

MATLAB EXPO

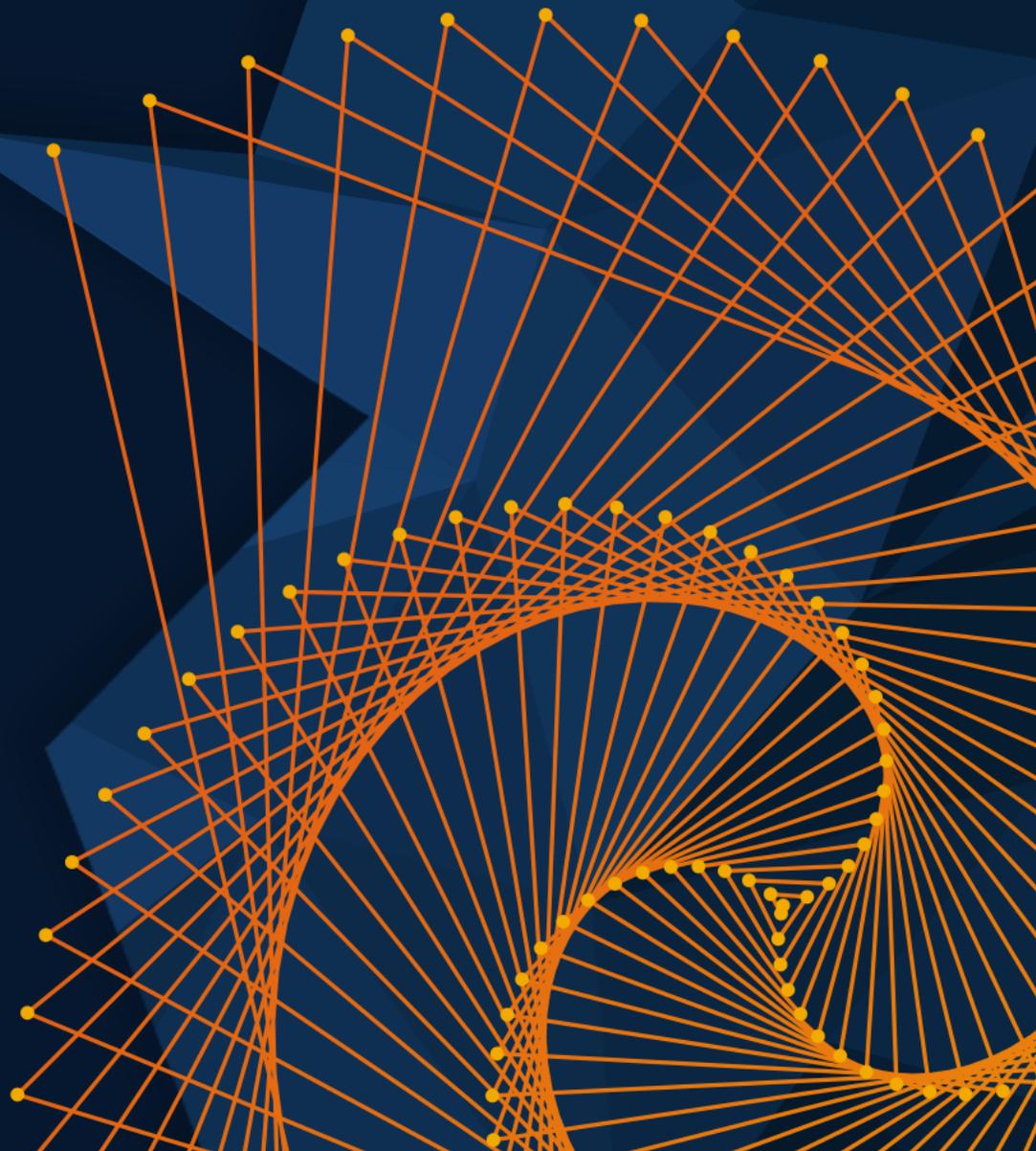
5月28日, 2024 | 北京

增量学习模型及其在多场景下的部署

马文辉, MathWorks 中国



MATLAB EXPO



增量学习

- 概念

增量学习 – Incremental Learning, 也可以称为在线学习(Online Learning)是机器学习的一个分支, 它涉及从数据流中进行连续的学习。

