

AID 方向《机器学习》课程大纲

机器学习概述 (0.5 天)	机器学习的定义	机器学习的学术化定义
		机器学习的一般化定义
		机器学习的工程化定义
	机器学习的优势	简化代码提升性能
		传统方法无法解决
		自动适应环境变化
		数据挖掘获得洞见
	机器学习的类别	有监督、无监督、半监督和强化学习
		批量和增量学习
		基于实例和模型的学习
	机器学习的问题	训练数据不够充分
		训练数据不够一般
		训练数据质量不高
		包含太多无关特征
		欠拟合或者过拟合
机器学习的评估	模型验证	
	模型测试	
机器学习流程 (2 天)	寻找数据	开放数据存储库
		开放数据元站点
		流行数据集页面
	大处着眼	选择框架
		性能指标
		确认假设
	获取数据	开发环境
		下载数据
		浏览数据
		建测试集
	探索数据	探索地理信息
		探索相关程度
		探索特征组合
	准备数据	数据清理
		文本类别
		转换器类
		特征缩放
		流水管线
	训练模型	线性回归
		决策树回归
交叉验证		
随机森林回归		

		支持向量机回归
	微调模型	网格搜索
		随机搜索
	分析模型	特征重要性
	测试模型	通过测试集评估系统
	应用模型	启动、监控和维护系统
分类问题 (1.5 天)	MNIST	手写数字数据集
		划分训练集和测试集
	二元分类	训练二元分类器
		手动交叉验证
		自动交叉验证
		傻瓜也能得高分
		混淆矩阵
		查准率、召回率和 F1 得分
		权衡查准率和召回率
		PRC 曲线
		高精度分类器
		ROC 曲线
	随机森林分类器	
	多元分类	训练多元分类器
		OVA 策略
		OVO 策略
		类别概率
		分类精度
		错误分析
	多标签分类	训练多标签分类器
		多标签分类器性能评估
多输出分类	训练多输出分类器	
回归问题 (1.5 天)	线性回归	预测函数
		向量形式的预测函数
		损失函数
		标准方程
		计算复杂度
	梯度下降	算法原理
		批量梯度下降
		随机梯度下降
		小批量梯度下降
	多项式回归	幂特征拓展线性回归
	学习曲线	模型性能因训练集大小而变
		高偏差与高方差
	正则化	岭回归
		Lasso 回归

		弹性网络
		提前停止
	逻辑回归	概率估算
		逻辑函数
		预测函数
		损失函数
		决策边界
	Softmax 回归	Softmax 分数
		Softmax 函数
		预测函数
损失函数		
决策边界		
支持向量机 (1 天)	线性 SVM 分类	小间隔分类和大间隔分类
		硬间隔分类和软间隔分类
	非线性 SVM 分类	幂特征拓展线性 SVM 分类
		多项式核函数
		径向基核函数
		计算复杂度
	SVM 回归	线性和非线性 SVM 回归
	工作原理	决策函数
		预测函数
		训练目标
		二次规划
		对偶问题
		核函数
在线 SVM		
决策树 (0.5 天)	决策树分类	训练决策树分类器
		决策树可视化
		预测原理
		估算类别概率
		CART 算法
		计算复杂度
		基尼不纯度和信息熵
		正则化
	决策树回归	训练决策树回归器
		损失函数
		正则化
决策树的不稳定性	决策树对训练数据的敏感性	
集成学习 (1.5 天)	算法投票	硬投票
		软投票
	样本子集	Bagging
		Pasting
		包外评估

	特征子集	Random Patches
		Random Subspace
	随机森林	训练随机森林分类器
		极端随机树
		特征重要性
提升法	自适应提升	
	梯度提升	
堆叠法	堆叠法集成学习原理	
数据降维 (1.5 天)	为什么需要降维	维度灾难
	常用方法	投影
		流形学习
	主成分分析	保留差异性
		主成分
		低维投影
		训练 PCA 转换器
		方差解释率
		合理维度
		PCA 压缩与解压缩
		增量 PCA
	随机 PCA	
	核主成分分析	训练 KPCA 转换器
		核函数与超参数
	局部线性嵌入	训练 LLE 转换器
	其它技巧	多维缩放算法
		等度量映射算法
t-分布随机近邻嵌入算法		
线性判别算法		

本课程共 10 天，要求学生具备如下基础知识：

- Python 语言
- 基于 Numpy、Pandas 和 Matplotlib 的数据分析及可视化