南昌大学
NANCHANG UNIVERSITY

统计机器学习

主讲人：彭振华

数学与计算机学院

2026年

成绩构成

总评成绩 = 平时成绩 * 30% + 期中成绩 * 10% + 期末成绩 * 60%（总评不低于期末卷面成绩）

- 平时成绩 = 考勤（30%） + 作业（40%） + 课堂表现（20%） + 课后讨论（10%）

- 缺勤1次扣10分/迟到早退每节课扣4分

  (不满1节课按1节算，每达3节课按缺勤记1次)

- 作业缺1次或抄袭扣10分/迟交1天扣2分，每次扣满10分为止

- 期末题型：选择题、填空题、简答题、计算题、综合题

个人简介——彭振华

数学与计算机学院 副教授

数学、统计学、计算机技术专业硕导

研究兴趣：

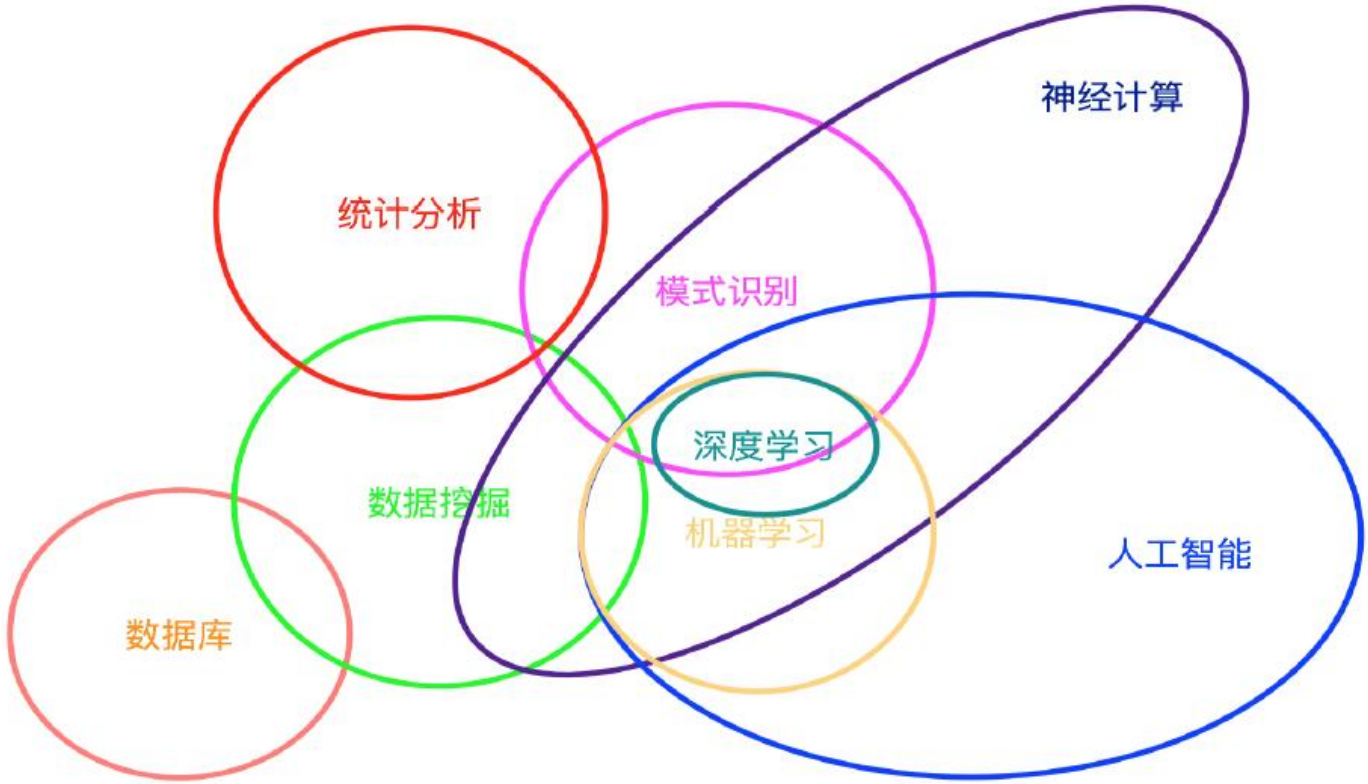
1. 非光滑非凸优化算法与理论（多目标规划、双层优化、DC规划、图像恢复、信道估计）
2. 智能决策（金融、交通、路径规划）
3. 机器学习（传统机器学习算法、神经网络算法、超参数学习）