- 特征

与传统的机器学习相比, 增量学习具有以下特征:

1. 增量学习环境中进行模型验证和超参数调优是困难的, 甚至是不可能的
2. 模型可以冷启动, 即对数据集知之甚少的时候可以上线模型
3. 模型可以从实时数据中学习模型参数, 并进行预测
4. 模型能够自适应数据的变化

机器学习的数据集假设

1. 训练数据集是静态的

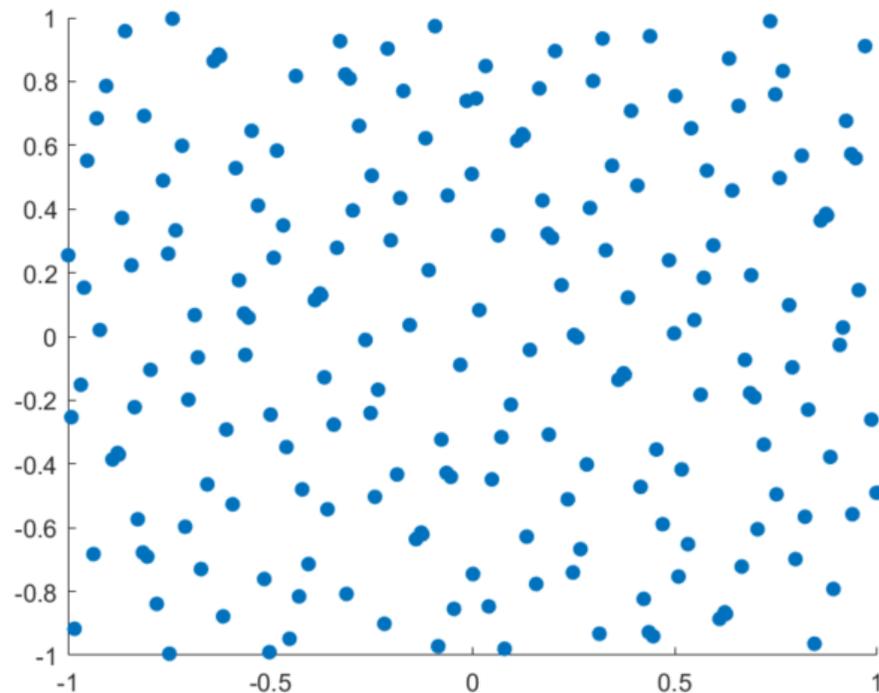
训练数据集随着时间的推移，其数据分布是不会发生变化的。

2. 独立同分布

数据集中的所有样本都是独立的，且来自相同的分布

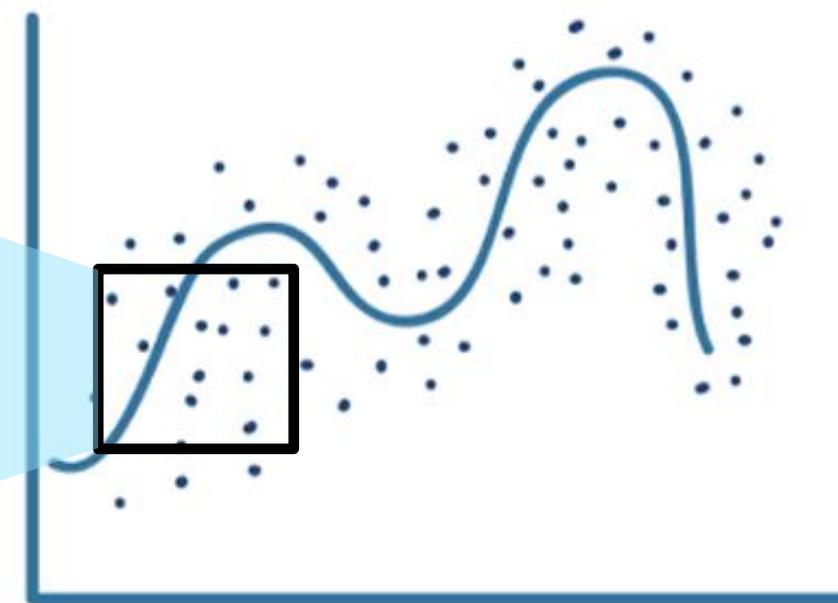
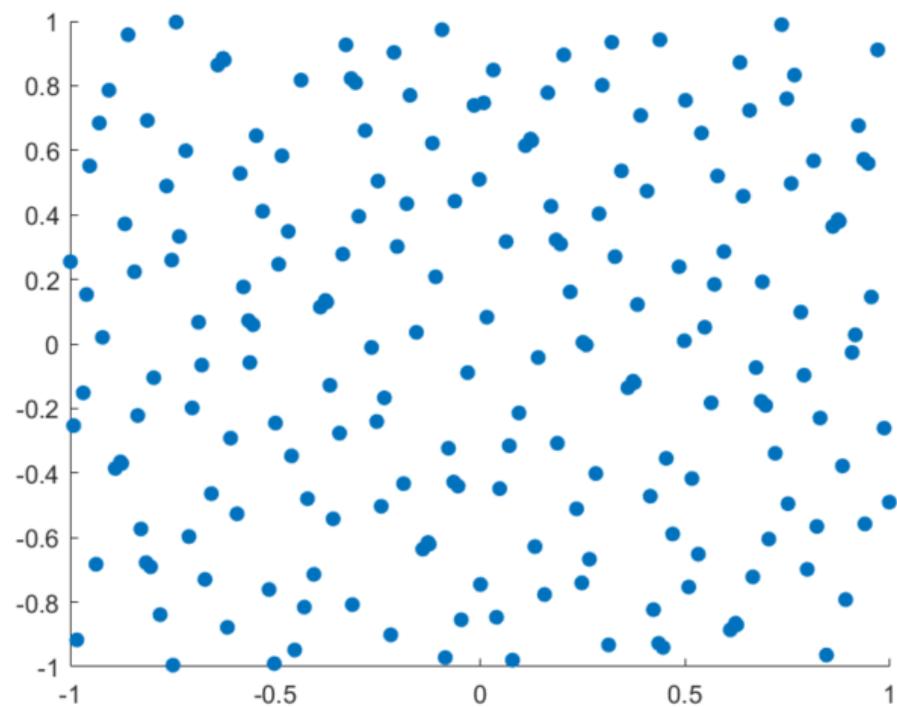
3. 分布一致性

