2022实际期末考点（九道大题）

NFL

交叉验证，留一，留一法无效举例

LR拓展到多分类，对数似然，倾斜数据集影响

剪枝影响（训练时间、测试时间、过拟合）

SVM硬间隔，计算

bagging流程，偏差方差

朴素贝叶斯，拉普拉斯，计算

感知机原理

距离度量四要素，合理和不合理各两个举例子

附加题：英文缩写写全程，10选5

一、绪论

归纳偏好一般性原则：奥卡姆剃刀：若有多个假设与观察一致，选择最简单的那个

NFL重要前提：所有问题出现机会均等（同等重要）（假设了f的均匀分布）

脱离具体问题，空谈什么学习算法好没有意义

- 泛化：训练的模型推广到新的问题上的能力
- 独立同分布（所有样本从同一分布中独立的抽样出来）

二、模型评估与选择

过拟合：经验误差小但泛化误差大

评估方法：（测试集）自助法&k折交叉验证：方差变大

性能度量：（评估性能优劣，衡量泛化能力）TP，FP，TN，FN（后面是预测结果），P，R=TPR，FPR（与TPR分子分母都不同）

AUC算面积，F1度量公式

比较检验：为什么需要?

泛化性能是由学习算法的能力（偏差）、数据的充分性（方差）以及学习任务本身的难度（噪声）所共同决定的

三、线性模型

> 形式简单，易于建模，机器学习思想，可解释性强

计算01线性回归求w,b最优解

对数几率回归：

- 实际是分类学习算法：找一个单调可微函数（对数几率函数）
- 优点：得到概率预测；凸：无需假设数据分布

LDA：同类近，异类远，新来的也投影过来

计算02LDA

- 贝叶斯决策理论：两类数据同先验、满足高斯分布且协方差相等，LDA可最优分类
- 当假设各类样例的协方差矩阵相同时，FDA 退化为线性判别分析 LDA。

ECOC

四、决策树

信息增益公式：前-后，增益率公式，基尼指数公式（不用log）

计算：剪枝，连续，缺失

预、后剪枝优缺点（时间+性能）

五、SVM

计算03对偶问题KKT

hinge损失，松弛变量

六、神经网络

链式法则

七、贝叶斯

MLE&MAP

朴素贝叶斯+拉普拉斯修正

贝叶斯网

八、聚类

K-Means代码

算法 1 k 均值算法

初始化所有簇中心 $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$;
repeat
    Step 1: 确定 $\{\boldsymbol{x}_i\}_{i=1}^m$ 所属的簇, 将它们分配到最近的簇中心所在的簇.

$$\Gamma_{ij} = \begin{cases} 1, & \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2 \leq \|\boldsymbol{x}_i - \boldsymbol{\mu}_{j'}\|^2, \forall j' \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

    Step 2: 对所有的簇 $j \in \{1, \cdots, k\}$, 重新计算簇内所有样本的均值, 得到新的簇中心 $\boldsymbol{\mu}_j$:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^m \Gamma_{ij} \boldsymbol{x}_i}{\sum_{i=1}^m \Gamma_{ij}} \tag{14}$$

until 目标函数 J 不再变化.

- E：计算期望：利用当前估计的参数值计算对数似然的期望值;
- M：最大化：寻找使E步产生的似然期望最大的参数值

九、集成

文字内容

Bagging

(贝叶斯最优错误率)

十、PCA证明