1. 统计机器学习：监督学习

机器学习基本概念

线性回归

提升算法

2. 机器学习基本概念

机器学习：从数据中学习知识(模式 和 规律)

像素图像 (数据矩阵空间) → 人、狗(高级语义空间)

1.原始数据中提取特征

2.学习映射函数𝑓

3.通过映射函数𝑓将原始数据映射到语义空间，即寻找数据和任务目标之间的关系

3. 机器学习的分类

监督学习：回归、分类

无监督学习：聚类、降维

强化学习：环境交互

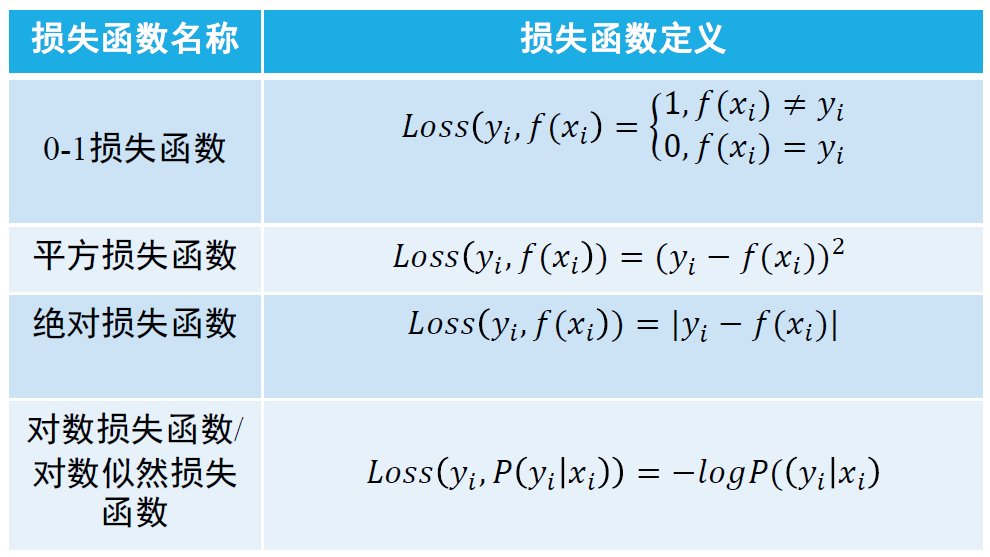
半监督学习：部分数据有标签，部分数据无标签

4. 监督学习的重要元素

标注数据：标识了类别信息的数据

学习模型：如何学习得到映射模型

损失函数：如何对学习结果进行度量



5. 经验风险与期望风险

经验风险(empirical risk)：

训练集中数据产生的损失。经验风险越小说明学习模型对训练数据拟合程度越好。

期望风险(expected risk):

当测试集中存在无穷多数据时产生的损失。期望风险越小，学习所得模型越好。

6. “过学习(over-fitting)”与“欠学习(under-fitting)”