训练数据集与预测数据集的分布是相同。



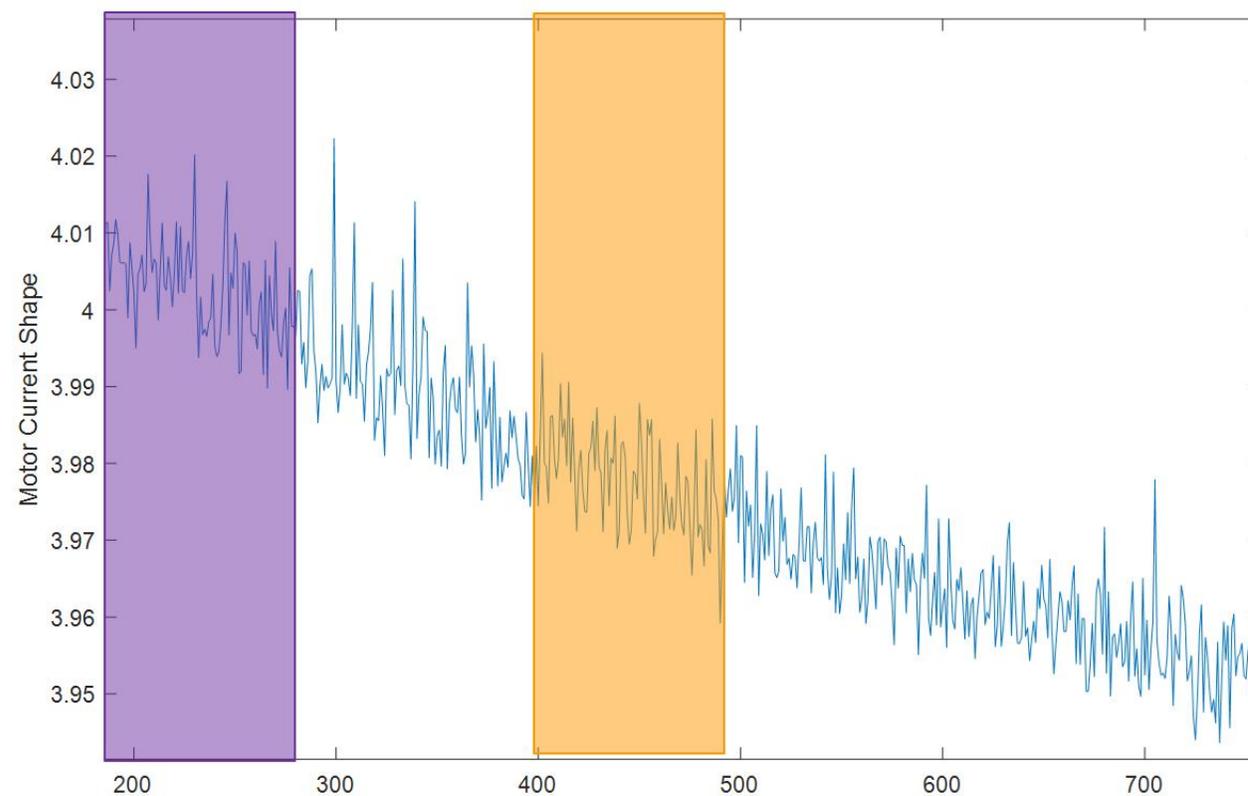
数据集假设存在的挑战

- 数据覆盖度问题



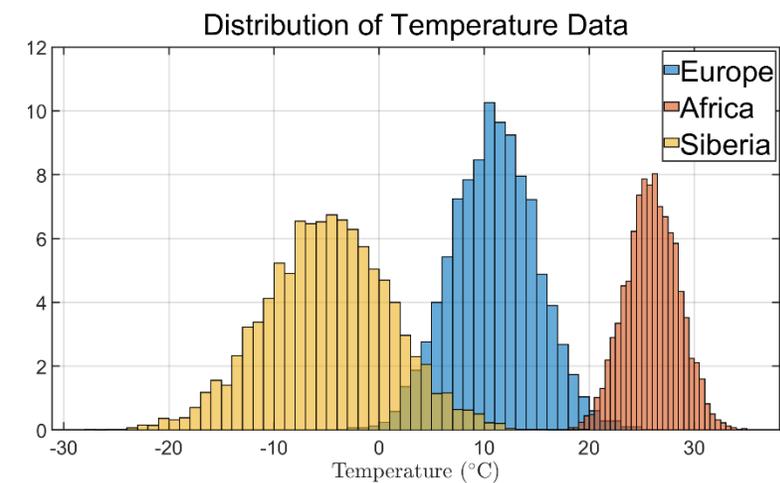
