四道大题：线性/逻辑回归；PCA与LDA；决策树；SVM

第一章

machine learning 的概念，什么时候用ml

机器学习种类的划分：如二分类，多分类 回归等等(选择题)

无监督 半监督 全监督 定义 区别 online batch 不考

特征种类 判定(选择题)

第二章

线性回归(大题)回归方程形式，

L2loss，

区别empirical,真实的error不可求。

||w||^2矩阵形式||Xw-y||，矩阵求导规则会给出。

加上L2后的解

逻辑回归

二分类

sigmoid 函数 比如求正类的概率

知道极大似然估计的最终形式

知道迭代优化两种方法的概念 确定性优化 随机优化

梯度下降 记住一阶和二阶展开式

梯度下降找到的最优解全局 or 局部 学习率 step size的影响

牛顿法更新规则 优缺点

逻辑回归 损失函数形式，给定Data

三种梯度下降 SGD加速

PCA LDA各一道大题

SVM bayes一道大题

第三章

PCA概念 特征选择和特征提取的区别

简述PCA过程（四步） 最大化方差 最小化误差

方差形式推导

如何选择特征向量

SVD概念(考选择题 如哪种任务适合SVD)

第四章

LDA 分类效果评价 拉格朗日求解 散度矩阵 如何选择

第五章

特征的概念

特征选择的三种算法

欠拟合和过拟合概念及相应的一些解决措施

常见的正则化形式如：L1 L2

第六章

概念 贝叶斯公式 高斯考小题 朴素贝叶斯

多元高斯分布 方差——>协方差

laplace smooth

最大似然估计

第七章

SVM

smo概念

正则化和不可分情形

硬间隔

软间隔

对偶问题的形式

nonlinear 核函数（计算量小）

高斯核函数的形式 gama C

核函数的条件

第八章

构建决策树 会给定公式（如信息熵）根据某种指标构建 也会用于判断哪个属性更佳

计算条件概率

一些基本公式要记

基尼系数 信息熵等等 计算的 选择小题

缺失值处理 常用的两种剪枝算法 基本思想

过拟合的两个原因:噪声点 有代表性的样本点过少

剪枝算法 举例子 预剪枝算法stop的三个条件

后剪枝数据集的划分，划分后各部分的作用要知道

ccp会考REP不考 ccp剪枝时考虑的两个因素(树的复杂度和training error)

参数化模型和非参数化模型的概念(选择判断题)

回归树 评价指标(概念性问题)

第九章

没有大题

集成学习的概念 集成模型work原因 概念题---选择判断

集成学习的算法(并行：bagging random;序列：adaboost 梯度提升决策树) 如：1000个数据点，有多少数据点未被采样 0.368 e^-1

随机森林：多样性的两个来源 ； 不考算法，但是要知道解决某个具体问题时如何做，如回归问题，分类问题怎么决策(投票法，求均值)

Boosting: 两个weight （alpha m的和每个训练器的)，数据样本点的重要性计算，误分类会变大。

adaboost使用的损失函数 指数损失函数 adaboost公式不用记 给出图来划分了边界回答那些数据点重要性变大即误分类样本点 记住alpha m的计算公式和误分类率

记结论:adaboost的training error随着集成个数增加指数级下降

其他基本不考

第十章

不考大题

GBDT回归任务的loss；GBDT的主要思想，回归模型拟合负梯度

objective function组成的两个部分，不会考具体公式，无法直接求导，索引不能使用传统的SGD等梯度下降方法。

review中的几个问题。

第十一章（未复习）

不考大题

model selection

交叉验证的概念，留出法的计算 混淆矩阵的概念

精准率 召回率 (TP,FP等)precision recal（二者关系） AUC ROC

超参数 验证集用来fine tune超参数 ，