参考书Hands on Machine Learning with Scikit Learn and Tensorflow pdf P56开始

PPT P11

1. Frame the Problem
2. Ask your boss what exactly is the business objective, building a model is probably not the end goal. V

在这里，我们的目标就是，对于输入的样本（具有所有的所需features或者缺失了某些features），尽量准确的预测这个样本是否患有心血管疾病（患有的风险）

1. what the current solution looks like (if any如果有的话)对于这个问题的现有解决方法. 比如房价在没有模型预测的时候就是专家预测。

在这里，是否患有心血管疾病的现有解决方案是通过去医院进行就诊，经过专业的医学检测与诊断来得出（数据集里面的label就是这样得到的）。我们这里可以通过一些较为易得的健康参数来预测一个人是否有心血管疾病。项目的意义

1. start designing your system

First, you need to frame the problem: is it supervised, unsupervised, or Reinforcement Learning? Is it a classification task, a regression task, or something else? Should you use batch learning(普通的机器学习，一堆数据进行训练得到weight) or online learning(update the weights after every input sample) techniques?

If the data was huge, you could either split your batch learning work across multiple servers (using the MapReduce technique, as we will see later), or you could use an online learning technique instead.

注：使用batch还是online的依据there is no continuous flow of data coming in the system, there is no particular need to adjust to changing data rapidly, and the data is small enough to fit in memory, so plain batch learning should do just fine.（三点）

这是一个supervised learning、classification task。我们可以都试试batch和online learning

1. Select a performance measure

由于这是一个classification problem，所以直接计算预测准确率即可。

Is the performance measure aligned with the business objective?

Yes.

1. What would be the minimum performance needed to reach the business objective?

没有明确要求。如果准确率和可靠性很高，则可以作为临床诊断；如果准确率和可靠性不那么高，可以作为风险评估。

1. What are comparable problems? Can you reuse experience or tools?

这是一个典型的classification problem，之前学过的聚类之类的模型都可以用。

1. Is human expertise available？

Yes。我们可以通过常识，很明显的判断出有些样本的某个feature数值明显超出正常的范畴。

1. How would you solve the problem manually?

首先进行数据的清洗，再应用各种模型，选取表现最好的模型。

1. List the assumptions you (or others) have made so far

血压抽烟喝酒可能对是否患有心血管疾病有较大影响。

1. Verify assumptions if possible

项目的意义：通过一些较为易得的健康参数来预测一个人是否有心血管疾病。

由于这个数据集采集的是一些比较宽泛的健康以及生活习惯数据，并不是非常专业的、有针对性的医学数据。因此，我们有这个数据集能够做的，只是：根据采集的数据判断这个人患心血管疾病的风险，而不能确切的预测这个人实际是否患有心血管疾病。

即使有些人及抽烟又喝酒又不爱运动，但是他就是没有心血管疾病，这种属于特例，不应在模型的考虑范围内。模型的最终目的是使得预测的准确率尽量高。

这个问题其实就是该数据features所提供的信息是否足够准确预测的问题。

1. Get the Data

P66

具体详见程序

1. 说明我们是从kaggle下的数据，该数据已经经过授权，
2. Create the Workplace V
3. Download the Data V
4. Convert the data to a format you can easily manipulate (without changing the data itself)
5. Take a quick look at the data structure V
6. Ensure sensitive information is deleted or protected (e.g., anonymized)

该数据集没有什么敏感信息

1. Check the size and type of data (time series, sample, geographical, etc.)

这一步好像也没有什么可做的

1. Create a test set （做EDA不能包括测试集）

Sample a test set, put it aside, and never look at it (no data snooping!)

使用哈希函数。哈希函数将incidences（样本）的identifier（比如id）转换成为哈希码，哈希码组合在一起就是哈希表。哈希表中的有些哈希码可能没有对应的incidence，有些一个哈希码可能对应多个incidence（冲突）。好的哈希函数应该使转化之后的incidence分布尽量均匀，冲突尽量少。

由于普通方法崩的原因就是测试集合可能不同，所以使用哈希函数，即使再次运行或者添加新的样本呢进来，由于哈希函数与每个样本的identifier都是固定的，所以得到的每个样本对应的哈希值是一直不变的，所以测试集里面的样本就一直是测试集样本。新的样本会根据哈希值以及之前制定的规则来决定是进训练集还是测试集。

