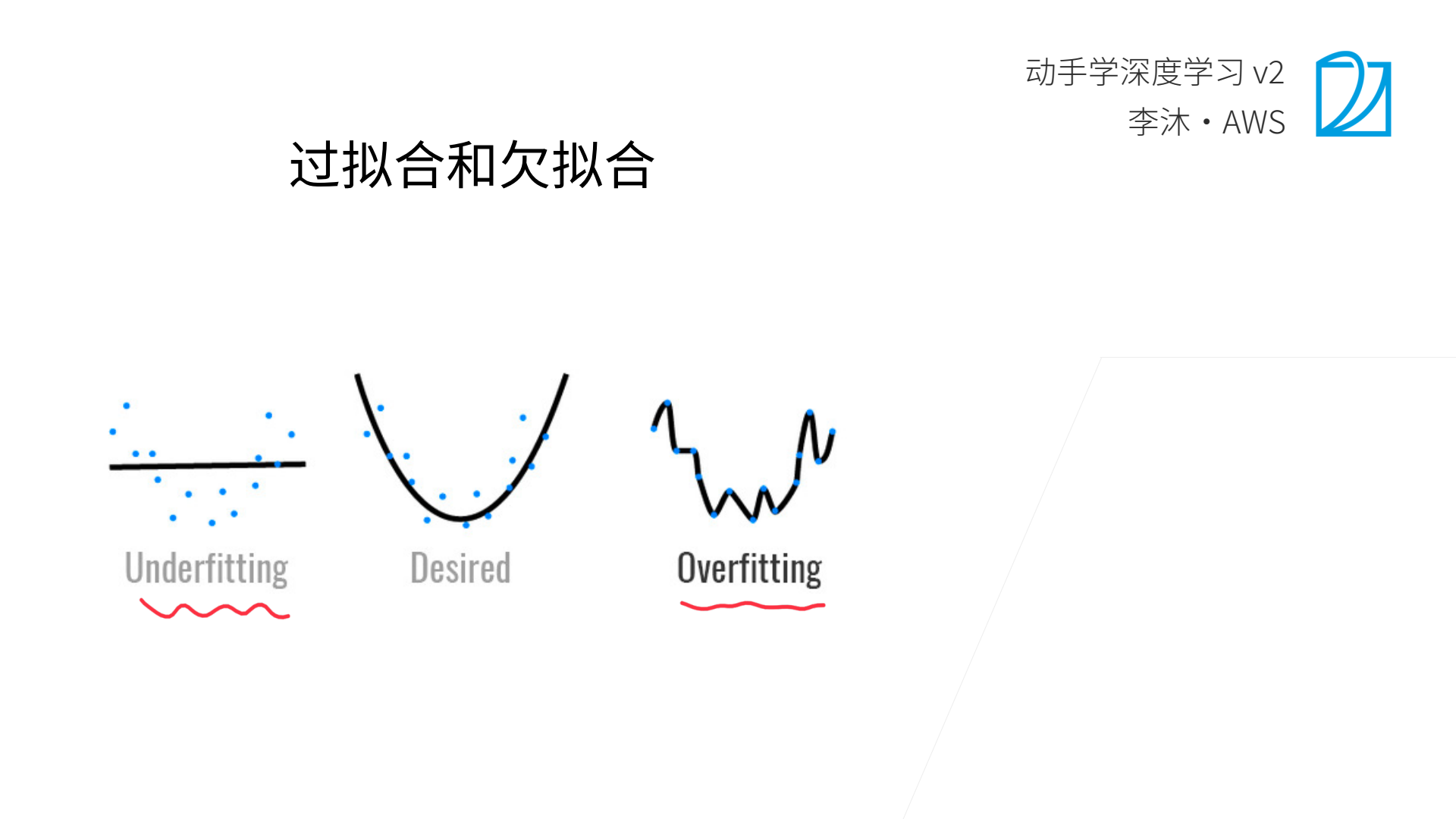


过拟合和欠拟合

过拟合和欠拟合

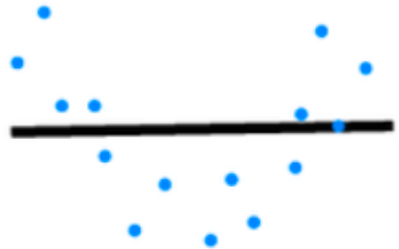
应该根据数据的复杂度来选择模型容量

数据

模型容量	简单	复杂
低	正常	欠拟合
高	过拟合	正常

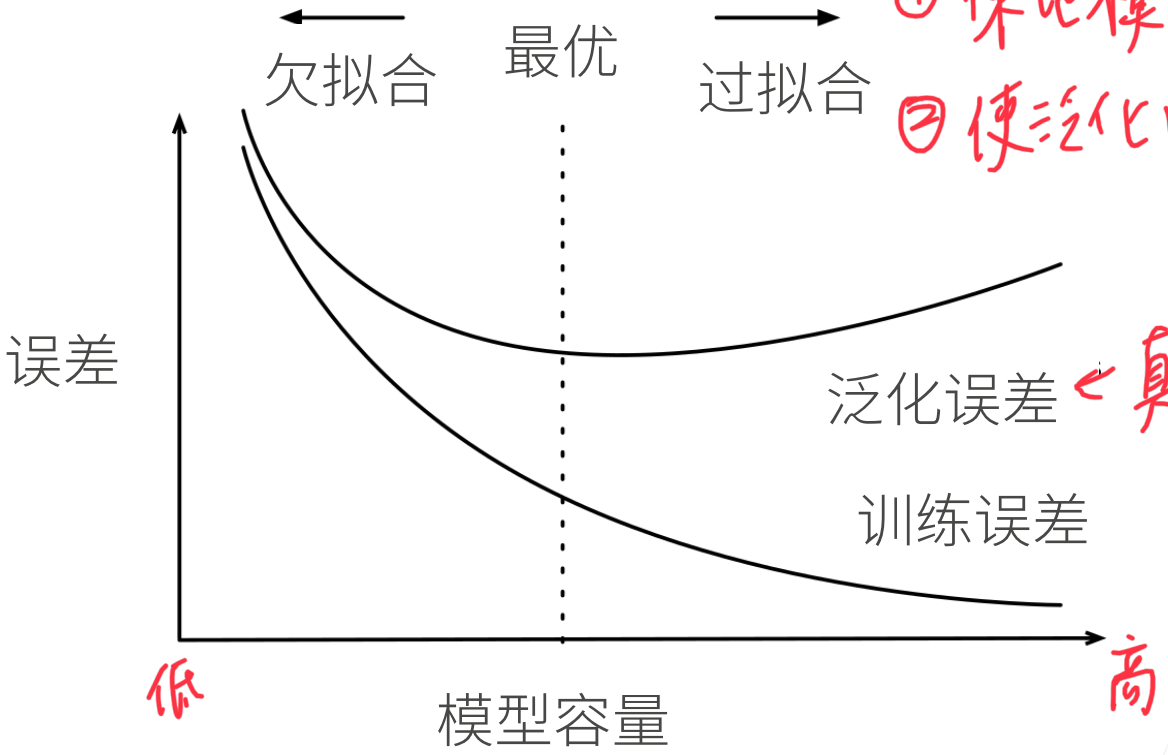
模型容量

- 拟合各种函数的能力
- 低容量的模型难以拟合训练数据
- 高容量的模型可以记住所有的训练数据



模型容量的影响

深度学习的核心：
① 保证模型足够大
② 使泛化误差尽可能小



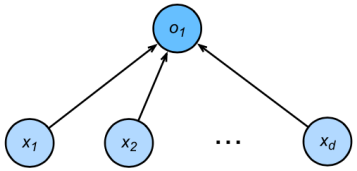
真正关心的误差

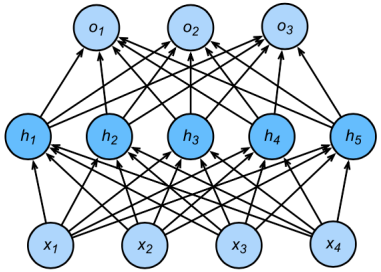
估计模型容量

- 难以在不同的种类算法之间比较
  - 例如树模型和神经网络
- 给定一个模型种类，将有两个主要因素
  - 参数的个数
  - 参数值的选择范围

$d+1$

$(d+1)m+(m+1)k$





VC 维

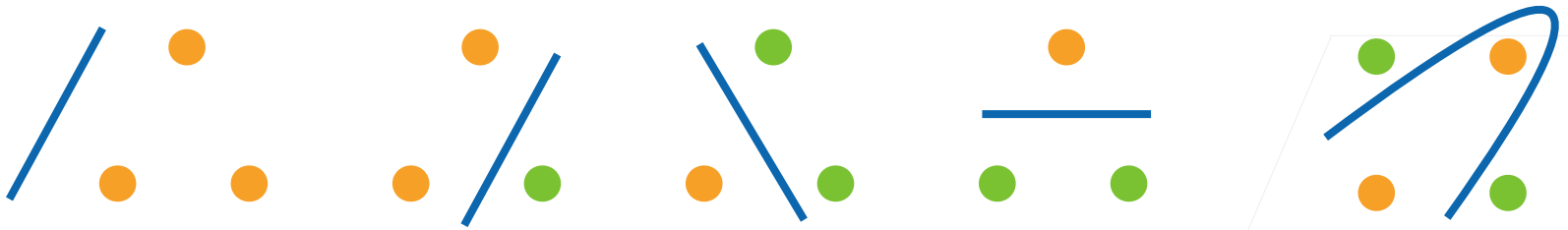
