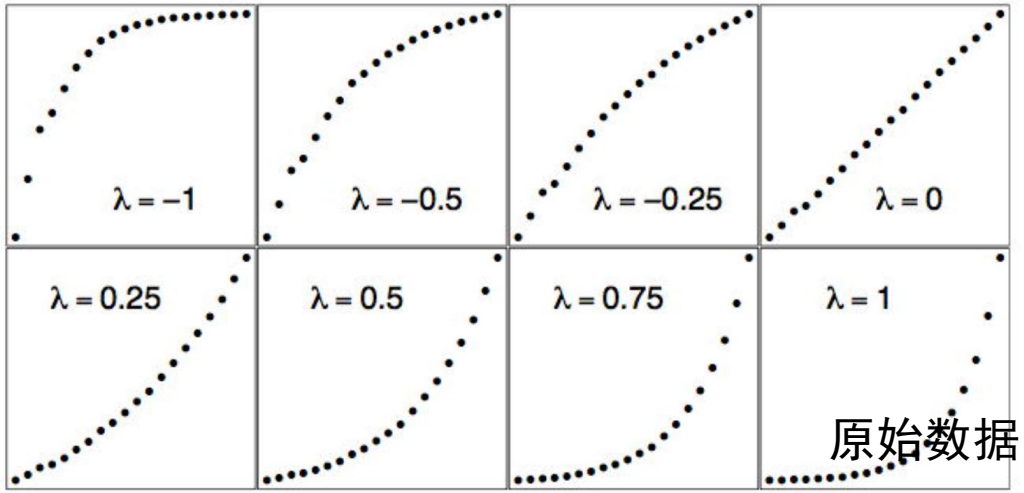
Tukey Ladder of Powers

- 考虑到λ的影响，如果$\lambda < 0$很可能会改变原始数据的变化趋势，因此，定义Tukey变换

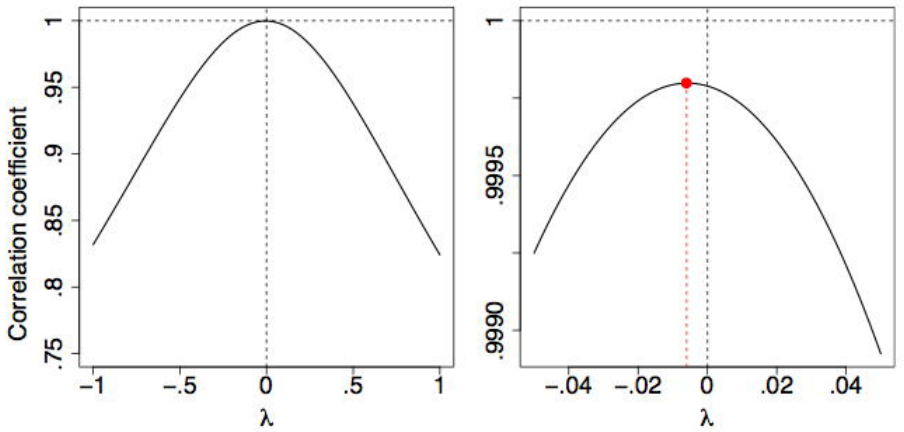
$$y = \begin{cases} x^{\lambda} & \text{if } \lambda > 0 \\ \log x & \text{if } \lambda = 0 \\ -\left(x^{\lambda}\right) & \text{if } \lambda < 0 \end{cases}$$

λ	-2	-1	$-1/2$	0	$1/2$	1	2
y	$-\frac{1}{x^2}$	$-\frac{1}{x}$	$-\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

