机器学习数据预处理过程

数据类型

- 正常的数据：非连续、非目录
- 差值数据： 两个值得差，这个差是有意义的
- 有序的数： 两个值是不同的（我理解的是类似类别的数据）
- 比率数据：两个值得比值型数据，0也是有意义的

	Nominal	Ordinal	Interval	Ratio
Countable	✔	✔	✔	✔
Order defined		✔	✔	✔
Difference defined (addition, subtraction)			✔	✔
Zero defined (multiplication, division)				✔

	Nominal	Ordinal	Interval	Ratio
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔

	Nominal	Ordinal	Interval	Ratio
Counts / Distribution	✔	✔	✔	✔
Minimum, Maximum		✔	✔	✔
Range		✔	✔	✔
Percentiles		✔	✔	✔
Standard deviation, Variance			✔	✔

数据探索

- 变量识别：确定数据模型的输入输出，辨别输入输出是不是连续性变量还是类别变量，证明此问题是回归还是分类
- 单变量分析
  - 连续型数据：均值、中值、最大值、最小值、范围分析、四分位数、四分位数的间距、方差、标准偏差、峰度、直方图、Box-plot
  - 类别数据：频数、直方图
- 多变量分析
  - 找出两个变量的关系
  - 散点图
  - 相关热图
  - 双向表：我们可以利用两个变量的关系创建一个数目和百分比数目的双向表
  - 叠堆柱状图
  - 卡方检验：这个检验是为了推导出变量之间关系的统计意义
  - Z检验：一般用于大样本(即样本容量大于30)平均值差异性检验的方法，它是用标准正态分布的理论来推断差异发生的概率，从而比较两个平均数的差异是否显著。
  - 方差分析

特征清洗

- 缺失值：可以选择从包含缺失值的数据集中删除元素，或者将值归为一个值
- 特殊值：数据集中包含NaN 、NA等数据，你需要把这些纸填充，有列均值填充，模型填充，在大多情况下，可以将其和缺失值采用相同的处理
- 异常值：就相当于离群点，可以删除也可以不删除，这是非常灵活的
- 明显不一致的值：比如说：年龄不可能是负数、未满18岁不能获取驾照等等看起来明显不对的值

缺失值计算

- 前向填充：利用缺失值前面的值填充缺失值
- cold-deck：利用其他数据集的填充这个数据集（不是特别理解，像是用了聚类的方法进行缺失值的填充）
- 均值填充：利用非缺失的均值来填充缺失值，有一个好处就是均值不会发生变化
- 回归：利用其他值训练一个回归模型进行缺失值的预测，这是模型填充

Package	Numeric			Hot deck			kNN	Longitudinal		
	mean	ratio	reg.	rand	seq	pmm		int	lsimo	LS
Amelia[12]	·	·	·	·	·	✓	·	·	·	·
Baboon[19]	·	·	·	·	·	·	·	·	·	·
cat[1]	·	·	·	✓	·	·	·	·	·	·
deducorrect[13]	·	·	·	·	·	·	·	·	·	·
eiPH[14]	✓	·	·	·	·	·	·	·	·	·
ForImp	✓	·	·	·	·	·	✓[5]	·	·	·
Hmisc[15]	✓	·	·	✓	·	✓	·	·	·	·
imputation[17]	✓	·	✓[3]	·	·	·	✓	·	·	·
impute[16]	·	·	·	·	·	✓	✓	·	·	·
mi[11]	·	·	✓[4]	✓	·	✓	·	·	·	·
mice[10]	✓	·	✓[4]	✓	·	✓	·	·	·	·
mix[1]	·	·	·	·	·	·	·	·	·	·
norm[18]	·	·	·	·	·	·	·	·	·	·
robCompositions[21]	·	·	✓[4]	·	·	·	✓	·	·	·
rrcovNA[20]	·	·	·	·	·	·	·	·	·	·
StatMatch[8]	·	·	·	✓	·	·	✓	·	·	·
VIM[24]	·	·	✓[4]	✓	·	·	✓	·	·	·
yaImpute[6]	·	·	·	·	·	·	·	·	·	·
zoo[22]	✓	·	·	·	·	·	·	✓	✓	·

some libraries

特征工程

- 分解：对时间序列进行分解；例如2012-12-8 12:44:32，可以将其分解为类别特征，一天的小时、一天的部分等
- 离散化
  - 连续特征：典型的处理方式：把它用值/长度离散化成多少个部分，还可以采用前总数据的k%的（等价于频率）
  - 类别特征：对于类别特征可以采用联合特征，与其他列进行联合得到特征
- 再定义：对于不符合某些分布，但是你想让他服从某些分布需要采用log平滑或者其他方式进行数据的再定义
- 交叉特征：利用已有的特征创建新的特征，联合特征、类别联合特征等，这是最好的方式进行特征的扩充