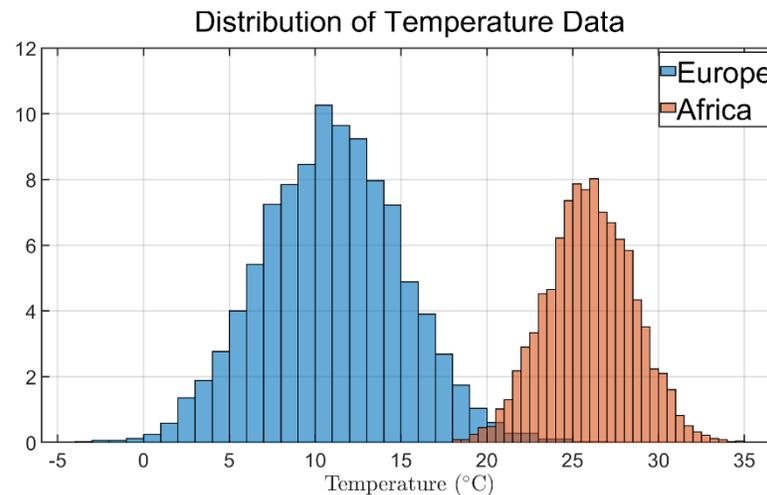
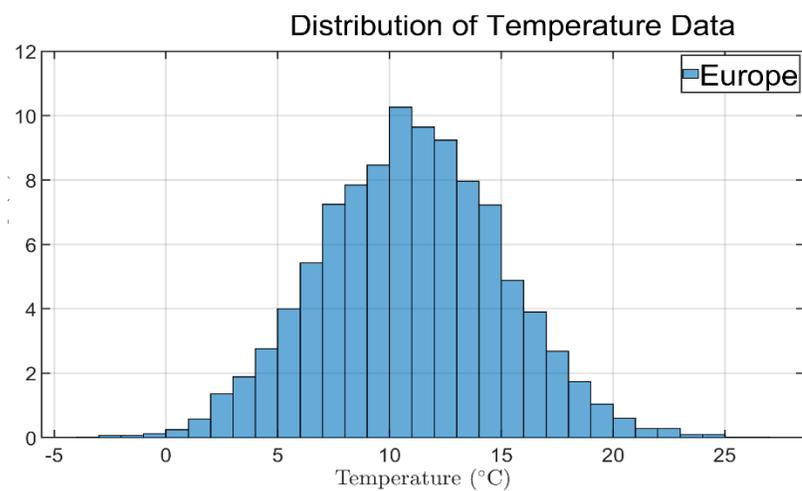
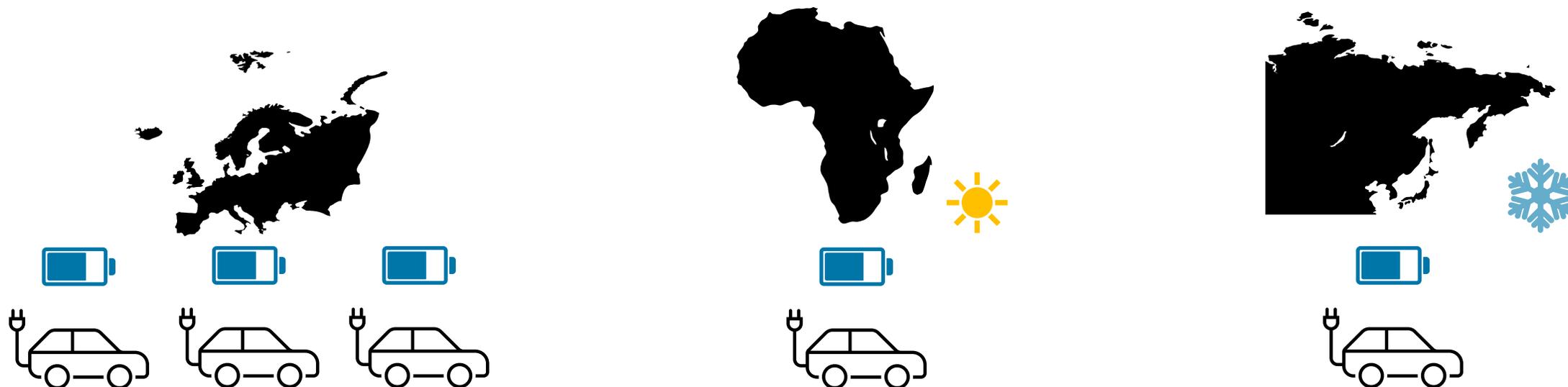
数据集假设存在的挑战

