**《21天入门机器学习》**

**训练营实践项目说明文档**

**如何参与实践项目：**

1.    请按下方具体实践项目要求完成作业；

2.    将模型训练结果、思路或心得、疑问提交在打卡小程序里对应的作业任务中（文字、语音均可）；

3.    可同时附带自己的 GitHub 地址链接来提交训练和测试代码（注意请在 readme 中写清如何运行，我们不负责调试哦~）；

4.    参与实践项目有机会赢取优秀学员奖，获得 99 元助学返现！

**如有疑问，可以微信咨询训练营小助手（gitchat2018）**

**【实践项目1】股价预测器**

**数据源：**

msft\_stockprices\_dataset.csv

msft\_stockprices\_dataset.csv 记录了从2014年9月9日到2018年9月10日的微软股价，包括每日最高价（High Price），最低价（Low Price），开盘价（Open Price），收盘价（Close Price）和交易量（Volume）

**要求：**

用本题提供的数据集训练一个模型，用来预测微软公司股票的收盘价。

**验证标准：**

预测准确率=（预测正确样本数）/（总测试样本数）\* 100%

可以人工指定一个 ErrorTolerance（一般是10%或者5%），

当 |预测值-真实值| / 真实值 <= ErrorTolerance 时，

我们认为预测正确，否则为预测错误。

**提示：**

1.    想一想这是一个什么样的任务，是分类任务还是回归任务？

2.    确定了任务类型，再来考虑用哪个模型好？

3.    如果在开始的时候不能确定哪种类型模型效果更佳，可以尝试多种模型。

4.    现在总共一份数据，是分成三份（训练集、验证集、测试集），还是只分成两份？按什么比例分？

**提高拓展：**

考虑一下如果用HMM预测股票收盘价该怎么做？

**【实践项目2】简历过滤器**

**数据源：**

employees\_dataset.csv

employees\_dataset.csv 为了某公司部分员工的简历信息，包含了：学历（degree）, 毕业院校（education）, 技能（skills）, 曾经工作过的公司（working\_experience）,和当前职位（position）几部分信息。

现在的职位有三种：dev（开发工程师），qa（测试工程师）和 manager（经理）

**要求：**

用本题提供的数据集训练一个模型，用来过滤求职者的简历；

用来判断求职者是：**不能被录取**、适合当**开发工程师**、适合当**测试工程师**，还是适合当**经理**

**验证标准：**精准率，召回率和 F1Score

**提示：**

1.    想一想这是一个什么样的任务，是分类任务还是回归任务？

2.    确定了任务类型，再来考虑用哪个模型好？

3.    如果在开始的时候不能确定哪种类型模型效果更佳，可以尝试多种模型。

4.    数据集分几份？怎么切分？

5.    选哪些 column 做特征？

6.    对于非数值类型特征，应该如何转化为数值型？

7.    本题中有几个 column（education，skills，working\_expereince）中包含多项内容，应该怎么处理？

8.    education 一列中列举的是具体原校名，可否考虑用985，211， normal 之类的枚举来代替具体院校？

**必须思考：**

要知道，本题的所有数据中是没有负例的，没有不能被录取的人的简历信息。在你开始指定 label 取值时，你是仅仅指定了 dev，qa 和 mananger 三种，还是也有一个 nohire？如果有 nohire，没有对应数据，怎么训练？这是现实中常见的情况，考虑一下在实践中如何解决。

**【实践项目3】电影评分器**

**数据源：**

movies\_dataset.csv

movies\_dtaset.csv 包含了某电影网站上的电影信息和评分信息，具体分为：中文电影名（chinese\_name），英文电影名（english\_name），导演（director），主演（starring），类型（type），发行日期（release\_date），评分（rate），投票数（votes），发行地区（region），播放次数（runtime），分级（certification），语言（language），发行公司（company）几部分信息。

**NOTE：**

本数据集和前两分不同，更接近真实场景，其中有一部分数据column有错位，还有一部分records本身就是无用信息，而且是这三份作业中唯一一份包含中文的数据集。

**要求：**

用本题提供的数据集训练一个模型，用来**预测**一部电影的评分。

**验证标准：**

预测准确率=（预测正确样本数）/（总测试样本数）\* 100%

**提示：**

1.    想一想这是一个什么样的任务，是分类任务还是回归任务？

2.    确定了任务类型，再来考虑用哪个模型好？

3.    如果在开始的时候不能确定哪种类型模型效果更佳，可以尝试多种模型。

4.    数据集分几份？怎么切分？

5.    本份数据需要预先**清洗！**

6.    选哪些column 做特征？

7.    对于非数值类型特征，应该如何转化为数值型？

8.    本题中有几个 column 中包含多项内容，应该怎么处理？

**延申扩展：**

如果我不是要预测打分，而只是想预测一部电影是“很不好看”，“一般”，还是“好看”，

那么，利用现有数据，应该怎么训练模型？

**如有使用疑问，可以微信咨询训练营小助手（gitchat2018）**