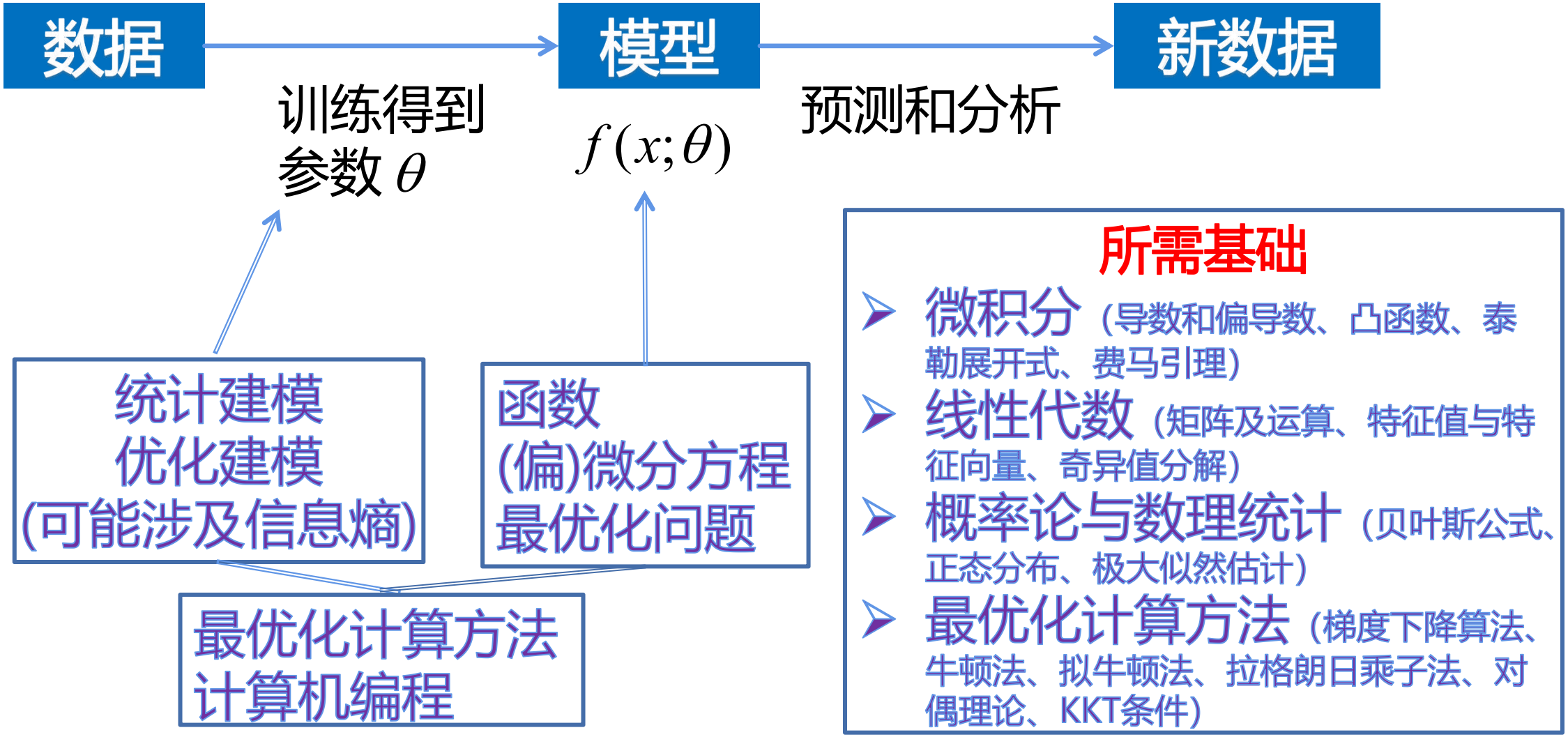
目录

CONTENTS

- **机器学习**：计算机利用已有的**数据**(经验)，得出了某种**模型**，并利用此模型**预测和分析**未来的一种方法。
- **换言之**，机器学习致力于研究如何通过**计算手段**，利用**经验**来改善系统自身的性能，从而在计算机上**从数据中产生“模型”**，用于对**新的情况**给出判断。
- **深度学习**：实现机器学习的一种技术
- **人工智能**：机器展现的人类智能
- **数据挖掘**：从大量的数据中通过统计、机器学习、专家系统和模式识别等诸多方法搜索隐藏于其中信息的过程。

数字、文字、图像、视频、音频





所需基础

- 微积分（导数和偏导数、凸函数、泰勒展开式、费马引理）
- 线性代数（矩阵及运算、特征值与特征向量、奇异值分解）
- 概率论与数理统计（贝叶斯公式、正态分布、极大似然估计）
- 最优化计算方法（梯度下降算法、牛顿法、拟牛顿法、拉格朗日乘子法、对偶理论、KKT条件）

