《智慧医疗》课程

二、机器学习基础

黄恩待

智能医学与生物医学工程研究院

2024-12-10

CONTENTS

01

基本概念

一、基本概念

机器学习：

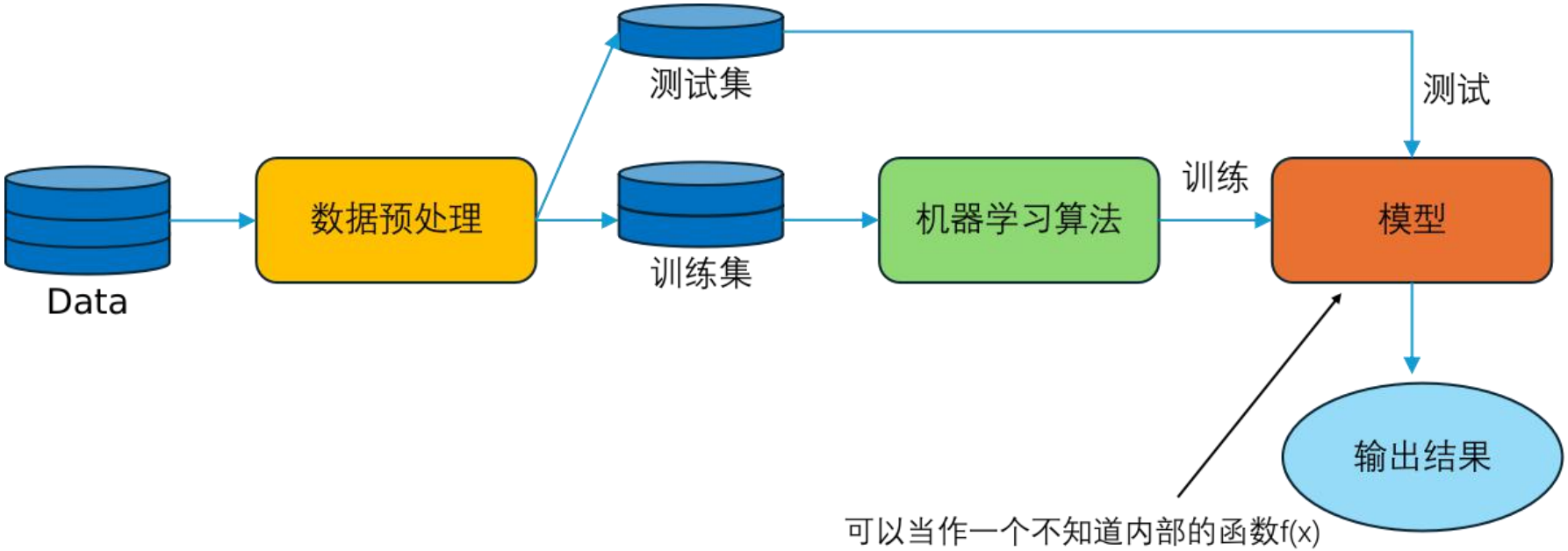
- 研究如何模拟和实现人类学习行为

深度学习：

- 以神经网络为代表的，层数很深的机器学习
- 实际中“机器学习“和“深度学习”不同

一、基本概念

机器学习主要流程:



一、基本概念

数据预处理:

- Excel存储数据（.csv文件更好），Python的Pandas包导入数据
- 通常数据量要是特征数量的10倍
- 将数据处理为机器可读形式
- One-hot编码(如果类别大于2)