寻找λ，使变化后关系（y_λ）趋于**线性**



原始数据



(x, y_λ)的相关性和λ的关系

Box-Cox变换用于噪声学习

- CE损失（针对真实数据）

$$R_{\mathcal{L}}(f) = \mathbb{E}_D[\mathcal{L}(f(\boldsymbol{x};\boldsymbol{\theta}), y_{\boldsymbol{x}})] = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} \boldsymbol{y}_{ij} \log f_j(\boldsymbol{x}_i;\boldsymbol{\theta}),$$

uniform

noise tolerant

- 噪声数据$D_\eta = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$，假设噪声和样本、类别无关

$$p(\tilde{y}_i = k \mid y_i = j, \boldsymbol{x}_i) = p(\tilde{y}_i = k \mid y_i = j) = \eta_{jk}$$

$$\eta_{jk} = 1 - \eta \text{ for } j = k$$

$$\eta_{jk} = \frac{\eta}{c-1} \text{ for } j \neq k$$

symmetric

- 针对噪声数据优化$R_{\mathcal{L}}^{\eta}(f) = \mathbb{E}_{D_\eta}[\mathcal{L}(f(\boldsymbol{x}), \tilde{y}_{\boldsymbol{x}})]$

Zhilu Zhang, Mert R. Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. NeurIPS 2018: 8792-8802

CE和MAE

- CE损失

$$R_{\mathcal{L}}(f) = \mathbb{E}_D[\mathcal{L}(f(\boldsymbol{x};\boldsymbol{\theta}), y_{\boldsymbol{x}})]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} \boldsymbol{y}_{ij}\log f_j(\boldsymbol{x}_i;\boldsymbol{\theta}),$$

- Nonsymmetric, sensitive to label noise

- MAE损失

$$\mathcal{L}_{MAE}\big(f(\boldsymbol{x}), e_j\big) = \|e_j - f(\boldsymbol{x})\|_1$$

$$= 2 - 2f_j(\boldsymbol{x})$$

- symmetric, noise robust

如何综合两个损失函数?

$$\sum_{i=1}^{n}\frac{\partial\mathcal{L}(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)}{\partial\boldsymbol{\theta}} = \begin{cases} \sum_{i=1}^{n} -\frac{1}{f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})}\nabla_\theta f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}) \text{ for CE} \\ \sum_{i=1}^{n} -\nabla_\theta f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}) \text{ for MAE/unhinged loss} \end{cases}$$

Generalized Cross Entropy Loss

- CE损失

$$R_{\mathcal{L}}(f) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} \boldsymbol{y}_{ij}\log f_j(\boldsymbol{x}_i;\boldsymbol{\theta}),$$

- MAE损失

$$\mathcal{L}_{MAE}\big(f(\boldsymbol{x}), e_j\big) = \|e_j - f(\boldsymbol{x})\|_1 = 2 - 2f_j(\boldsymbol{x})$$

- 使用**Box-Cox变换**

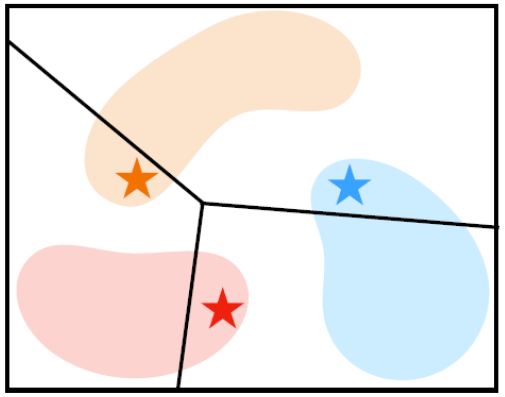
$$\mathcal{L}_q\big(f(\boldsymbol{x}), e_j\big) = \frac{\big(1 - f_j(\boldsymbol{x})^q\big)}{q}$$