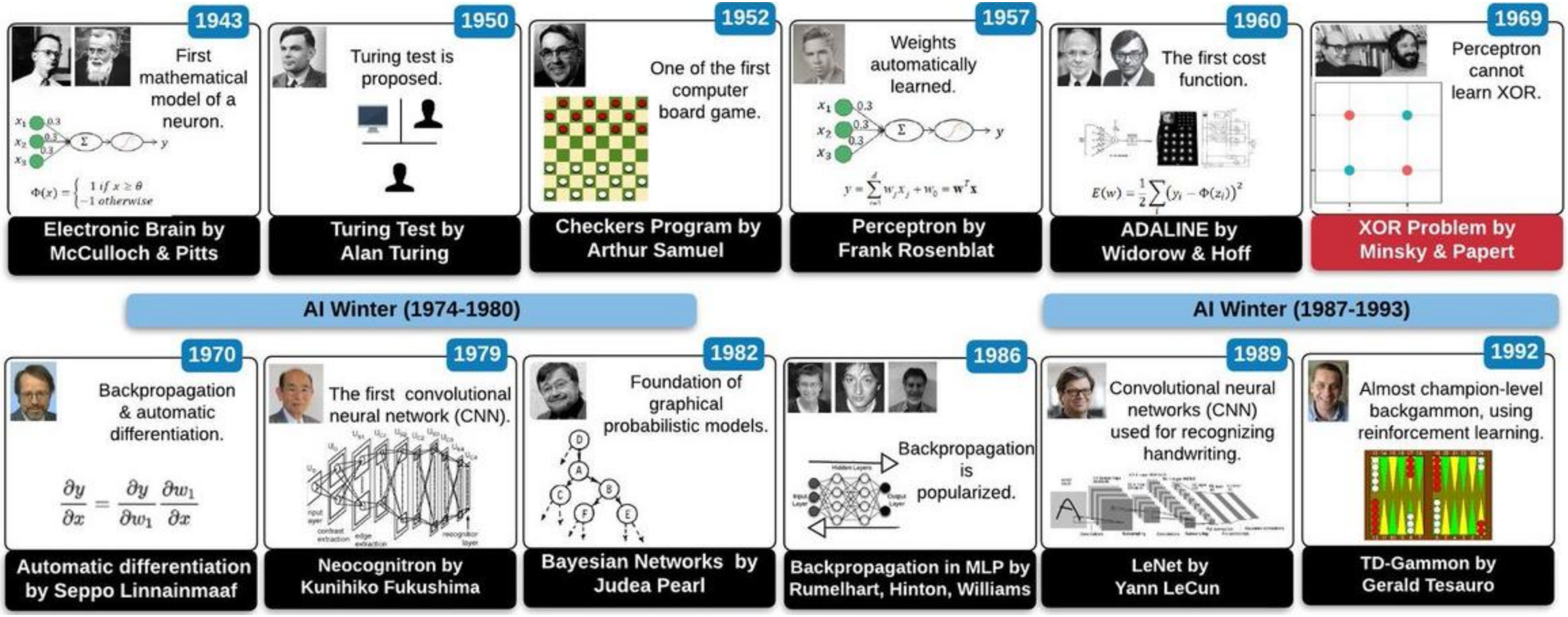
算法或理论	用到的数学知识点
贝叶斯分类器	随机变量，贝叶斯公式，随机变量独立性，正态分布，最大似然估计
决策树	概率，熵，Gini 系数
KNN 算法	距离函数
主成分分析	协方差矩阵，散布矩阵，拉格朗日乘数法，特征值与特征向量
流形学习	流形，最优化，测地线，测地距离，图，特征值与特征向量
线性判别分析	散度矩阵，逆矩阵，拉格朗日乘数法，特征值与特征向量
支持向量机	点到平面的距离，Slater 条件，强对偶，拉格朗日对偶，KKT 条件，凸优化，核函数，Mercer 条件
logistic	概率，随机变量，最大似然估计，梯度下降法，凸优化，牛顿法
随机森林	抽样，方差
AdaBoost 算法	概率，随机变量，极值定理，数学期望，牛顿法
隐马尔科夫模型	概率，离散型随机变量，条件概率，随机变量独立性，拉格朗日乘数法，最大似然估计
条件随机场	条件概率，数学期望，最大似然估计
高斯混合模型	正态分布，最大似然估计，Jensen 不等式