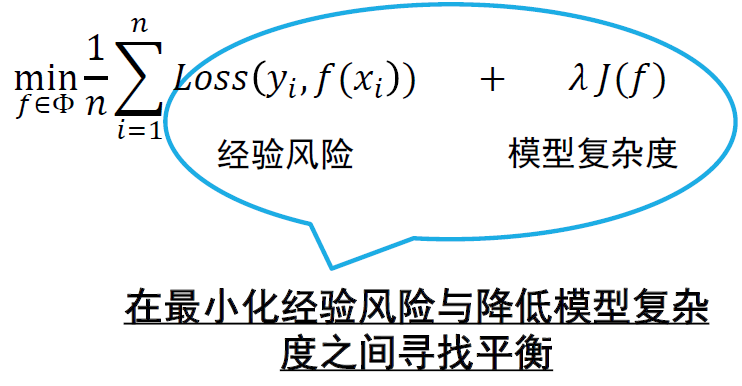
7. 结构风险

经验风险最小化：仅反映了样本数据的数据特征

期望风险最小化：无法得到全量数据来进行计算

结构风险最小化(structural risk minimization)：

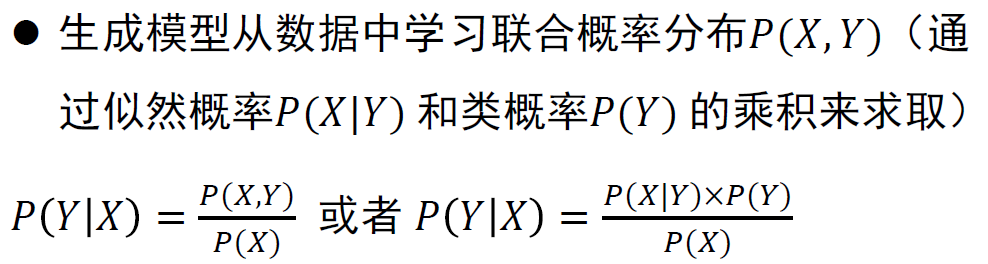
为了防止过拟合，在经验风险上加上表示模型复杂度的正则化项(regulatizer)或惩罚项(penalty term ) ：



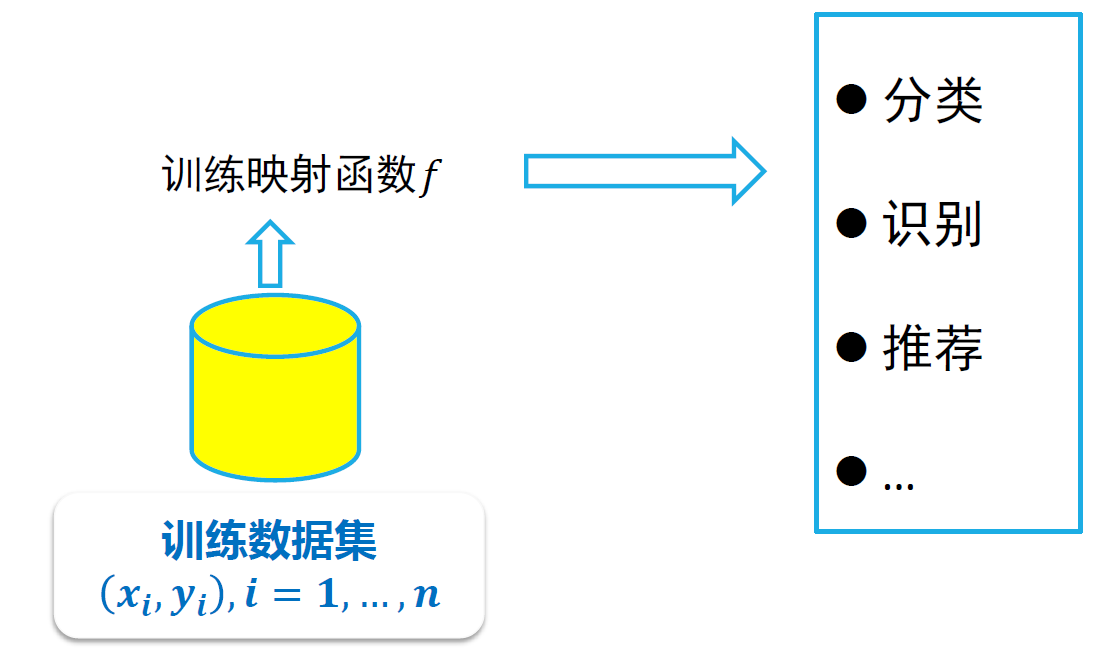
8. 判别模型与生成模型

监督学习方法又可以分为生成方法(generative approach)和判别方法(discriminative approach)

所学到的模型分别称为生成模型(generative model)和判别模型(discriminative model)



