**软件学院2021年机器学习课程编程作业1**

1. **[时间序列预测]**

使用以下3个数据集: 商品销量数据集(commodity.txt)，用电量数据集(electricity.txt)和交通流量数据集(traffic.csv)，数据读取方式可参考load\_data.ipynb。这3个数据集均为时间序列数据集，记录了多个商品/用户/传感器在一段时间内销量/用电量/读数的变化；数据格式为的矩阵，其中，T为时间序列长度，N为商品/用户/传感器的数量，时间序列的采样间隔是固定的，分别为天/小时/5分钟。本题需要选择至少3种回归模型，实现根据时间序列的历史信息，预测下一时刻的值。具体要求如下:

1. **对时间序列数据做预处理**

当前的序列数据无法直接用于机器学习模型的训练与预测，需要通过预处理将其转化为的形式，其中，y为t时刻时间序列的值，为输入特征，从这一时段的时间序列X中提取获得，可以包括时间序列的历史值、最大值、最小值等信息。

使用宽度为的滑动窗口生成多组，如图1所示:

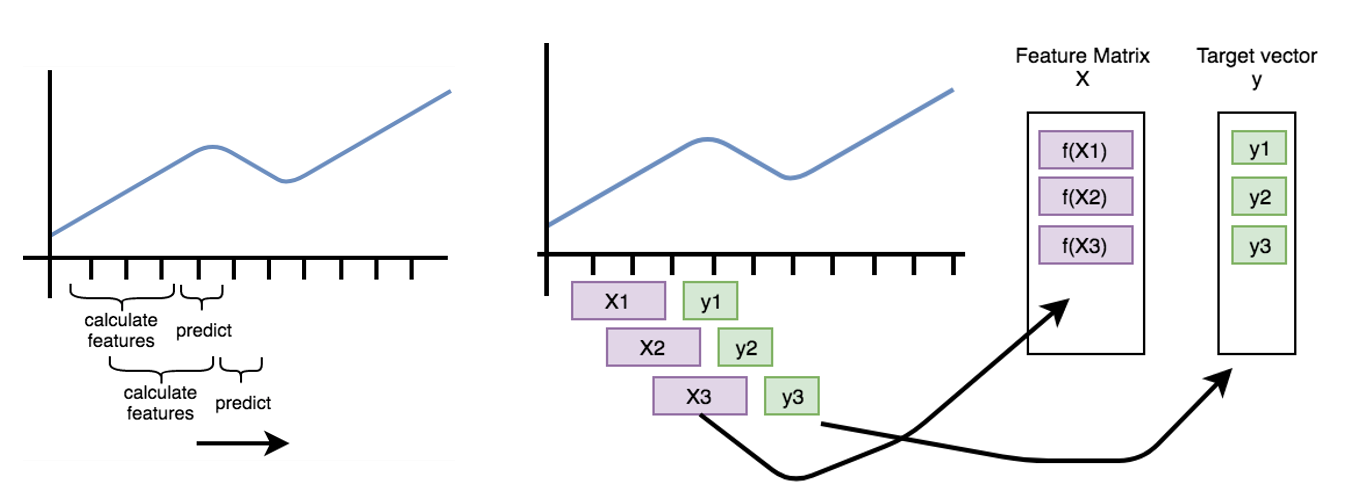


图1：数据预处理

最终获得条处理好的数据样本，将其按时间顺序排列，选择后条作为测试数据，其余用于训练和验证。

1. **模型训练与测试**

选择至少3种回归模型在训练集上训练，通过验证集调试模型参数，最后在测试集上测试，使用RMSE作为误差评价指标，比较多个模型的性能。

1. **(可选) 交通数据集的路网信息**

对于交通数据，不同路段的时间序列数据存在复杂的空间依赖关系，例如，路网距离较近的路段的交通流量数据，往往存在一定关联。能否将路网拓扑信息作为辅助信息应用于交通数据的预测呢? 可以尝试相关的模型或在其他模型的基础上做一定的改进。

W.csv是traffic数据集对应的路网数据，traffic数据集包含228个路段的交通流量信息，W.csv是一个的矩阵W，是根据不同路段的路网距离计算获得，计算方法为:

1. **[Logistic回归]**

使用height\_weight\_gender.csv数据集，实现Logistic回归算法，将height和weight作为输入特征，预测gender信息。

可参考template.ipynb, 补全相应的代码，完成以下任务:

1. 通过线性回归完成分类任务(可参考lab2的代码), 并将分类结果可视化，画出样本点和决策边界，如图2所示;
2. 通过Logistic回归完成分类任务(使用梯度下降法)，并将分类结果可视化；
3. 通过Logistic回归完成分类任务(使用牛顿法)，并将分类结果可视化；
4. 为b)中的Logistic回归方法增加正则项, 并查看随着增大，的变化情况。

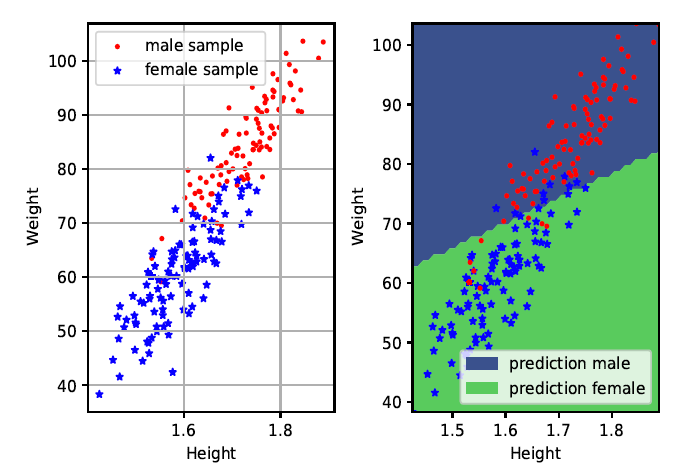


图2：分类结果可视化示例

**其他说明：**

提交日期: 2021年5月3日 23:59, 提交至canvas的作业->project1，作业命名为: 组号\_project1.zip，提交内容包括:

1. 题目1的代码和实验报告(包括对数据预处理、特征选择、用到的模型的说明，以及不同模型的性能比较与分析讨论)；
2. 题目2的代码和实验报告(包括预测结果的可视化和说明)。

**注: 不要提交数据集！**