Human-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

一、基本概念

01

监督学习

回归与分类

02

无监督学习

聚类与降维

03

强化学习

策略优化与奖励机制

一、基本概念

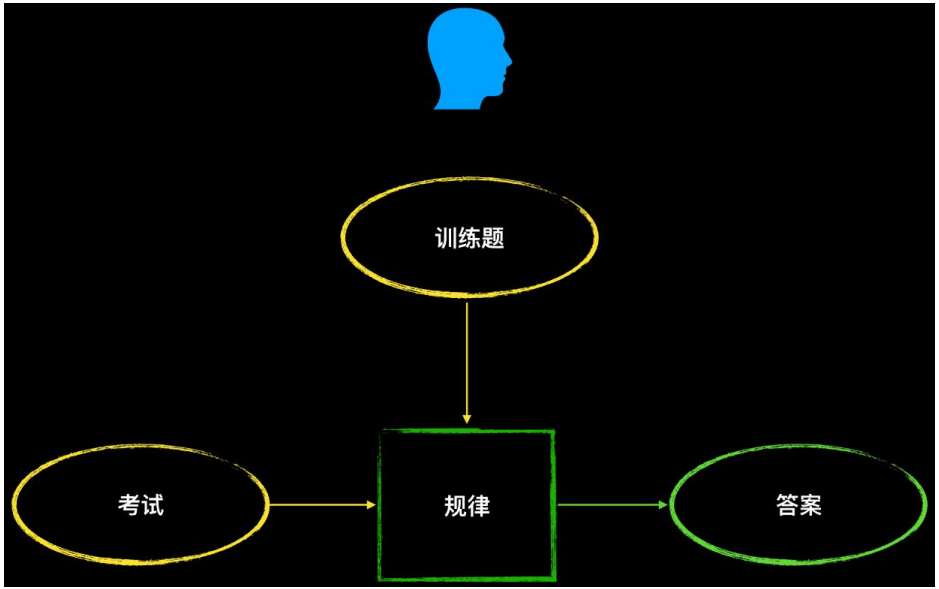
经典定义：利用经验改善系统自身的性能

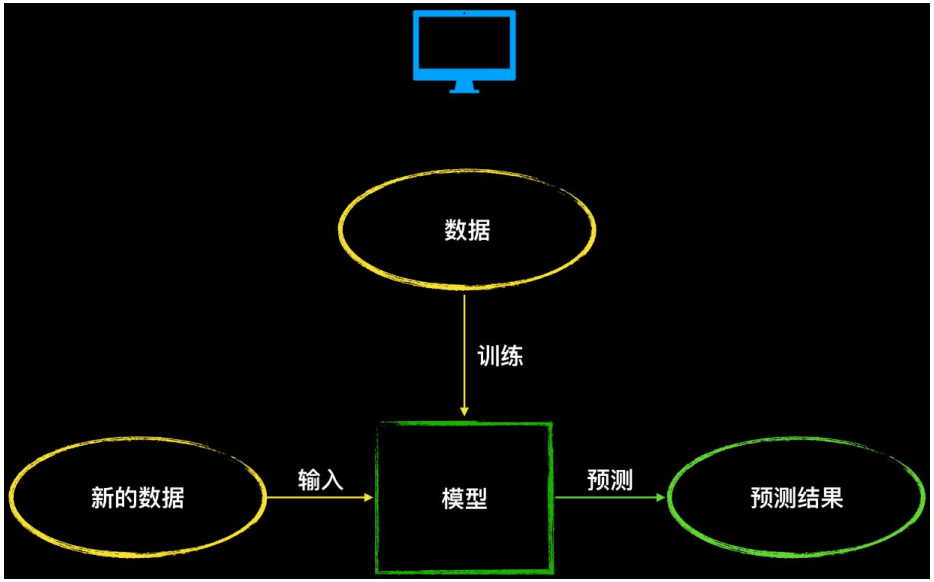
经验 → 数据

随着该领域的发展，目前主要研究智能数据分析的理论和算法，并已成为智能数据分析技术的源泉之一

一、基本概念

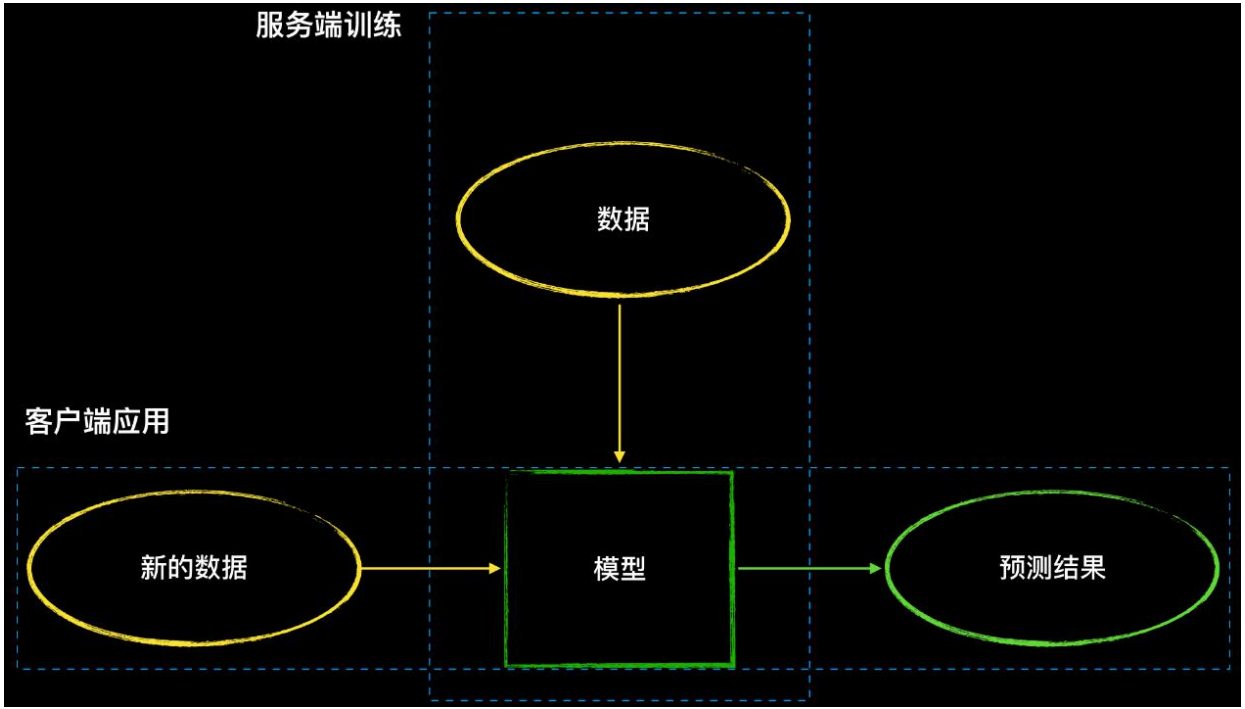
- 学习“规律”





一、基本概念

- 学习“规律”



一、基本概念

- 学习“规律”
- ≈寻找一个函数f(x)

Speech Recognition

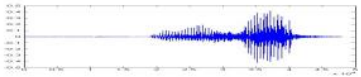
f() = “How are you”

Image Recognition

f() = “Cat”

Playing Go

f() = “5-5” (next move)