9. 监督学习



10. 线性回归分析

分析不同变量之间存在关系的研究叫回归分析

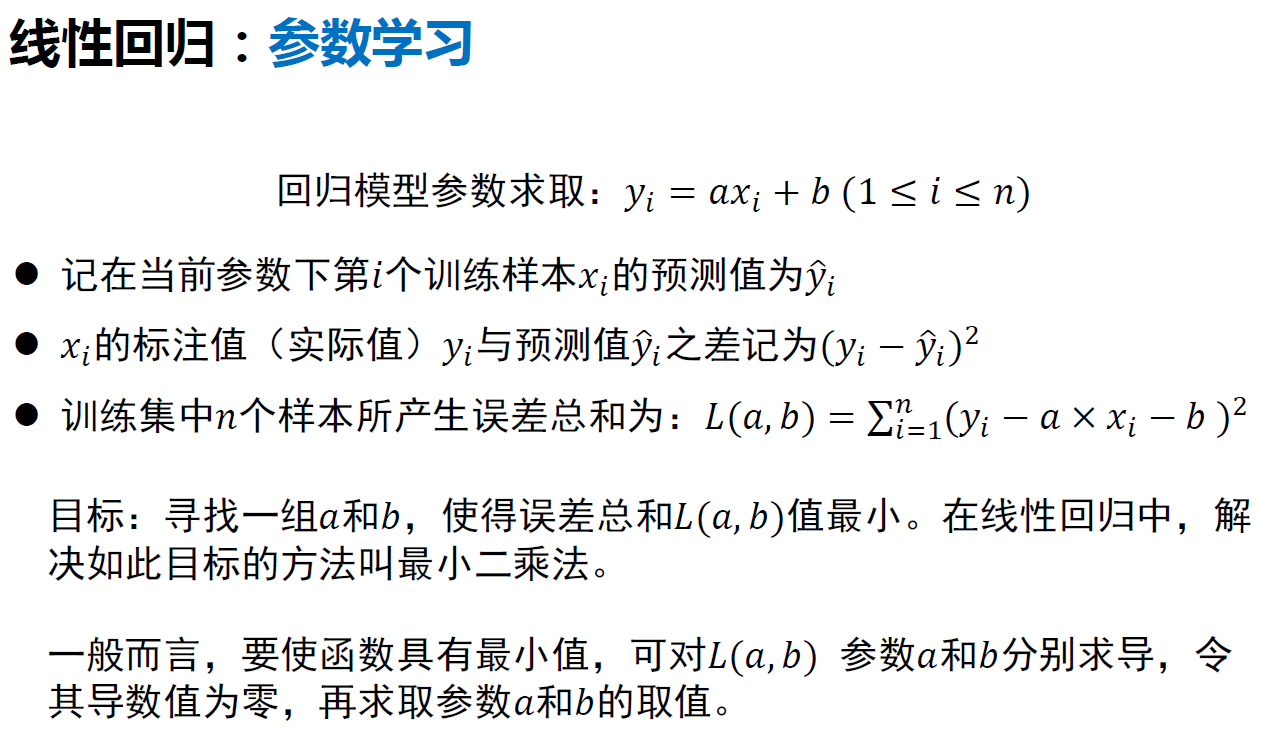
刻画不同变量之间关系的模型被称为回归模型