➢ 机器学习典型应用领域

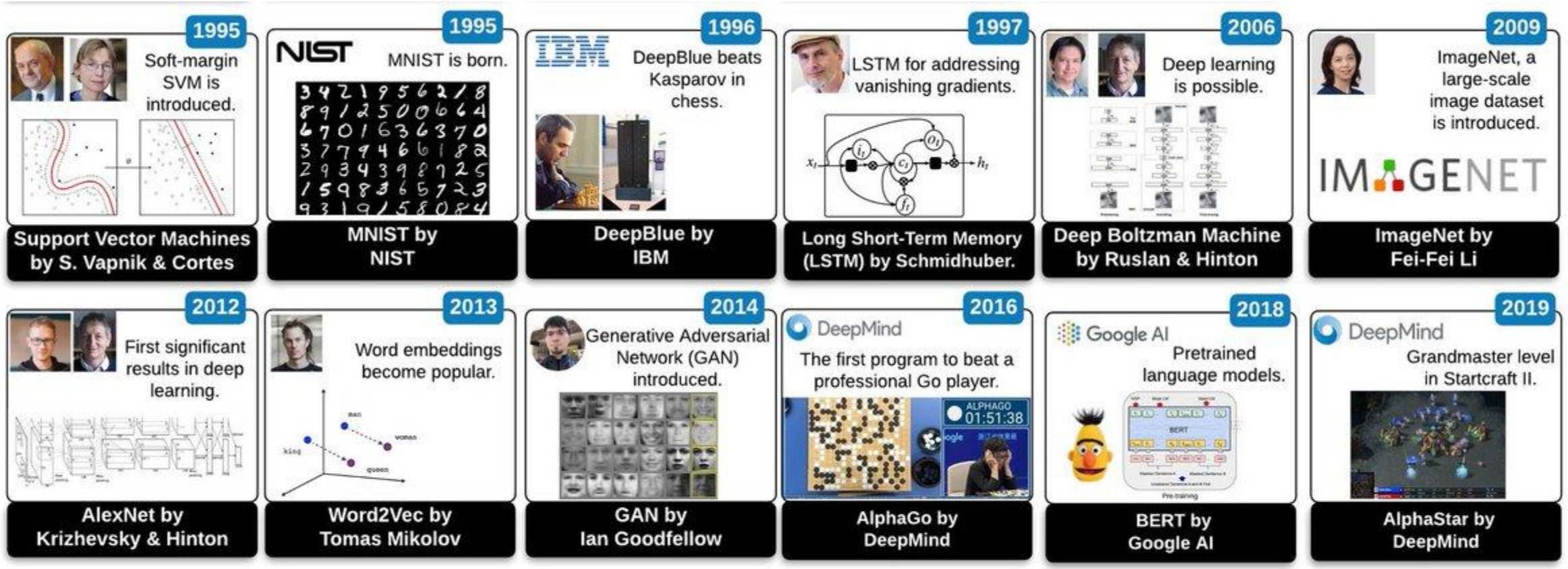
- 医疗健康
  （疾病诊断、个性化治疗、药物研发）
- 金融科技
  （风险控制、资产定价与投资、个性化服务）
- 智能交通与自动驾驶
  （交通流量预测、智能导航、自动驾驶）
- 智能制造
  （质量控制、设备维护）
- 自然语言处理
  （语音识别与翻译、情感分析与文本分类、智能家居控制）