一、基本概念

“文献筛选”的故事

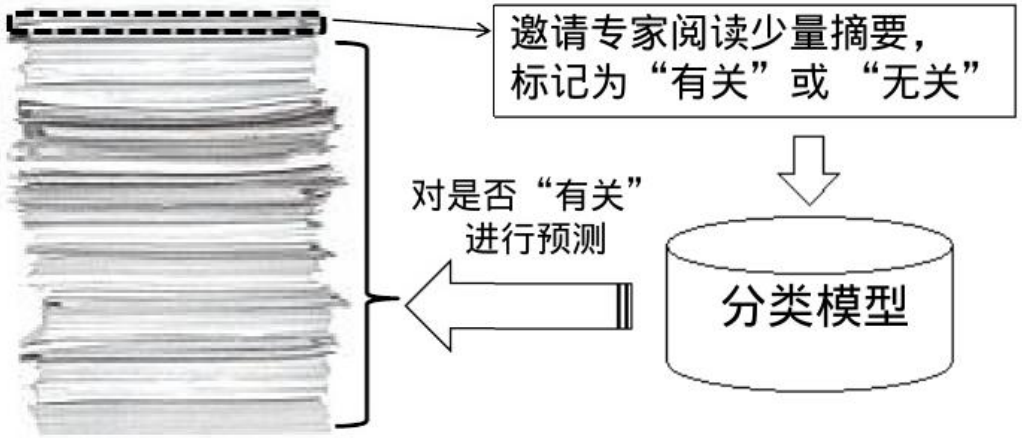
- 在一项关于婴儿和儿童残疾的研究中，美国Tufts医学中心筛选了约 33,000 篇摘要。尽管Tufts医学中心的专家效率很高，对每篇摘要只需 30 秒钟，但该工作仍花费了 250 小时
- 每项新的研究都要重复这个麻烦的过程!
- 需筛选的文章数在不断显著增长!

a portion of the 33,000 abstracts

一、基本概念

“文献筛选”的故事

为了降低昂贵的成本，Tufts医学中心引入了机器学习技术



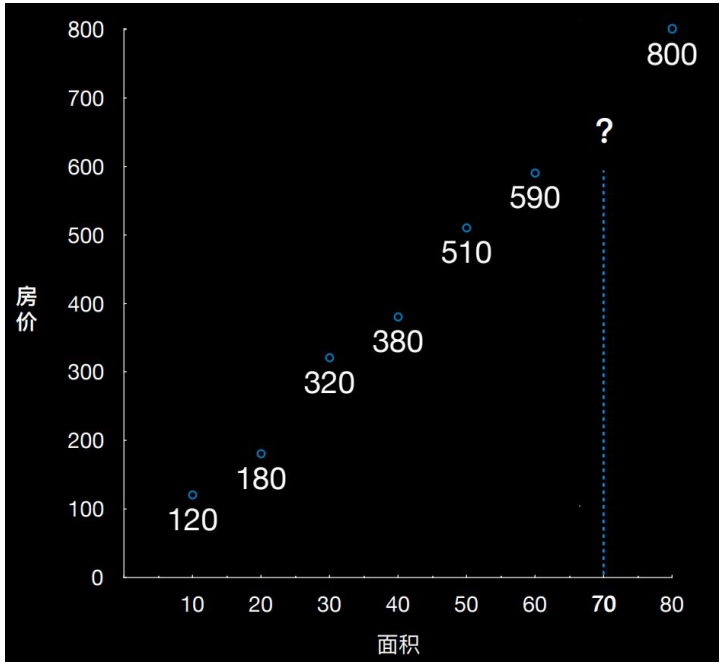
人类专家只需阅读 **50** 篇摘要，系统的自动筛选精度就达到 **93%**

人类专家阅读 **1,000** 篇摘要，则系统的自动筛选敏感度达到 **95%**

(人类专家以前需阅读 **33,000** 篇摘要才能获得此效果)

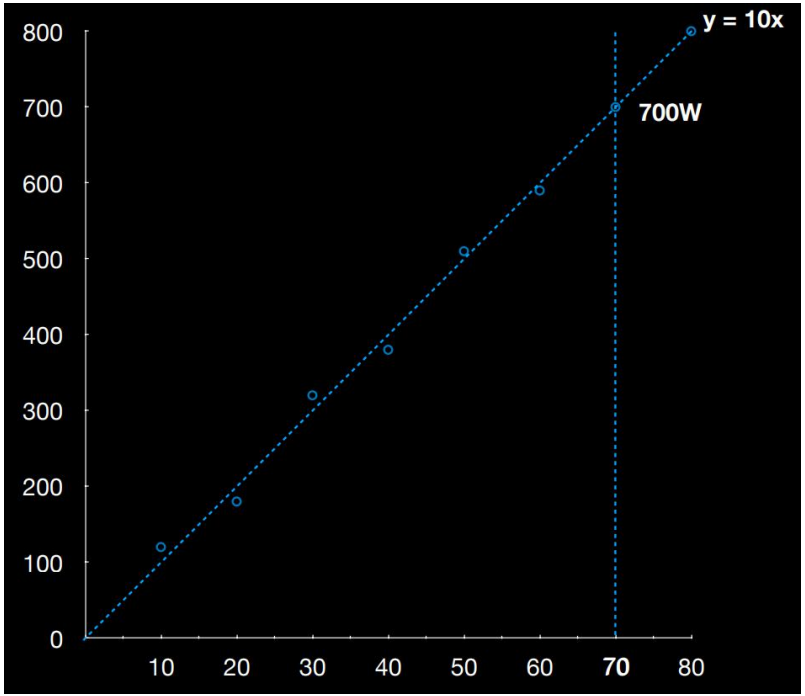
一、基本概念

- 已知10平，20平、……60平，80平房子的价格
- 求70平房子的价格?