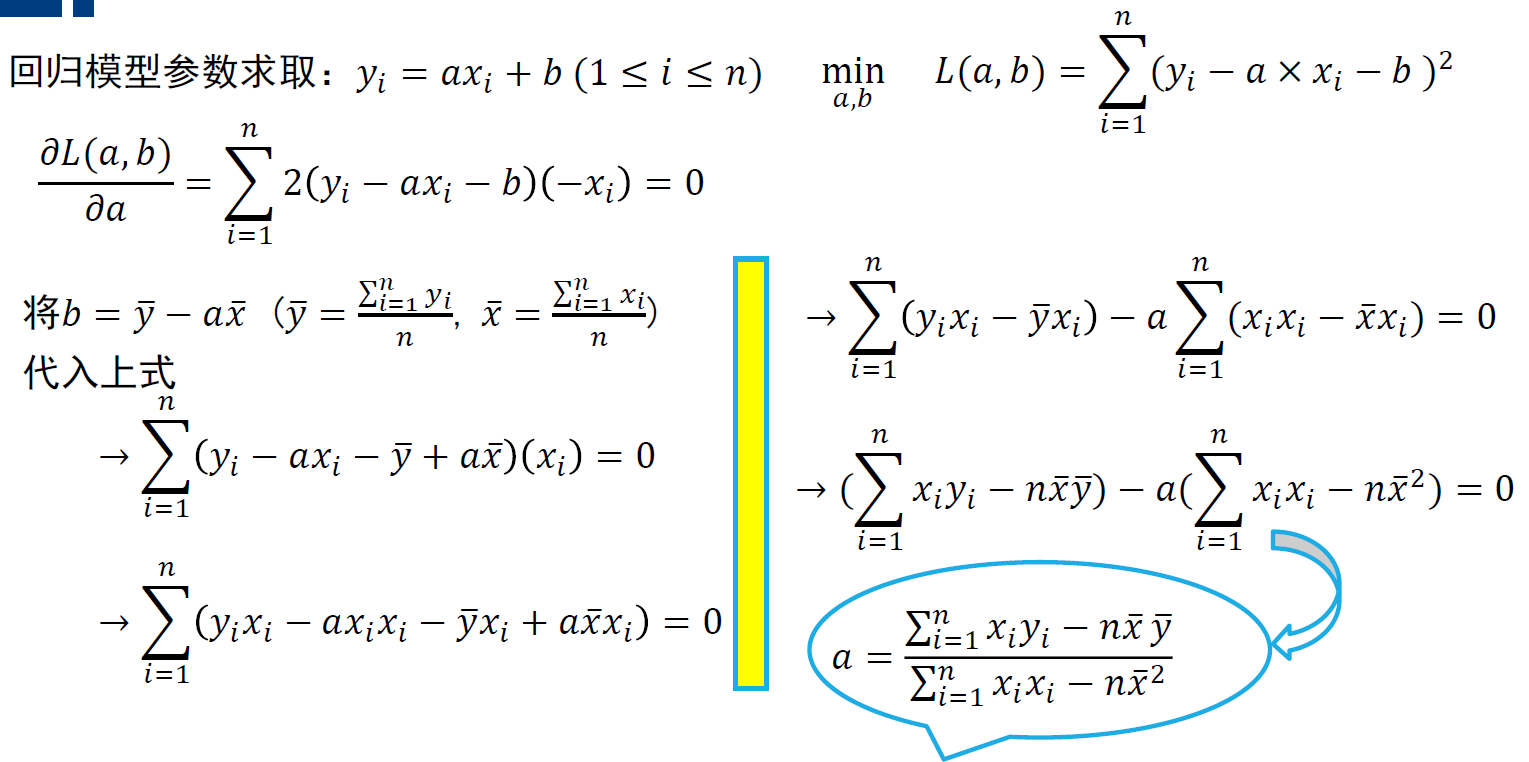
如果这个模型是线性的，则称为线性回归模型

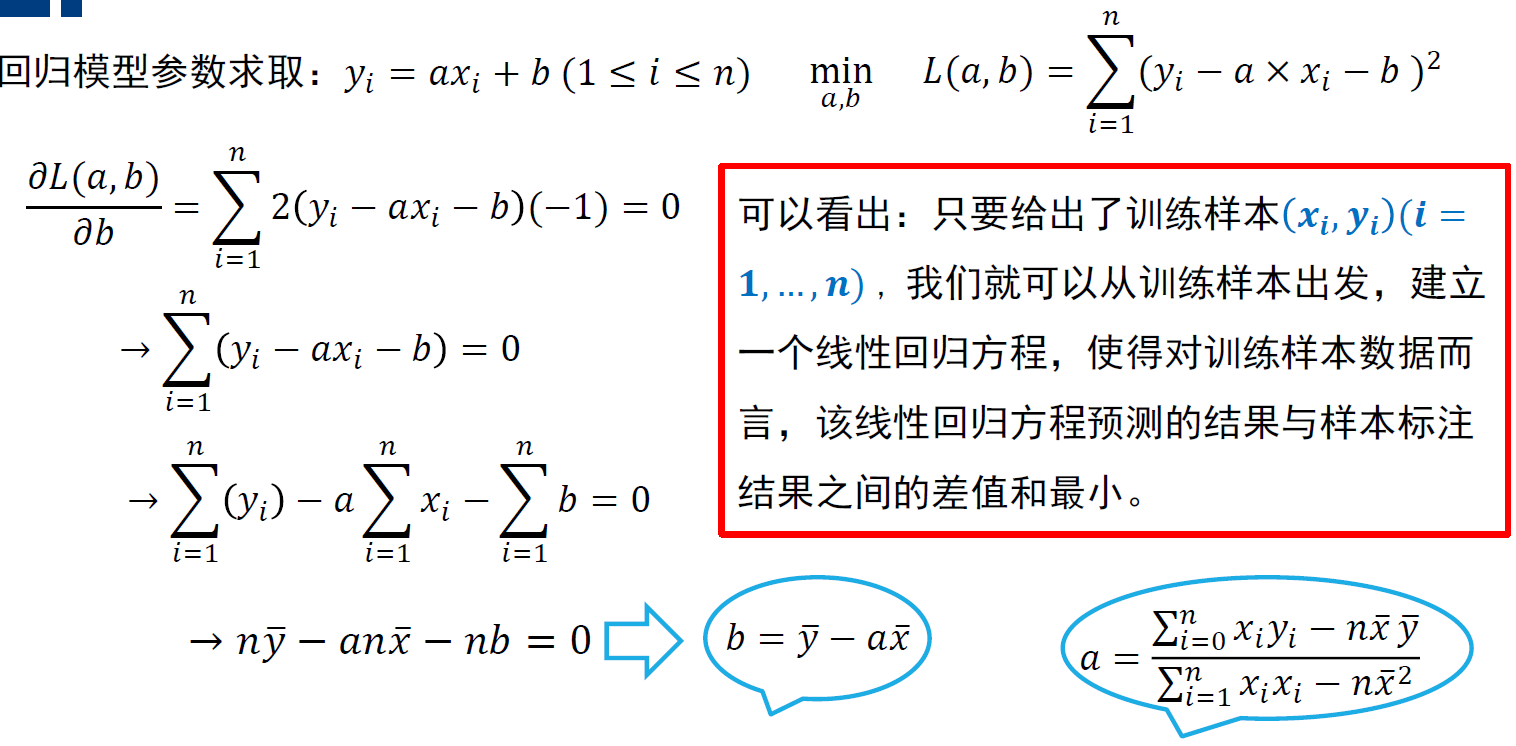
英国著名生物学家兼统计学家高尔顿：

父母平均身高每增加一个单位, 其成年子女平均身高只增加0.516个单位，它反映了这种“衰退(regression)”效应（“回归”到正常人平均身高）

11. 参数学习







12. 提升算法

Boosting (adaptive boosting, 自适应提升)

