这个文档是在干什么?

早在本科数模期间，我就觉得网上关于机器学习的方法五花八门，但却缺少一个总体性的框架，可以让我沿着机器学习的一般流程，掌握各种各样的方法。现在，立志成为一名数据科学家的我觉得总体的脉络图对于我后续的学习实在是至关重要，于是便写了这个文档。

更新日志

第一次更新 2023.5.25

1. 构建文档基本框架
2. 查阅文档，归纳基本概念与一般流程

有关基本概念

详情参见(29条消息) 机器学习：基本概念_机器学习的概念_燕双嘤的博客-CSDN博客

一般流程

详情参见(29条消息) 最全的机器学习模型训练全流程_kuokay的博客-CSDN博客
特征选择有关内容可以看(29条消息) 特征选择方法最全总结！_Datawhale的博客-CSDN博客
特征工程有关内容可以看(29条消息) 【特征工程】呕心之作——深度了解特征工程_CS正阳的博客-CSDN博客
概括而言：

流程	流程细节	流程方法	流程意义	指标与算法	意义
处理数据	探索性数据分析(Exploratory Data Analysis, EDA)	描述性统计	对数据集形成初步的印象	平均数	捕捉数据集分布的“核心”特征，容易受极端值影响。
				中位数	捕捉数据集分布的“核心特征”
				标准差/方差	捕捉数据集分布的离散情况

				皮尔逊相关系数/协方差	捕捉两个变量之间的线性相关性
		可视化分析		热力图	捕捉变量之间的相关性
				箱型图	捕捉数据集分布的分位数特征，也反映了数据集的离散情况
				折线图	捕捉数据的变化情况
				直方图	捕捉数据的分布情况
				饼图	捕捉不同类别的占比情况
				散点图与简单的回归直线	捕捉数据的分布情况
		数据整形		过滤	捕捉某一部分数据的特征
				分组	捕捉某一部分数据的特征
	数据清洗	处理缺失值	去除掉数据中的“杂质”：要么直接删了，要么找东西来代替	删除	若缺失值影响不大
				修正	用平均数、中位数、众数、插值或回归的方法处理
		处理异常值		盖帽原则：是否位于平均值三倍标准差外	识别异常值
				箱型图	
				删除或无视	若异常值影响不大
				修正	用平均数、中位数、众数、插值或回归的方法处理
	数据变换	分类数据	将分类数据变换用于机器学习	虚拟变量	适合二分的情况，如男女
				独热编码	适用于多分的

					情况
				哑编码	
		数学变换	更改数据范围	取对数	平滑数据
				平方	扩大数据，适合非负情况
				开方	缩小数据，适合非负情况
				差分	平滑数据
				Z-Score	让数据落到有限的范围
				最大最小归一化	
		连续属性离散化	用类别特征代替个体特征	聚类分析	用类别特征代替个体特征
	数据分割	数据分割	更高效地利用有限的数据集	训练集、测试集与验证集	按一定比例分割数据
		“折”		k-folds	将数据分为k折，验证k次，综合考虑这k种结果以得到验证结果
搞模型	选择模型	分类	分类	KNN	分类
				决策树	
				决策树ensemble-随机森林	
				SVM	
				k-means	
				Logit回归	
				朴素贝叶斯	
				ANN	
		回归	预测	多元回归	预测
				岭回归	
				LASSO回归	

		Ensemble	集成多个算法，以期达到更好的效果	Ensemble	常用，集成多个算法，以期达到更好的效果
	特征工程	特征提取	降维过程	PCA	线性方法降维
				LDA	
				ICA	获得相互独立的属性
				LLE	非线性方法降维，保留局部性质
				LE	
				SNE	非线性方法降维，保留全局性质
				t-SNE	
		特征选择	筛选有用的特征，排掉冗余特征	过滤法	在把数据丢到模型前，先进行一波筛选
				包裹法	将后续学习器的性能作为特征筛选标准
				嵌入法	学习器自主选择特征
	模型调优	性能指标	衡量模型效果	交叉验证与其他性能指标	不同模型有不同性能指标
		调参	让模型结果更好看	调参	调参以调整模型效果