PDF P76-77

在分测试集和训练集的时候，要注意：对于重要的特征，需要按照这个特征的类别进行分层抽样。比如血糖这个特征很重要，血糖这个特征有三个类别低、中、高，那么在分训练集和测试集的时候，在训练集（测试集）中血糖低、中、高这三类样本的比例应该保持与样本总体的血糖低中高的比例一致。

如果该重要的特征是连续特征（如收入），则要注意：对于该特征的分层（比如分成收入很低、低、中、高、很高五个等级），每层都需要have a sufficient number of instances，否则the estimate of the stratum’s importance may be biased，即不要分太多层，每层需要足够大（样本数足够多）。当有少部分样本偏离主要分布很远的时候，可以采用：1.对所有数值除以一个数，使样本分布更紧密；2.设定大于多少的值全部归为一类。

1. **Discover and Visualize the Data to Gain Insights（Explore the Data）**

P79

1. Visualizing geographical data V

More generally, our brains are very good at spotting patterns on pictures, but you may need to play around with visualization parameters to make the patterns stand out.

通过画scatter来看不同attributes之间的关系

1. Looking for correlations V

计算每个attributes之间的标准相关系数 standard correlation coefficient

缺点：P84相关系数仅测量线性相关（“如果x上升，则y通常上升/下降”）。 它可能会完全错过非线性关系（例如，“如果x接近零，则y通常会上升”）。 请注意，尽管底行的轴显然不是独立的，但底行的所有图如何具有等于零的相关系数：这些是非线性关系的示例。 另外，第二行显示了相关系数等于1或–1的示例； 请注意，这（相关系数）与斜率无关。 例如，以英寸为单位的高度与以英尺或纳米为单位的高度的相关系数为1。

1. Experimenting with Attribute Combinations V

Some attributes have a tail-heavy distribution(分布范围很广), so you may want to transform them (e.g., by computing their logarithm（取对数）).

通过对于attributes的组合，如BMI和MAP（平均动脉压），得到对target（即cardio）更高的相关系数

注：1.以上一二三并不需要全部做。（只是帮助你更好的理解这个项目数据）

2.以上过程是一个循环的过程，在做了一个prototype之后如果有其他更好的想法可以重复以上步骤进行更合理的处理。

四、Prepare the Data for Machine Learning Algorithms (EDA重点) P87

不要手动的做，要写成函数函数。（原因：方便下次调用。逐渐建立自己的函数库。）

1. 预处理数据前把feature和target(label) 分开
2. Data cleaning

\*dealing with missing data

1.扔掉对应样本数据

2.扔掉这个feature

3.补充失去的数据替换（中位数median，均值，零）；要将用来替换的数存下来，用来补充/替换test set以及实际预测中样本的缺失值。

应用：使用sklearn的imputer。

imputer = Imputer(strategy="median") # 设定填补的数据形式（中位数）

housing\_num = housing.drop("ocean\_proximity", axis=1) # 去除text feature

imputer.fit(housing\_num) # 学习各个feature的中位数

imputer.statistics # 将各个feature的中位数存在这

*housing\_num.median().values # 调看所有features的中位数（与上面对比）*

**X = imputer.transform(housing\_num) # 将中位数填充进缺失数据处**

注意：中位数仅能用于数值型feature，在填补前应先去除text feature。在实际应用模型时，由于无法确定新进来的样本数据是否是完整无空缺的，因此也要对进来的样本数据进行补充操作。

1. Handling Text and Categorical Attributes 处理文本与分类属性 P90

将文本变为数字。但是这种方法存在问题：0与1会比0与4更加相似（越接近就认为越相似）。

feature\_engineering\_for\_machine\_learningP93为了处理这种情况，使用one-hot encoding(为每个类别创建一个二进制属性，只有在对应位置为1，其他位置为0)，得到的是SciPy sparse matrix。当这个attribute的取值category很多的时候，生成的one-hot encoding很稀疏，存储那么多的0很浪费存储空间，可将其转换为（密集）NumPy array。

转换顺序：text categories to integer categories to one-hot vectors。可用LabelBinarizer一步到位。

应该要先做四，再做三

关于两个feature highly correlated及Multicollinearity（多重共线性）的问题阐述和解决方法：https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4

要做的事：

1. 去除异常值：boxplot、三倍标准差。

去除一些组合异常值，如身高很矮，但是体重异常的大

1. 把去除的值用中位数补上
2. Category attributes 转换为 onehotcode

算法：

1.用聚类求出两类的样本中心，计算测试样本对于两个样本中心的距离

二分类：样本离哪个聚类中心近

患病可能性：样本距离两个中心的距离以及两个中心之间的距离，这三个距离的某种关系得出。