对于一个复杂的分类任务，可以将其分解为若干子任务，

然后将若干子任务完成方法综合，最终完成该复杂任务。

将若干个弱分类器(weak classifiers)组合起来，

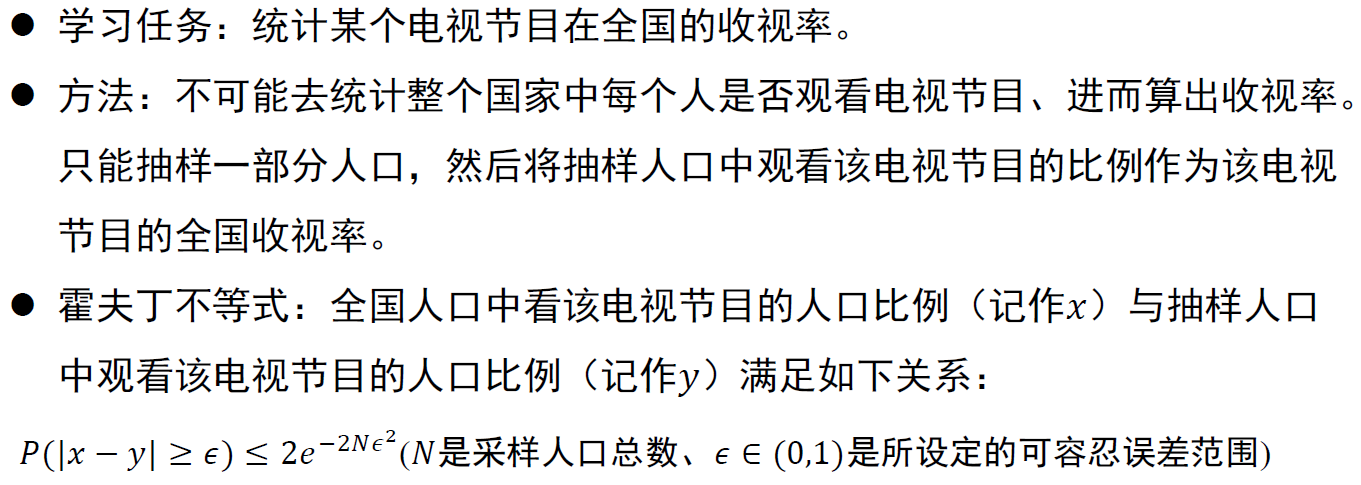
形成一个强分类器(strong classifier)。

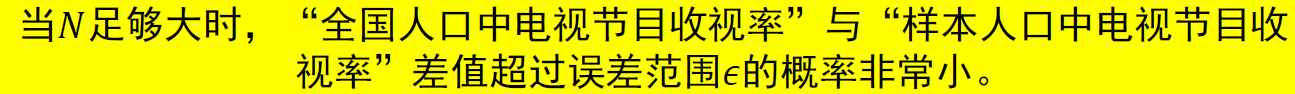
13. 计算学习理论(Computational Learning Theory)

可计算：什么任务是可以计算的？图灵可停机即可计算

可学习：什么任务是可以被学习的、从而可被学习模型来完成？

14.计算学习理论：霍夫丁不等式(Hoeffding’s inequality)





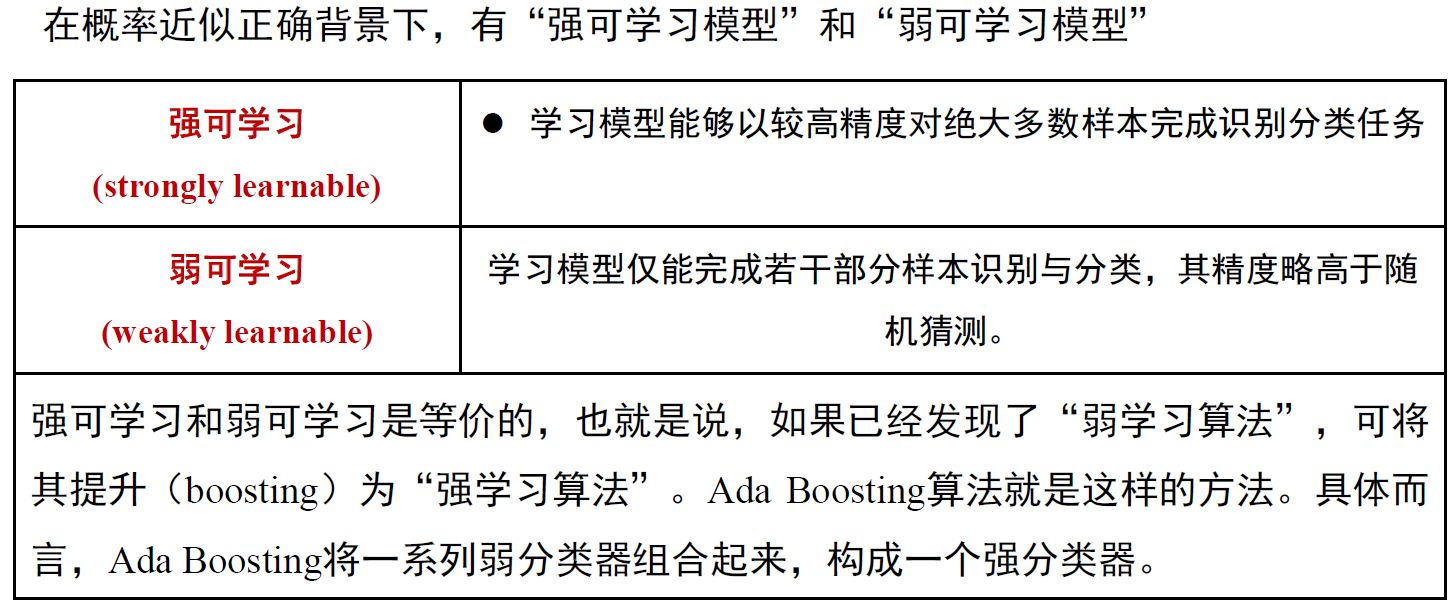
15. 计算学习理论：概率近似正确(probably approximately correct, PAC)