其中，$q \in (0,1]$

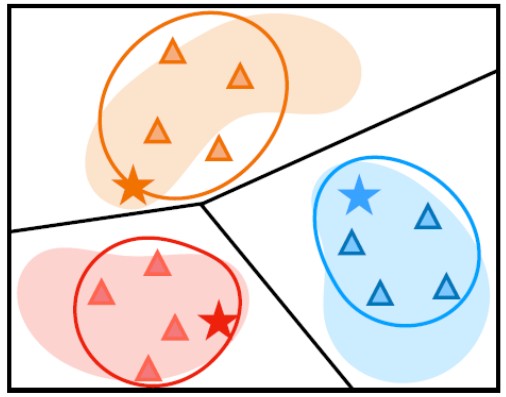
- $q \to 0$，变化为CE，$q = 1$, 变化为MAE

$$\frac{\partial \mathcal{L}_q(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)}{\partial \boldsymbol{\theta}} = f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})^q\left(-\frac{1}{f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})}\nabla_\theta f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})\right) = -f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})^{q-1}\nabla_\theta f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}),$$

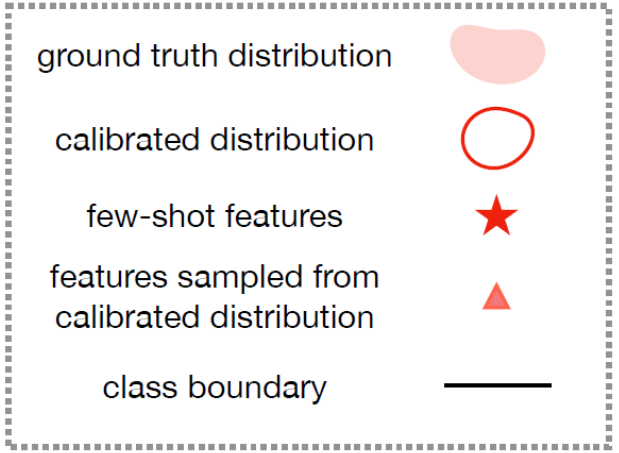
Tukey变换用于小样本学习



Classifier trained with few-shot features



Classifier trained with features sampled from calibrated distribution



Algorithm 1 Training procedure for an N-way-K-shot task

Require: Support set features $\mathcal{S} = (\boldsymbol{x}_i, y)_{i=1}^{N \times K}$

Require: Base classes' statistics $\{\boldsymbol{\mu}_i\}_{i=1}^{|C_b|}, \{\boldsymbol{\Sigma}_i\}_{i=1}^{|C_b|}$

1: Transform $(\boldsymbol{x}_i)_{i=1}^{N \times K}$ with Tukey's Ladder of Powers as Equation 3

2: **for** $(\boldsymbol{x}_i, y_i) \in \mathcal{S}$ **do**

3: Calibrate the mean $\boldsymbol{\mu}'$ and the covariance $\boldsymbol{\Sigma}'$ for class y_i using $\boldsymbol{x}_i$ with Equation 6

4: Sample features for class y_i from the calibrated distribution as Equation 7

5: **end for**

6: Train a classifier using both support set features and all sampled features as Equation 8

$$\tilde{\mathbf{x}} = \begin{cases} \boldsymbol{x}^{\lambda} & \text{if } \lambda \neq 0 \\ \log(\boldsymbol{x}) & \text{if } \lambda = 0 \end{cases}$$

$$\boldsymbol{\mu}' = \frac{\sum_{i \in \mathbb{S}_N} \boldsymbol{\mu}_i + \tilde{\boldsymbol{x}}}{k+1}, \boldsymbol{\Sigma}' = \frac{\sum_{i \in \mathbb{S}_N} \boldsymbol{\Sigma}_i}{k} + \alpha$$

$$\mathbb{D}_y = \{(\boldsymbol{x}, y) | \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \forall (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{S}^y\}.$$

$$\ell = \sum_{(\boldsymbol{x}, y) \sim \tilde{S} \cup \mathbb{D}_{y, y \in \mathcal{Y}^{\mathcal{T}}}} - \log \Pr(y | \boldsymbol{x}; \theta)$$

Shuo Yang, Lu Liu, Min Xu. Free Lunch for Few-shot Learning: Distribution Calibration. ICLR 2021