特征编码

在数据挖掘领域中，所有的特征都是数值型的，因此我们会需要一些编码来对非数值型的数据进行编码

Human	Penguin	Octopus	Alien
1	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	1	0
0	0	0	1

One-Hot Encoding

Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

Label Encoding

使用One-hot编码时要确保所有的特征都是线性独立的

特征标准化或者比例缩放

由于原始数据的取值范围相差很大，在一些机器学习算法中，如果没有标准化，目标函数就不能正常工作。使用特征缩放的另一个原因是，梯度下降比不使用特征缩放的速度要快得多

- 归一化：这是一中将数据映射一个固定范围的方法，比如【0,1】获取【-1，1】

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- 方法
  - 标准化：特征标准化使得数据中的每个特征值都具有零均值(当减去分子中的均值)和单位方差

$$x' = \frac{x - \bar{x}}{\sigma}$$

  - 缩放到单位长度：缩放特征向量的分量，使完整的向量长度模为1

$$x' = \frac{x}{||x||}$$

特征选择

- 相互关系：训练特征之间应该不想关，然后训练特征和要预测的结果应该高度相关
  - 协方差：衡量两个随机变量变化多少

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- 降维
  - PCA：主成分分析：是一种静态的统计过程，将一组可能相关的观测值采用正交变换成一组不相关的过程。PCA把原先的n个特征用数目更少的m个特征取代，新特征是旧特征的线性组合，这些线性组合最大化样本方差，尽量使新的m个特征互不相关，从旧特征到新特征的映射捕获数据中的固有变异性。
    - 画出每一个特征的方差，选择最大的方差作为新特征
  - SVD：SVD是一个实数或复杂矩阵的因式分解，通过对极坐标分解的扩展，将一个正半正态矩阵的特征点的特征值推广到任何m * n矩阵。它在信号处理和统计方面有许多有用的应用

$$\mathbf{M} = \mathbf{U\Sigma V}^*$$

where

- $\mathbf{U}$ is an $m \times m$ unitary matrix (if $K = \mathbb{R}$, unitary matrices are orthogonal matrices),
- $\mathbf{\Sigma}$ is a diagonal $m \times n$ matrix with non-negative real numbers on the diagonal,
- $\mathbf{V}$ is an $n \times n$ unitary matrix over K, and
- $\mathbf{V}^*$ is the conjugate transpose of $\mathbf{V}$.

- 特征重要性
  - 过滤：过滤方法是一种经常使用的方法，主要就是利用某些方法计算特征列于目标列的值，是一种简单的判断特征重要性的方法，计算时间很快，鲁棒性较好，但是不能仅根据一个来判断特征的重要性
    - 相关性分析
    - 线性判别分析
    - 方差分析
    - 卡方检验
  - 封装：选择特征的子集进行评估
    - 前向选择
    - 后向消元
    - 回归特征消除
    - 遗传算法
  - 集成：结合了前两种特征选择算法的优点，感觉类似xgboost就是使用这个方法进行特征重要性的评估
    - Lasso 回归 L1进行特征重要性的评估，L1是一个稀疏解的过程，重要特征系数不为0，不重要特征系数为0，计算的是为x向量各个元素绝对值之和。服从拉普拉斯分布
    - Ridge 回归，L2进行特征重要性的选择，为x向量各个元素平方和的1/2次方，L2范数又称Euclidean范数或Frobenius范数，服从高斯分布

数据分割

- 训练集：例如：训练多层感知机参数，利用数据集找到最优的权重参数，当使用反向传播时候
- 测试集：用来训练的数据集：在最终的模型上进行预测测试集，用来评估分类器的好坏
- 验证集：一组仅用于评估训练好的分类器性能的示例：在多层感知机的例子中，我们使用验证机来寻找最优参数、隐藏单元、权重参数等
- 交叉验证：一组用来调整分类器参数的例子：在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。