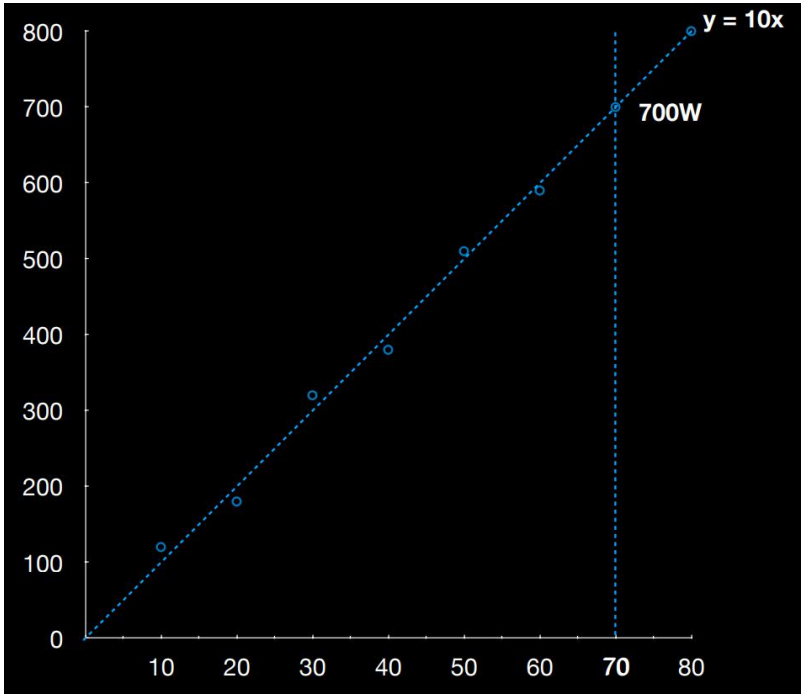
一、基本概念

- 已知10平，20平、……60平，80平房子的价格
- 求70平房子的价格?



一、基本概念

- 已知10平，20平、……60平，80平房子的价格
- 求70平房子的价格?
- 模型一定是线性的?
- 只有一个因素?
- 我怎么知道拟合出来的规律是正确的?



一、基本概念

最小化误差函数：

$$\text{Min} \quad \frac{1}{m}\sum_{x=1}^{m}(f(x_i)-y_i)^2$$

即求两个参数k和b，使上述函数取得最小值（How？）

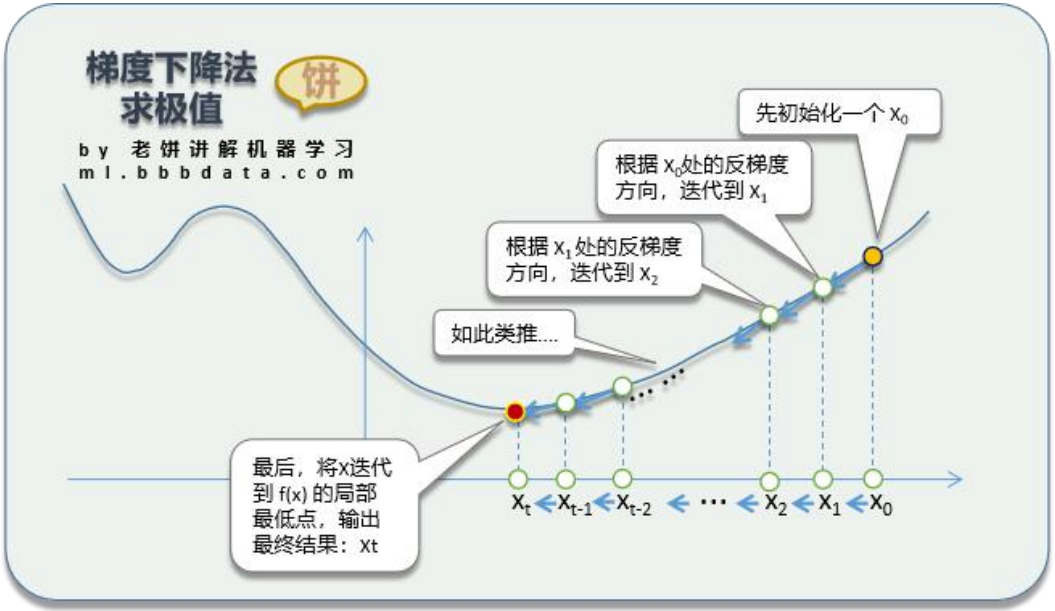
$$\text{Min} \quad \sum_{x=1}^{m}(kx_i+b-y_i)^2$$