- 数据动态性导致数据分布的差异和变化



数据集假设存在的挑战

- 训练数据集与预测数据集分布的不一致性



机器学习需要适应数据集的变化

- 漂移

- 数据漂移：输入数据的统计特征或概率分布发生变化；
- 概念漂移：在机器学习领域，概念漂移特别指输入数据（特征）和输出数据（标签或数值）之间关系的变化。

- 漂移检测

- 在流式数据中，数据持续性的，因此检测漂移需要在线或增量的方法。
- 统计检测：通过对数据流中连续的数据块进行统计测试来检测分布的变化，可以帮助识别数据分布是否发生了显著变化；
- 性能指标检测：针对机器学习模型的概念漂移，检测模型的性能指标（如准确率，残差等），如有显著下降，表明发生概念漂移；

增量学习

增量学习方法是灵活的，它可以实时适应数据漂移，随着模型训练更多数据（更多的数据到来），其预测准确性接会逐步提高。相较于传统机器学习，增量学习方式发生以下变化：

1. 评估模型：利用真实数据集跟踪模型的性能。
2. 检测漂移：检测数据统计特征分布的变化显著性；
3. 训练模型：当真实标签可用，或当前模型出现退化时，对模型进行训练并更新参数；
4. 生成预测：利用最新模型进行预测。

MATLAB中增量学习

- 数据（或概念）漂移检测

<code>DriftDetectionMethod</code>	Incremental drift detector that utilizes Drift Detection Method (DDM) (自 R2022a 起)
<code>HoeffdingDriftDetectionMethod</code>	Incremental concept drift detector that utilizes Hoeffding's Bounds Drift Detection Method (HDDM) (自 R2022a 起)

- `DriftDetectionMethod` (DDM)

DDM方法主要关注于错误率的变化。其基于基于假设：

1. 当模型稳定时，模型的错误率应该相对稳定，
2. 而当模型发生变化时，模型的错误率会增加。
3. DDM通过监控在数据流的连续样本上的模型的错误率，来检测概念漂移。

MATLAB中增量学习

- 数据（或概念）漂移检测

`DriftDetectionMethod`

Incremental drift detector that utilizes Drift Detection Method (DDM) (自 R2022a 起)

`HoeffdingDriftDetectionMethod`

Incremental concept drift detector that utilizes Hoeffding's Bounds Drift Detection Method (HDDM) (自 R2022a 起)

➤ HoeffdingDriftDetectionMethod (HDDM)

HDDM方法是针对统计量，基于Hoeffding's Bound计算预期的波动范围。

1. 以一定的置信水平设置波动区间
2. 如果数据流中新到达的数据的统计量与之前的统计量的差值超出了这个波动范围，那么就可以判断出现了概念漂移。

MATLAB中增量学习

➤ 漂移检测统计理论保证：

1. 只要样本的类分布是静态的，学习算法的错误率（会随着样本数量的增加（ i 的增加）而减少。
2. 大数定律：随着试验次数的增加，样本平均值会趋近于期望值。
3. 中心极限定理：当样本量足够大时，独立同分布的随机变量的平均值将近似呈正态分布。这意味着：在静态环境中，随着样本量的增加，可以更准确地估计算法的性能（错误率）。