- 统计学习理论的一个核心思想
- 对于一个分类模型，VC等于一个最大的数据集的大小，不管如何给定标号，都存在一个模型来对它进行完美分类

Vladimir **V**apnik

Alexey **C**hervonenkis

线性分类器的 VC 维

- 2 维输入的感知机，VC 维 = 3
  - 能够分类任何三个点，但不是4个 (xor)



- 支持 N 维输入的感知机的 VC 维是 $N + 1$
- 一些多层感知机的 VC 维 $O(N \log_2 N)$

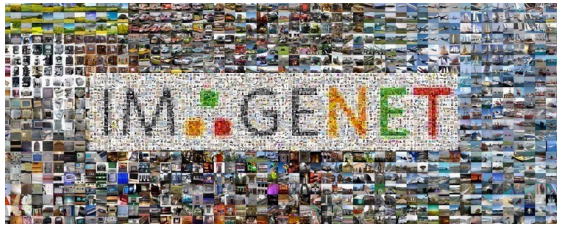
VC 维的用处

- 提供为什么一个模型好的理论依据
  - 它可以衡量训练误差和泛化误差之间的间隔
- 但深度学习中很少使用
  - 衡量不是很准确
  - 计算深度学习模型的 VC 维很困难

数据复杂度

- 多个重要因素
  - 样本个数
  - 每个样本的元素个数
  - 时间、空间结构
  - 多样性

VS



根据数据复杂度来选择模型容量

总结

- 模型容量需要匹配数据复杂度，否则可能导致欠拟合和过拟合
- 统计机器学习提供数学工具来衡量模型复杂度
- 实际中一般靠观察训练误差和验证误差