一、基本概念

最小化误差函数：

方法一：求导数，取导数等于0的点（难求解）

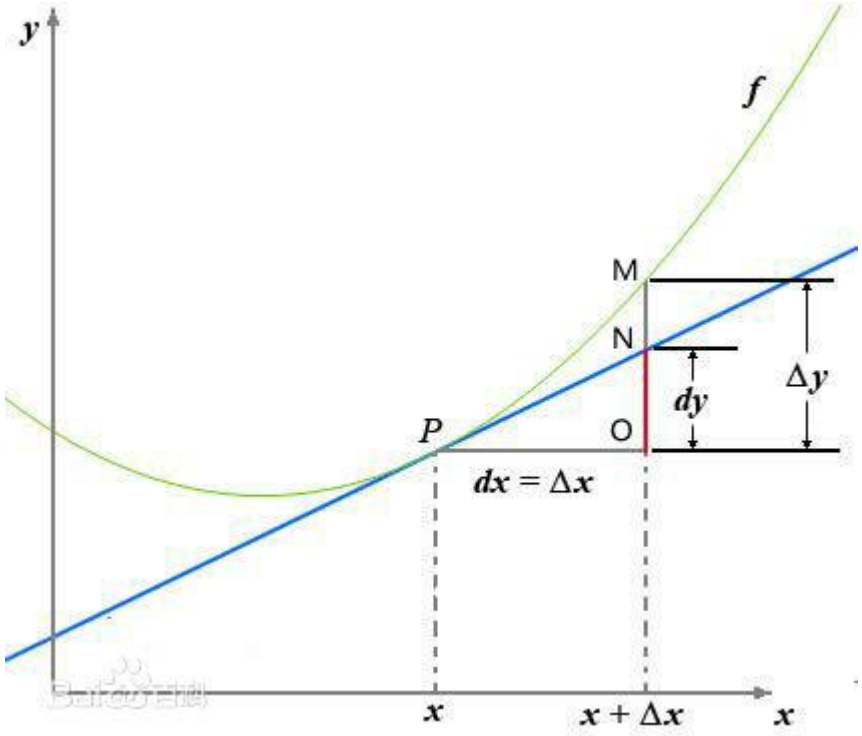
方法二：不断的靠近最小值（即学习的过程，梯度下降）



一、基本概念

导数：

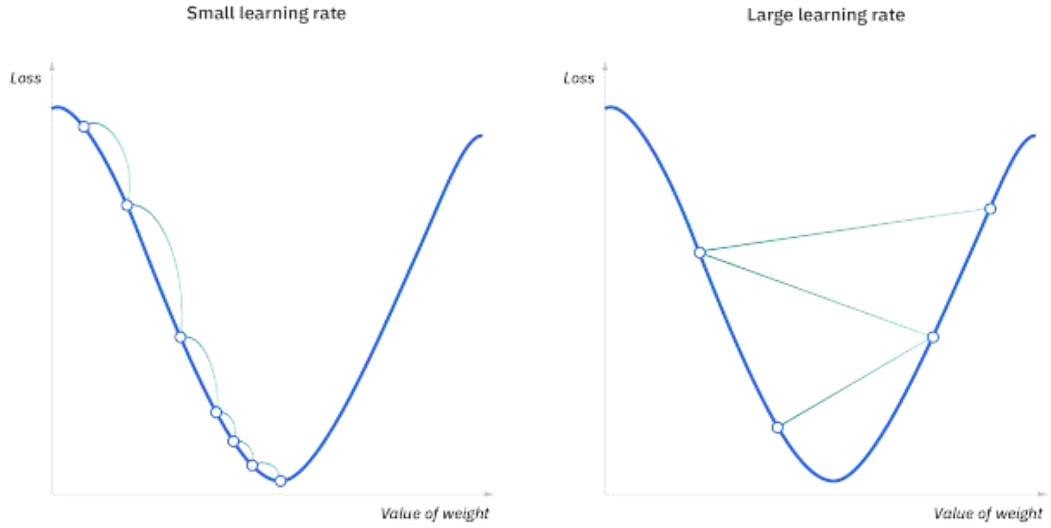
$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

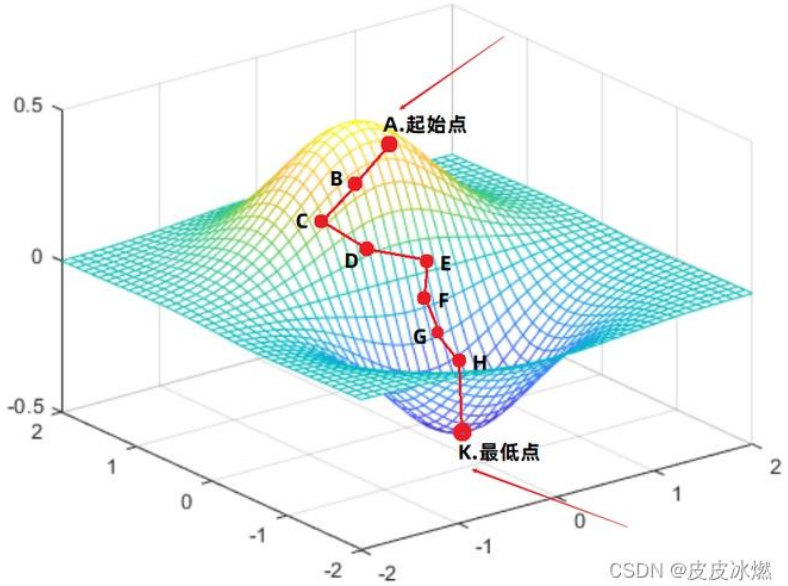


一、基本概念

梯度下降：

很小的步长（学习率）





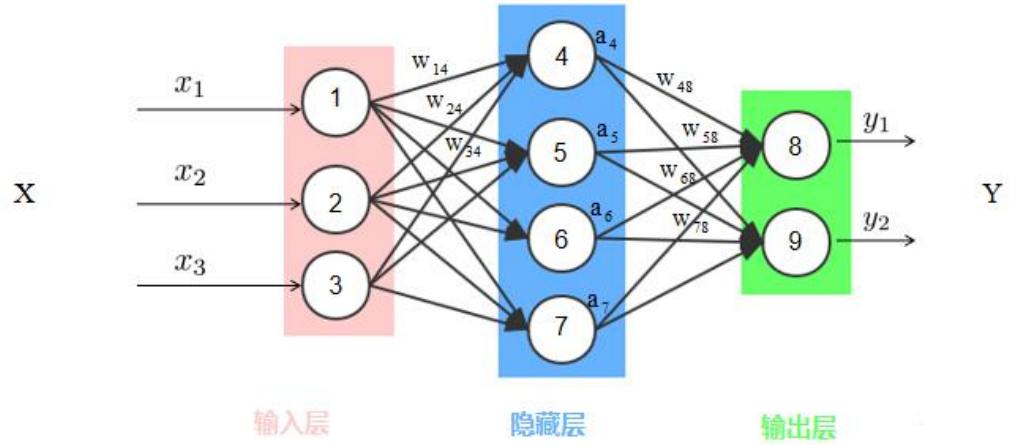
02

机器学习模型

二、机器学习模型

神经网络:

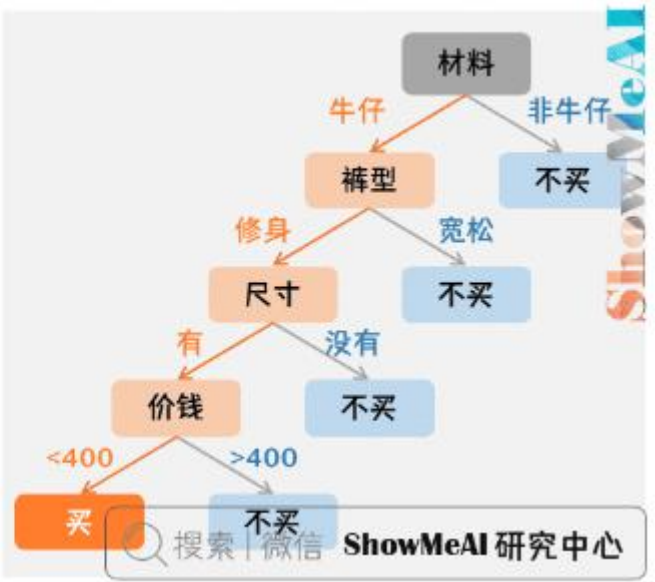
- 模仿人脑神经元结构的计算模型
- 神经元、激活函数组成



二、机器学习模型

决策树及随机森林模型：

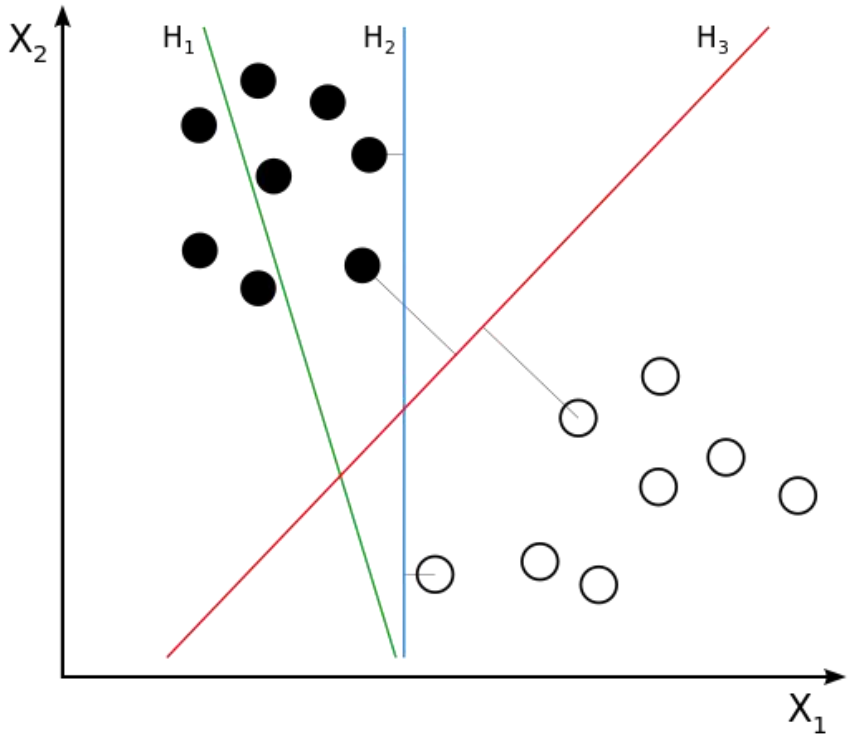
- 模仿人的决策（如Yes/No）



二、机器学习模型

支持向量机(分类)：

- 模仿人脑神经元结构的计算模型
- 神经元、激活函数组成



03

机器学习评估

三、机器学习评估

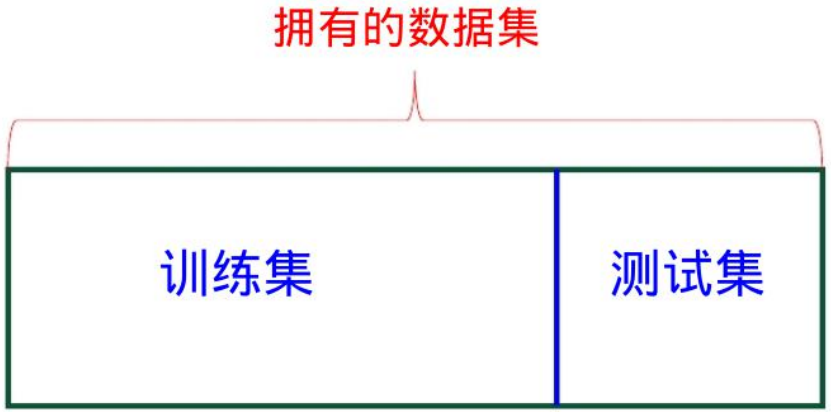
如何获得测试结果?

- 留出法 (hold-out)
- 交叉验证法 (cross validation)
- 自助法 (bootstrap)

三、机器学习评估

如何获得测试结果?