问题：如果得到完成该任务的若干“弱模型”，

是否可以将这些弱模型组合起来，

形成一个“强模型”，该“强模型”产生误差很小呢？

这就是概率近似正确（PAC）要回答的问题。

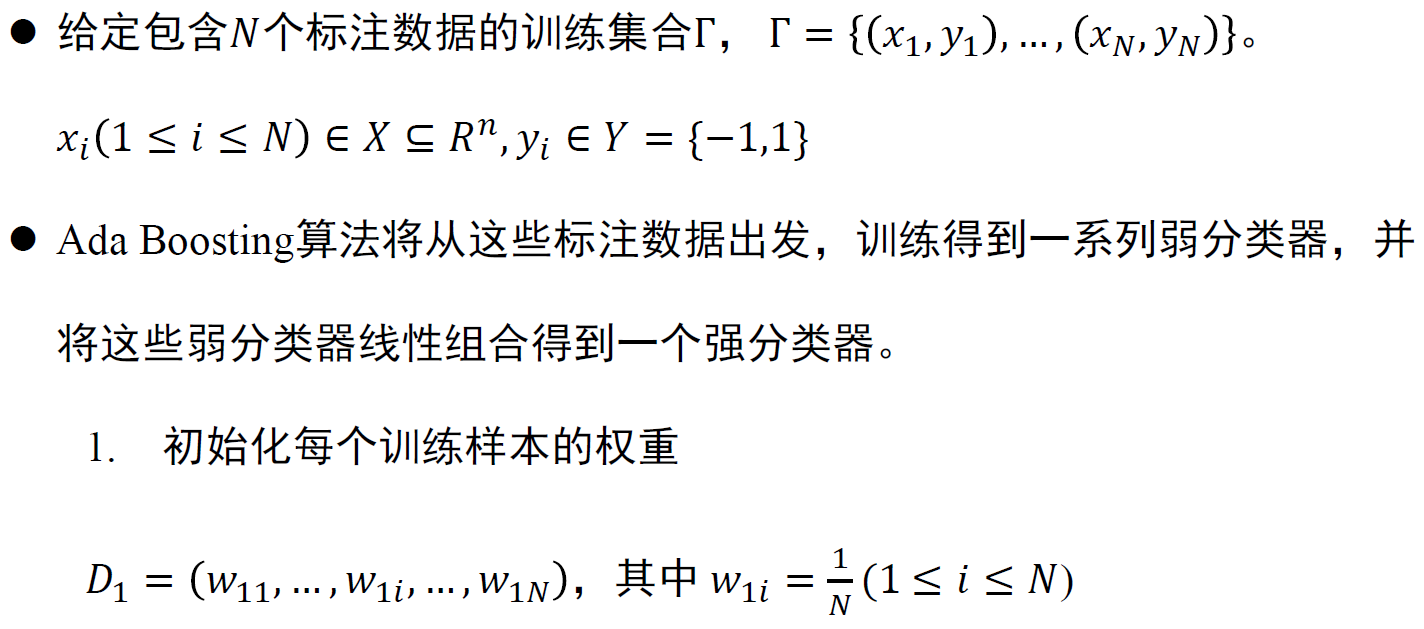


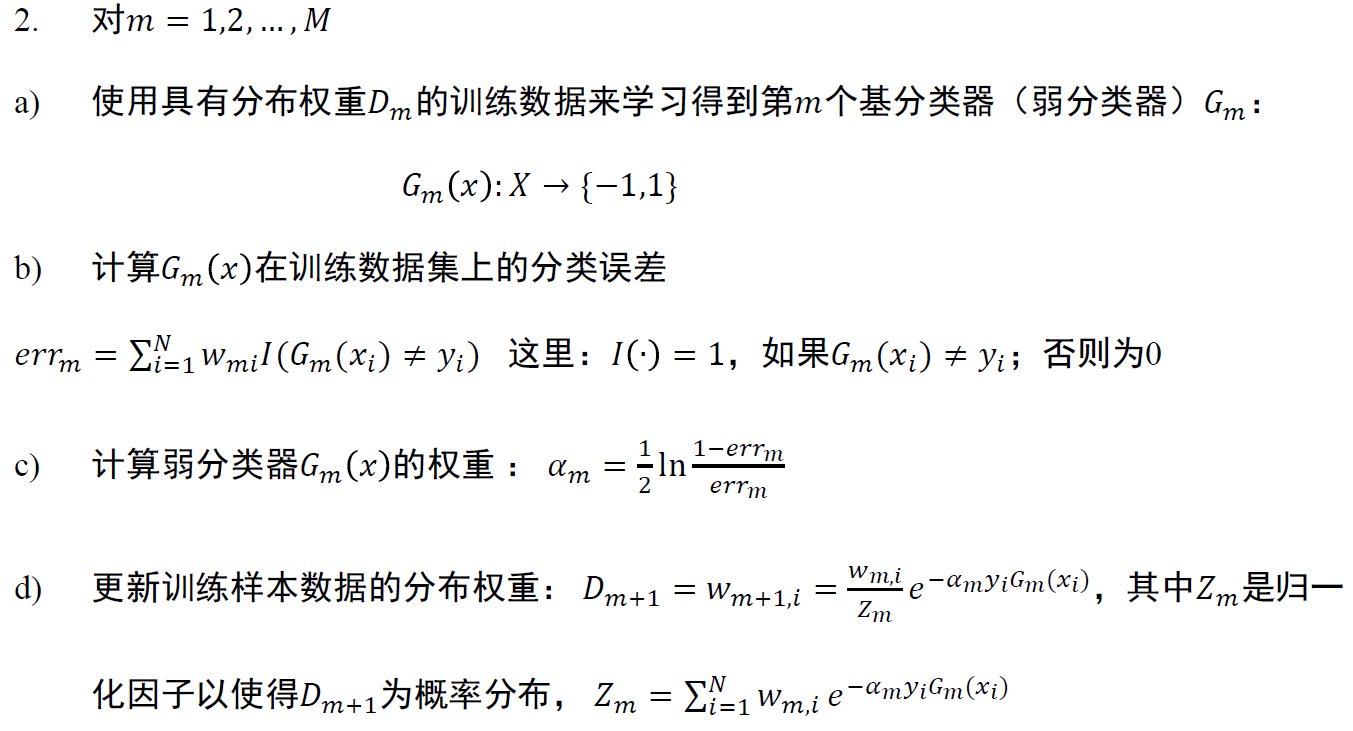
16. Ada Boosting

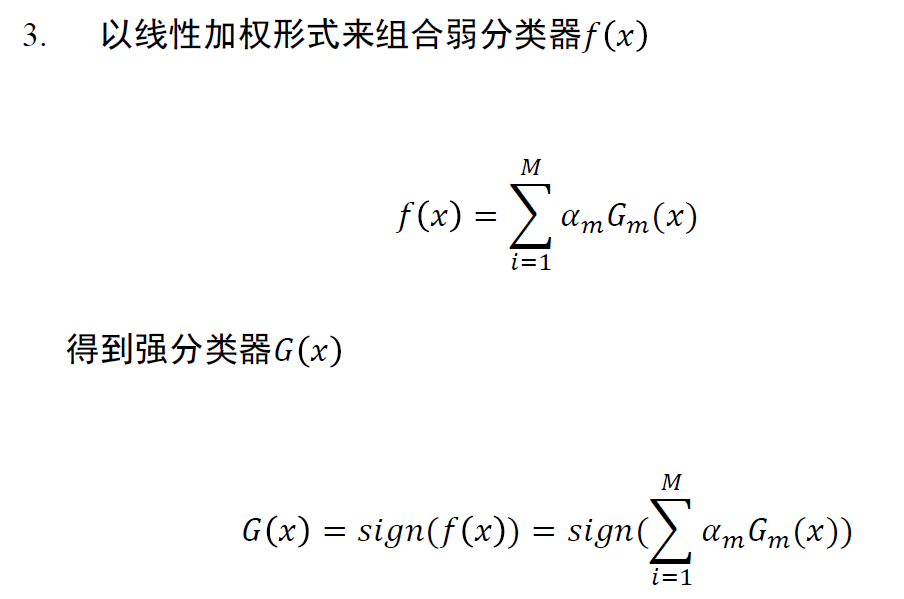
🟎Ada Boosting算法中两个核心问题：

🟎在每个弱分类器学习过程中，如何改变训练数据的权重：提高在上一轮中分类错误样本的权重。

🟎如何将一系列弱分类器组合成强分类器：通过加权多数表决方法来提高分类误差小的弱分类器的权重，让其在最终分类中起到更大作用。同时减少分类误差大的弱分类器的权重，让其在最终分类中仅起到较小作用。







17. Ada Boosting算法解释