MATLAB中增量学习

- 模型构建 - 增量学习模型

Model Object	Objective	Model Type
<code>incrementalClassificationECOC</code> R2022a	Multiclass classification	Error-correcting output codes (ECOC) model with binary learners
<code>incrementalClassificationKernel</code> R2022a	Binary classification	Linear support vector machine (SVM) and logistic regression with Gaussian kernels
<code>incrementalClassificationLinear</code> R2020b	Binary classification	Linear SVM and logistic regression
<code>incrementalClassificationNaiveBayes</code> R2021a	Multiclass classification	Naive Bayes with normal, multinomial, or multivariate multinomial predictor conditional distributions

Incrementally Fit and Track Performance

<code>fit</code>	Train kernel model for incremental learning (<i>Since R2022a</i>)
<code>updateMetrics</code>	Update performance metrics in kernel incremental learning model given new data (<i>Since R2022a</i>)
<code>updateMetricsAndFit</code>	Update performance metrics in kernel incremental learning model given new data and train model (<i>Since R2022a</i>)

<https://www.mathworks.com/help/releases/R2023b/stats/incremental-learning-overview.html>

MATLAB中增量学习

- 模型构建 - 将机器学习模型转换为增量学习模型

Traditionally Trained Convertible Model Object	Conversion Function	Model Object for Incremental Learning
ClassificationECOC and CompactClassificationECOC	incrementalLearner	incrementalClassificationECOC
ClassificationKernel	incrementalLearner	incrementalClassificationKernel
ClassificationSVM and CompactClassificationSVM	incrementalLearner	incrementalClassificationLinear
ClassificationLinear	incrementalLearner	incrementalClassificationLinear
ClassificationNaiveBayes	incrementalLearner	incrementalClassificationNaiveBayes
RegressionKernel	incrementalLearner	incrementalRegressionKernel
RegressionSVM and CompactRegressionSVM	incrementalLearner	incrementalRegressionLinear
RegressionLinear	incrementalLearner	incrementalRegressionLinear

R2022a

MATLAB中增量学习

- 带有漂移检测的增量学习模型

`incrementalDriftAwareLearner`

R2022b

增量学习模型

漂移检测方法

Classification	Regression
<code>incrementalClassificationKernel</code>	<code>incrementalRegressionKernel</code>
<code>incrementalClassificationLinear</code>	<code>incrementalRegressionLinear</code>
<code>incrementalClassificationECOC</code>	
<code>incrementalClassificationNaiveBayes</code>	

`DriftDetectionMethod`

`HoeffdingDriftDetectionMethod`

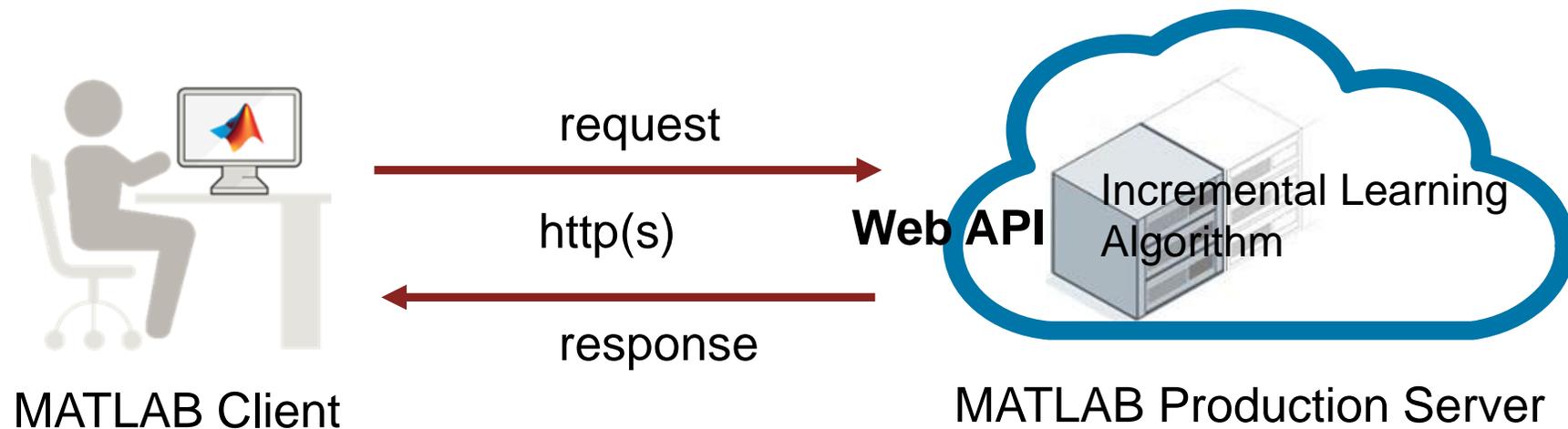
<https://www.mathworks.com/help/releases/R2023b/stats/incrementaldriftawarelearner.html>

