20240819-0821

最近对自己的学习状态不太满意，被动学习取代了主动学习，经过反思决定开始写三天一篇的生活学习日志，会包括：

1. time logger:每天精确到分钟的动向记录
2. 学习日志：记录每天学习的笔记，进展
3. 生活记录和一些想法

0819

---

time logger
0530 up
0540 默写
0554 meditation
0630 原则
0700 gym
0755 eat
0808 数据分析课
0850 sleep
0921 数据分析课
1022 song
1045 数据分析课
1200 eat，和妈辩论
1330 podcast
1350 数据分析
1530 写作
1600 机器学习课
1630 开手机，找旅游计划
1800 关手机，吃饭
1912 singing practice
2040 reading
2140 sleep

0820
TIME LOGGER

0550 up
0600 meditation
0622 默写
0700 gym
0800 eat，youtube（黑神话悟空测评，中国新闻），shower
0840 数据分析
0940 relax，接电话
1020 pandas
1200 足球比赛集锦
1230 吃饭
1250 数据分析
1320 ml西瓜书
1440 relax
1500 ml
1550 writing

0530 up
0535 medi 默写
0610 reading
0700 gym
0800 数据分析
0813 分析
0829 数分
0940 relax
1000 数分
1100 闲逛
1200 relax
1400 房子， travel plan
1540 writing

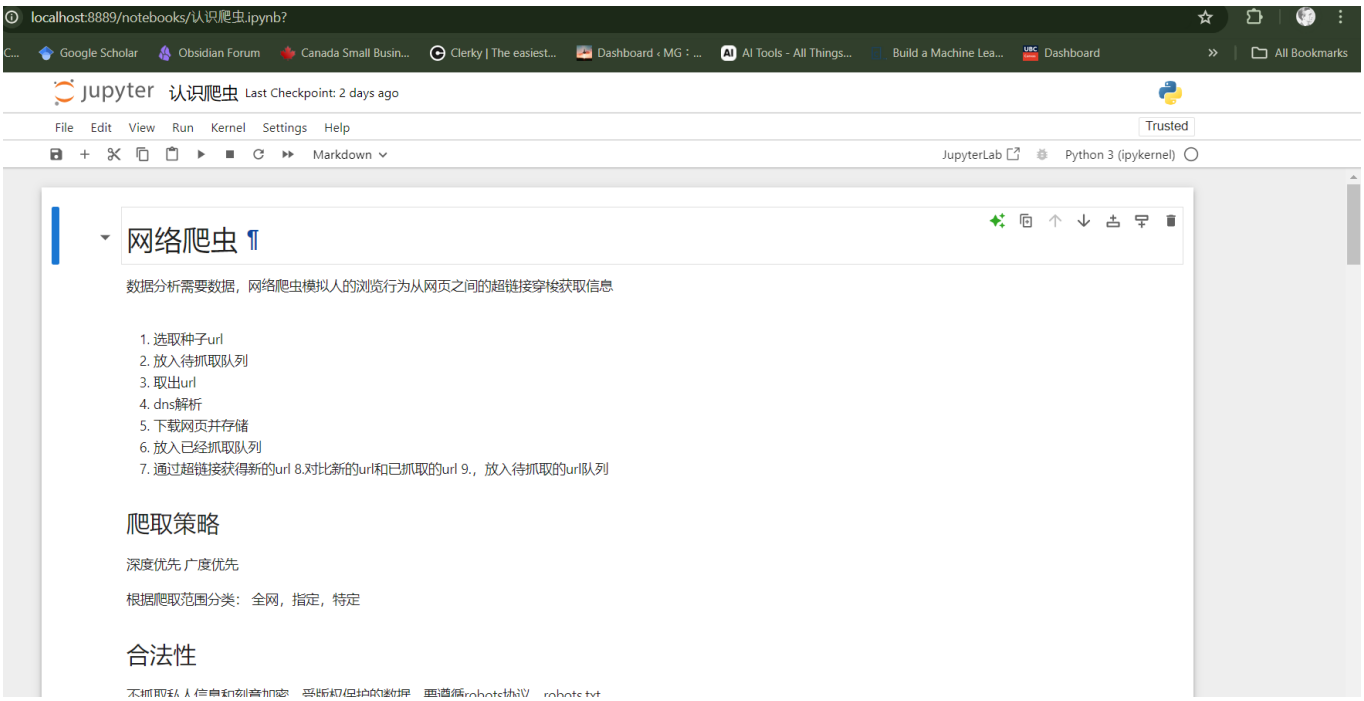
太快地把学校的宿舍租出去了，其实问的人很多，可以抬高价格的
8.18号基本上确定了开学之后的几位常驻室友，后面要把大家拉一个群然后起一个酷酷的房屋名，裂空座（?
开学之后有好多事情可以做呀，比如自媒体方面，我买了一副周易算命的牌，这个可以录一期给外国人算命的节目，二次穷游，也可以来一次节目！其他节目内容包括我的想法记录，社会实验，学校“名人”采访，还可以有之前的心理学性格报告

