**Week 4-6：大型数据集的算法可行性验证**

本阶段以及后续数个任务都将基于NYC Taxi and Limousine Commission提供的TLC行程数据及相关数据。

在进行分析之前，我们需要了解NYC的出租车体系。在NYC，所有类型的出租车辆都必须由TLC进行管辖，并且将行程数据提交给该机构进行分析、调度、处理。

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars and luxury limousines), commuter vans, and paratransit vehicles. The Commission's Board consists of nine members, eight of whom are unsalaried Commissioners. The salaried Chair/ Commissioner presides over regularly scheduled public commission meetings and is the head of the agency, which maintains a staff of approximately 600 TLC employees.

Over 200,000 TLC licensees complete approximately 1,000,000 trips each day. To operate for hire, drivers must first undergo a background check, have a safe driving record, and complete 24 hours of driver training. TLC-licensed vehicles are inspected for safety and emissions at TLC's Woodside Inspection Facility.

NYC TLC选择将这些Trip Record开放给社会进行分析和研究。在NYC，出租车的种类有以下几种：

* Yellow taxis are traditionally hailed by signaling to a driver who is on duty and seeking a passenger (street hail), but now they may also be hailed using an e-hail app like Curb or Arro. Yellow taxis are the only vehicles permitted to respond to a street hail from a passenger in all five boroughs.
* Green taxis, also known as boro taxis and street-hail liveries, were introduced in August of 2013 to improve taxi service and availability in the boroughs. Green taxis may respond to street hails, but only in the areas indicated in green on the map (i.e. above W 110 St/E 96th St in Manhattan and in the boroughs).
* FHV data includes trip data from high-volume for-hire vehicle bases (bases for companies dispatching 10,000+ trip per day, meaning Uber, Lyft, Via, and Juno), community livery bases, luxury limousine bases, and black car bases.

Uber作为一个提供网约车服务的大型企业，其行程数据被囊括在High Volume For-Hire Vehicle Trip Records中。

本节任务需要成员团队完成在大型数据集上的上游机器学习任务算法可靠性验证，要求团队A完成以下两个算法的可行性探究。

**任务1（3天）**

第一阶段，我们提供TLC在2021年全年的HVFHV行程数据、NYC taxi zone ID的对应表和边界图、Python使用手册、数据字典，以及NYC的日度气候数据。

你需要接触到python的地理绘图：<https://pypi.org/project/folium/>

第一阶段，我们需要对一月份的数据进行可视化，根据fhvhv\_tripdata\_2021-01数据，以6:00~18:00作为日间，分别生成：

1. 整个月的起点热度图
2. 整个月的到达点热度图
3. 整个月日间的起点、到达点热度图
4. 整个月夜间的起点、到达点热度图
5. 尝试以每个zone的中心点作为标定点，并且随机偏移以避免线路完全重叠，生成每趟行程的直线polyline（用户详细的上落客地点为隐私，TLC无权收集），以此能够可视化地观察行程趋势

并且通过图片分析上述内容中你的发现。

通过肉眼和人工的可视化，为后期算法辅助的数据分析铺垫基础。

**任务2（5天）**

从这节任务开始，我们假设你已经具备基本的机器学习基础，即掌握了KNN、CART、Bayes、LR、SVM、RF、Boosting等基础算法的原理和使用，我们将会在后续的内容中使用机器学习的算法辅助进行业务分析。

如果你还没有系统性地学习过机器学习，推荐根据以下系列教程进行入门：

<https://ailearning.apachecn.org/#/docs/ml/1>

这一个任务将会是监督学习的回归：旨在通过一月份的数据结合当日气候——将天气数据与tripdata进行连接，构建一个以日为单位，分时的订单客流高峰预测器，以预测每个zip code区域在当时的**预计去程客流**。

如果数据量不足以完成265个区域的训练，可以加入更多的月份数据。

分时预测器能够对宏观运营的调度起到建议和优化，并且能够让司机的行程利润最大化，同时避免不必要的拥堵。

在该回归任务中，Pickup\_datetime和PULocationID将会是最核心的自变量，其余数据则需要按工程需求适当舍去以降低多余的算法性能消耗。

我们推荐使用基本的Lasso回归选出基本的高相关度特征后，再将选择出的特征送入更加高级的算法进行训练，比如：CART→GBDT→XGBoost或者RNN→LSTM→GRU。

你需要在调优获取最优的预测结果后，报告其优势、算法性能消耗，并且横向对比所有你使用过的算法的预测表现，最后使用全部数据训练出最终能够投入下游使用的模型。

**任务3（7天）**

除了关注宏观的系统调度之外，我们还计划从用户个体特征来优化利润。

在NYC，Uber的计价方案包含以下组成：

* A base rate
* Tolls and surcharges
* Dynamic pricing
* A booking or marketplace fee
* Route based adjustments
* Applicable promotions
* Pool discounts

除了起价和路程费、燃油附加费以外，Uber Rides同时加入了一个Dynamic Pricing项，用来动态调整价格，该项目也是盈利的关键点之一。

在第16天以及第19天的任务1中，我们已经简要探讨了行程费用跟起终点、时间、气候等方面的大致联系，现在我们需要秉承这个思路，挖掘当中的经济价值。

在这环节，你需要完成一个聚类任务：通过聚类区分出高价值客户，即判断什么类型的订单用户愿意为之付出更高的价格，Find out which order was Urgent。

这一任务甚至与订单起落地点都有相当密切的关联，通过聚类算法寻找出“紧急订单/高价值订单”。我们能够在行程后为这样的高价值订单适当增加Dynamic Pricing。

在该任务中，基础的K-means及其衍生算法的效果可能一般，但我们依旧推荐你从kmeans开始练手，探寻基础的订单画像。

对于业务级别的模型，你需要尝试DBSCAN、凝聚层次聚类，有条件尝试使用Deep Cluster算法。

该任务有相当多的细节可供探究，比如如何确定该任务的Loss计算？数据集中的价格是已经发生过的订单，如何探寻该订单如果增加收费乘客是否愿意接受？真实的模型推理过程中，价格既不是输出，也不是输入，那么它在这个任务中究竟起什么样的作用？