Demos

- 增量学习模型的云端部署
- 利用Simulink搭建增量学习模型并部署

增量学习模型的云端部署

- 将增量学习模型部署到MATLAB Production Server上，并提供Web API的访问



MATLAB Production Server

Client applications

Native client Application

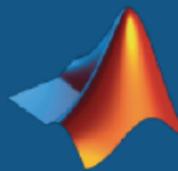
- Java
- .Net
- C/C++
- Python

MWHttpClient
object

Web/Mobile Application

Any language that can make
REST calls

Server


MATLAB
Production Server

Parameterized
return values

HTTP(S)

By default
over port
9910
(changeable)

HTTP(S)

JSON input/output
payload

增量学习模型的嵌入式部署

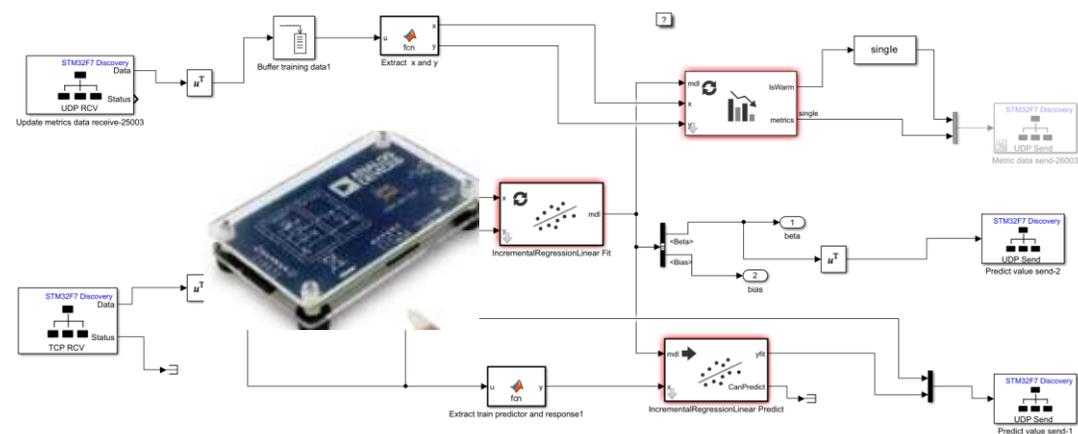
- 利用Simulink搭建增量学习模型仿真和代码生成



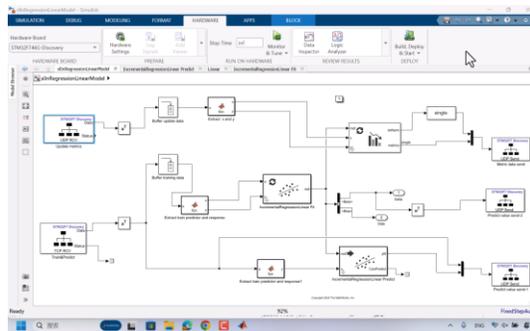
MATLAB Client



TCP/UDP



Simulink 仿真



MATLAB EXPO

谢谢



© 2024 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See [mathworks.com/trademarks](https://www.mathworks.com/trademarks) for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