- 留出法 (hold-out)
- 交叉验证法 (cross validation)
- 自助法 (bootstrap)



注意:

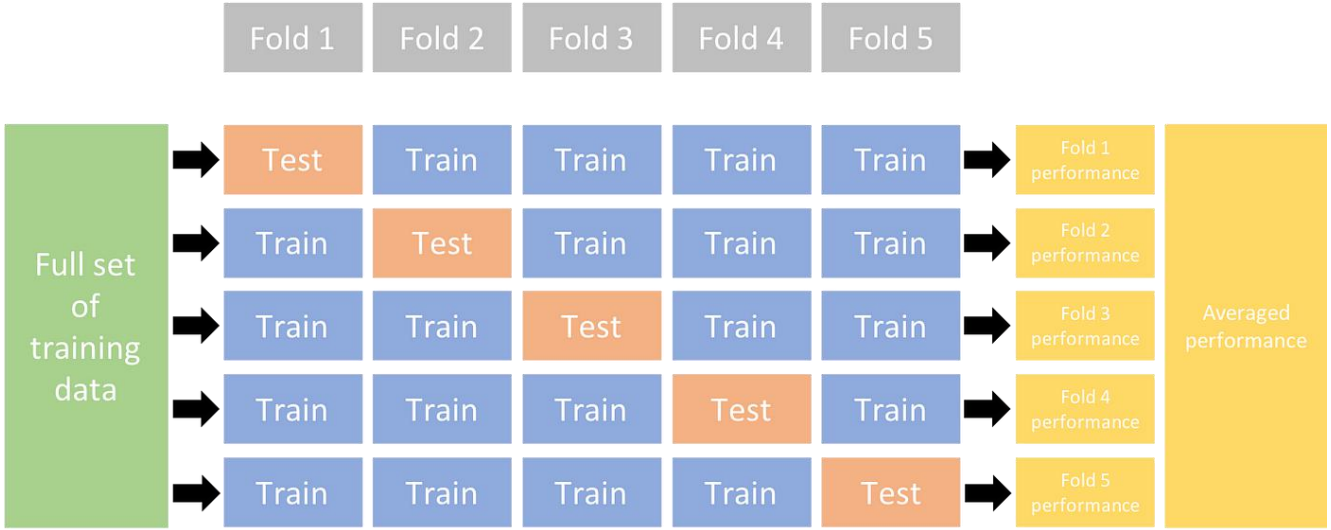
多次重复划分 (例如: 100次随机划分)

测试集不能太大、不能太小 (例如: 1/5~1/3)

三、机器学习评估

如何获得测试结果?

- 留出法 (hold-out)
- 交叉验证法 (cross validation)
- 自助法 (bootstrap)



三、机器学习评估

如何获得测试结果?

- MAE
- MSE/RMSE
- R^2

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

三、机器学习评估

如何获得测试结果?

- Accuracy
- Recall

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP
	Negative	FN	TN

期望
TP和**TN**越大越好
FN和**FP**越小越好

真实值=Positive，　预测值=Positive　(TP = True　Positive)　√
真实值=Positive，　预测值=Negative　(FN = False Negative)　×
真实值=Negative，　预测值=Positive　(FP = False　Positive)　×
真实值=Negative，　预测值=Negative　(TN = True　Negative)　√

三、机器学习评估

如何获得测试结果?

- Accuracy
- Recall

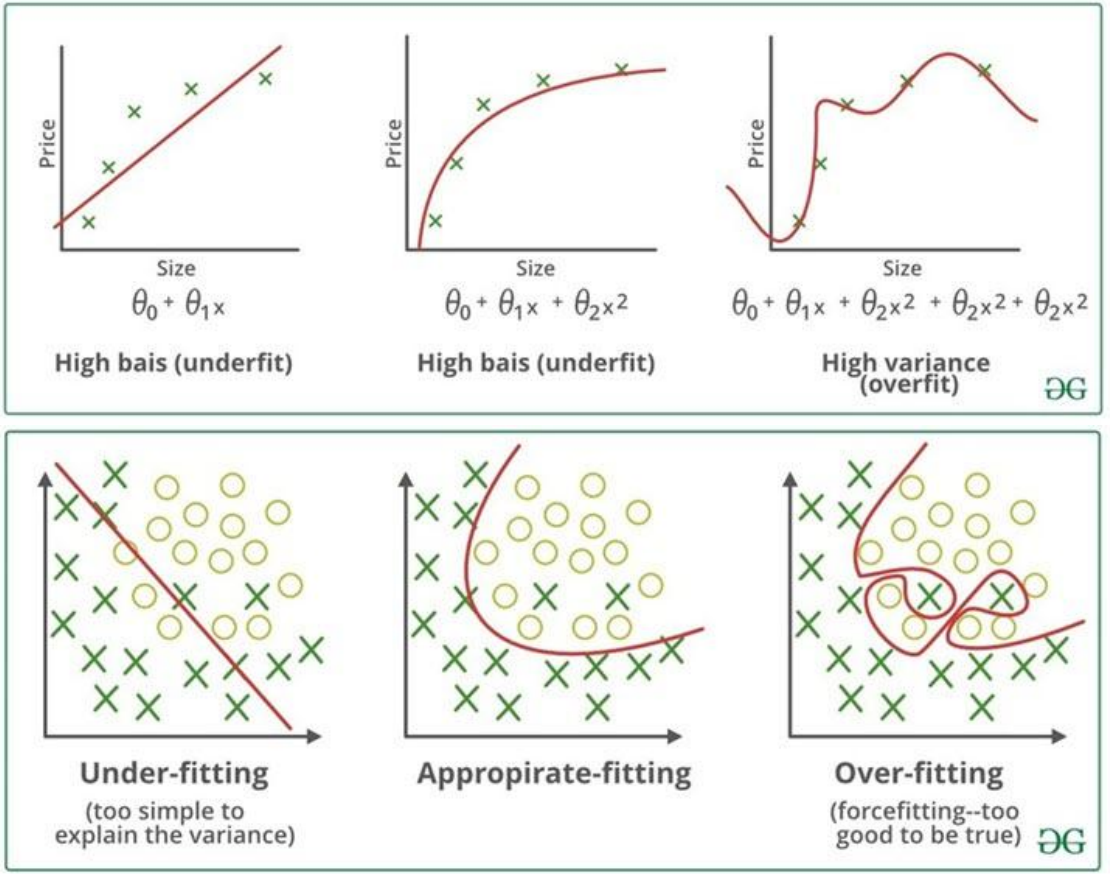
混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP
	Negative	FN	TN

	公式	意义
准确率(ACC) Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	模型正确分类样本数占总样本数比例(所有类别)
精确率(PPV) Positive Predictive Value	$\text{Precision} = \frac{TP}{TP + FP}$	模型预测的所有**positive**中，预测正确的比例
灵敏度/召回率(TPR) True Positive Rate	$\text{Reall} = \frac{TP}{TP + FN}$	所有真实**positive**中，模型预测正确的**positive**比例
特异度(TNR) True Negative Rate	$\text{Specificity} = \frac{TN}{TN + FP}$	所有真实**negative**中，模型预测正确的**negative**比例

三、机器学习评估

如何评估学习情况?