房子的事情终于告一段落了，最后确定了erica，小蔺，小王，小胡和我组成开学五人长租阵容，对这个结果还算满意的，我买了三国杀象棋什么的，开学的生活一定很丰富！

这两天家里在计划出去旅游两三天的事情。我发现我的想法很复杂，一方面觉得去旅游一两天在价值上是值得的，另一方面又对牺牲两天学习时间很抵触，不过我喜欢这个矛盾的心理。

这几天学习状态还是不错的，不过到了周三的中午开始感觉身体很疲惫，这也和平常娱乐活动较少有关，先暂且不做调整，等到了温哥华试验一段时间再看。

学习日志



机器学习（西瓜书）

0820
第一章

基本术语

数据集：记录的集合
示例/样本“单个记录
属性/特征：反应某方面的性质
属性空间/样本空间/输入空间：由属性的取值范围组成的坐标轴组成的空间
特征向量：属性空间的某个点的坐标向量

学习就是从数据中得到模型，训练用的是单个训练样本聚集而成的训练集，学到的潜在规律是假设，要和真相对比。监督学习需要标记，也就是结果，瓜到底是好是坏。

有的时候解决的是离散问题，叫做分类。如果是二分类，要建立x对y的映射关系，y的取值便是{-1,1}or{0,1},多任务则|y|>2
有的时候面对的是连续的值，这种任务叫做回归。y=R

学习过后使用模型在测试样本上测试。

分类和回归主要是监督学习，无监督学习则主要面对聚类的问题，为的是发现数据集本身的规律，无需标记。

我们的目标是把学到的用到预测新样本上，这叫做泛化能力，意味着模型可以适应很大的样本空间

假设空间

通过样本学习是一种inductive reasoning
我们可以把学习理解为在所有假设组成的空间里搜索与训练集匹配的过程。假设空间的大小：（x1+1）(x2+1)(..xn+1)+1.这是要把通配符和”没有好瓜“考虑进来，或者说是要考虑那个在视野之外的0.
可能会有多个假设符合训练集，组成一个版本空间

归纳偏好

那如何选择假设呢？这取决于学习算法的偏好。
可以偏好尽可能special，or general，or simple(occam's razor)
不同的偏好选择不同的算法，适合不同的具体情境。nfl原理就是在说，一个算法在某类问题表现更好，就一定在某类问题表现较差。因此一定不能脱离具体问题。

2. 模型评估与选择

我们的目标是泛化能力强，那我们怎么在没有无限unseen instances情况下进行评估呢

经验误差，过拟合

错题数/总数=错误率
在训练集上出现的叫经验误差，这个是我们可以控制的
在新样本上出现的叫泛化误差
学习力太强导致过度拟合（把个例当作规律），不足导致欠拟合
这就在减少经验误差和防止过度拟合上有了矛盾：学的越猛，经验误差越低，过度拟合越多。某种程度上说，各类算法都在尝试给出找到"平衡点"，这也是学习各类算法时一个思考的点。

评估方法

要解决三大核心问题：评估方法，性能度量，比较检测
评估方法要回答怎么获得测试集。

测试集应该尽可能和训练集互斥

留出法

将数据集一分为二，用于训练和测试，

1. 切分时主义要分层采样以保证数据分布的一致
2. 多次进行。
3. 测试集在占比0.2-0.33之间

交叉验证

把数据集D分为k个子集，每次用一个子集测试，k-1个子集训练，进行k次，一般也会进行多次重新切割，返回均值