- 欠拟合(underfit)
- 过拟合(overfit)



三、机器学习评估

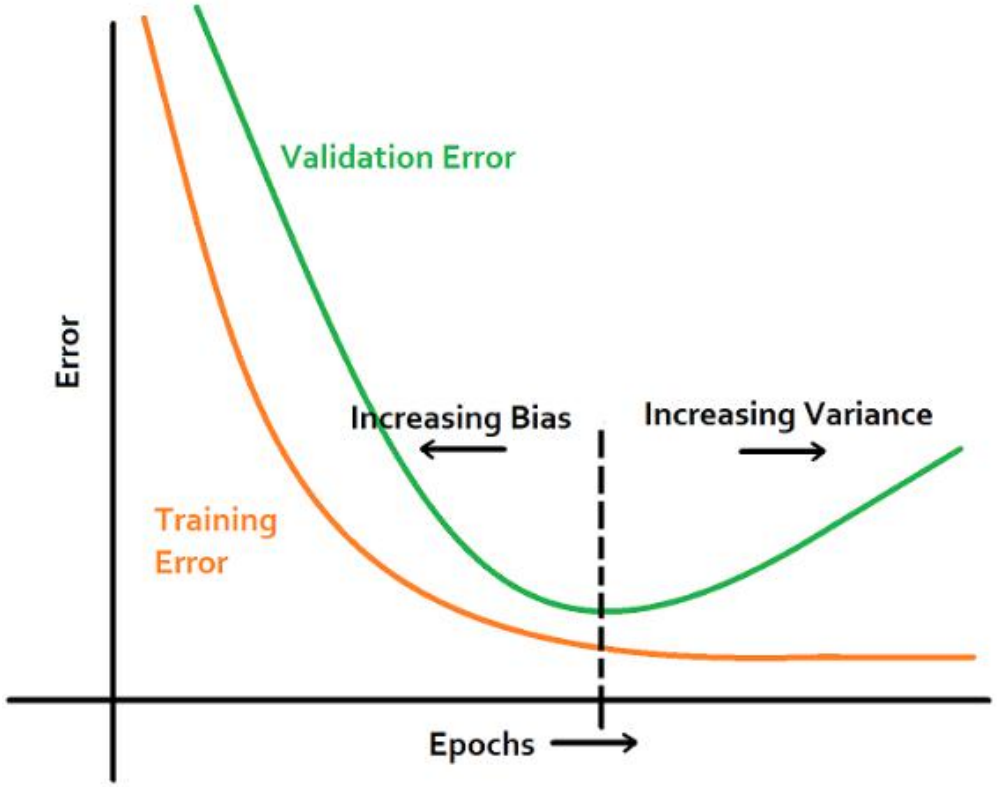
欠拟合

- 模型复杂化
- 增加更多的特征
- 增加训练数据往往没有用
- 降低正则化约束

三、机器学习评估

过拟合

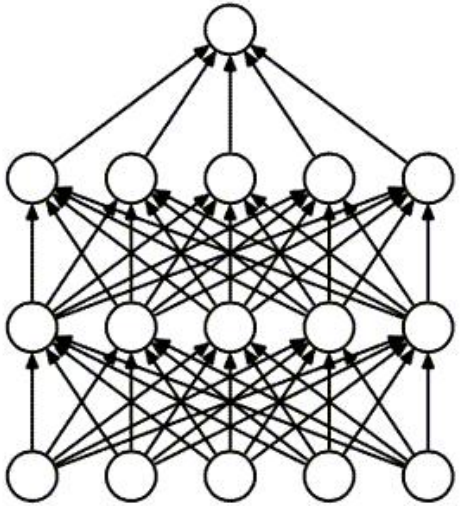
- 增加更多数据（最直接）
- 降低模型复杂度
- 降低特征的数量（如降维）
- Early stopping（提前终止）
- Dropout

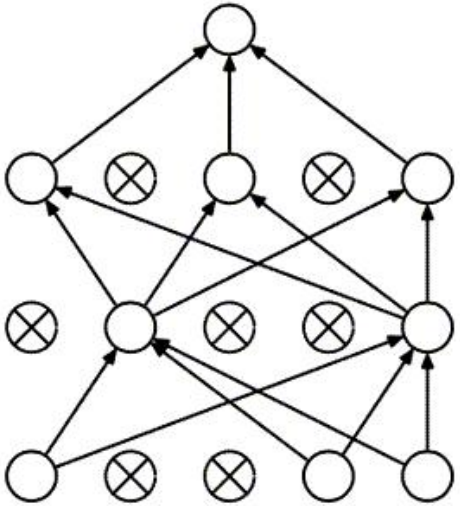


三、机器学习评估

过拟合

- 增加更多数据（最直接）
- 降低模型复杂度
- 降低特征的数量（如降维）
- Early stopping（提前终止）
- Dropout





THE END

THANKS

黄恩待

智能医学与生物医学工程研究院

huangendai@nbu.edu.cn