➢ 机器学习发展史

总的来说，人工智能经历了**逻辑推理**、**知识工程**、**机器学习**三个阶段。



➢ 机器学习发展史



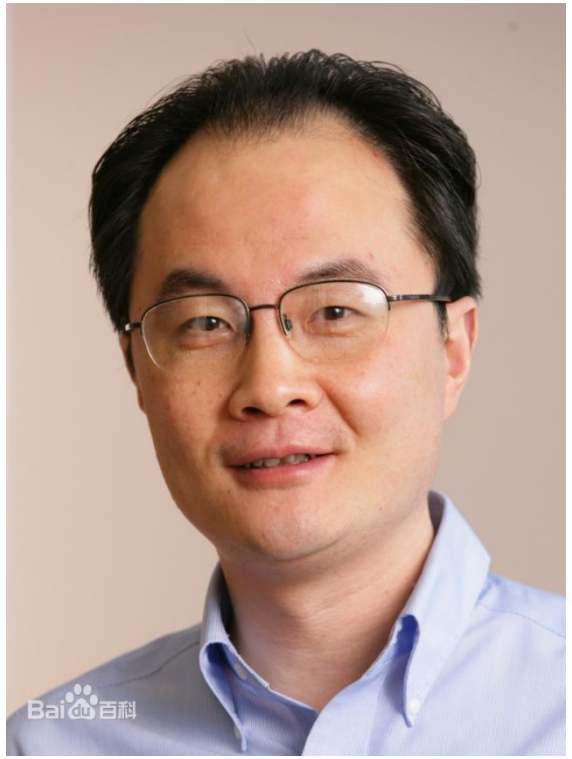
Parisa Rashidi, May 2020. CC BY 4.0

➢ 机器学习界的执牛耳者

杨立昆（Yann LeCun）
杰弗里·欣顿（Geoffrey Hinton）
本吉奥（ Bengio ）
共同获得了2018年计算机科学的最高奖项——**ACM图灵奖**。

Andrew Ng
中文名**吴恩达**，斯坦福大学副教授，前“百度大脑”的负责人与百度首席科学家。

➢ 机器学习界的国内泰斗



李航，现任字节跳动科技有限公司人工智能实验室总监，北京大学、南京大学客座教授，IEEE 会士，ACM 杰出科学家，CCF 高级会员。
代表作：《统计学习方法》

周志华，南京大学计算机科学与技术系主任 、人工智能学院院长。
代表作：《机器学习》（西瓜书）

徐宗本，中国科学院院士，西安交通大学数学与统计学院教授。提出了稀疏信息处理的L(1/2)正则化理论。发现并证明机器学习的“徐-罗奇”定理。

参考书籍

[1] Andrew Ng. Machine Learning[EB/OL]. StanfordUniversity,2014.https://www.coursera.org/course/ml

[2] 李航. 统计学习方法[M]. 北京: 清华大学出版社,2019.

[3] 周志华. 机器学习[M]. 北京: 清华大学出版社,2016.

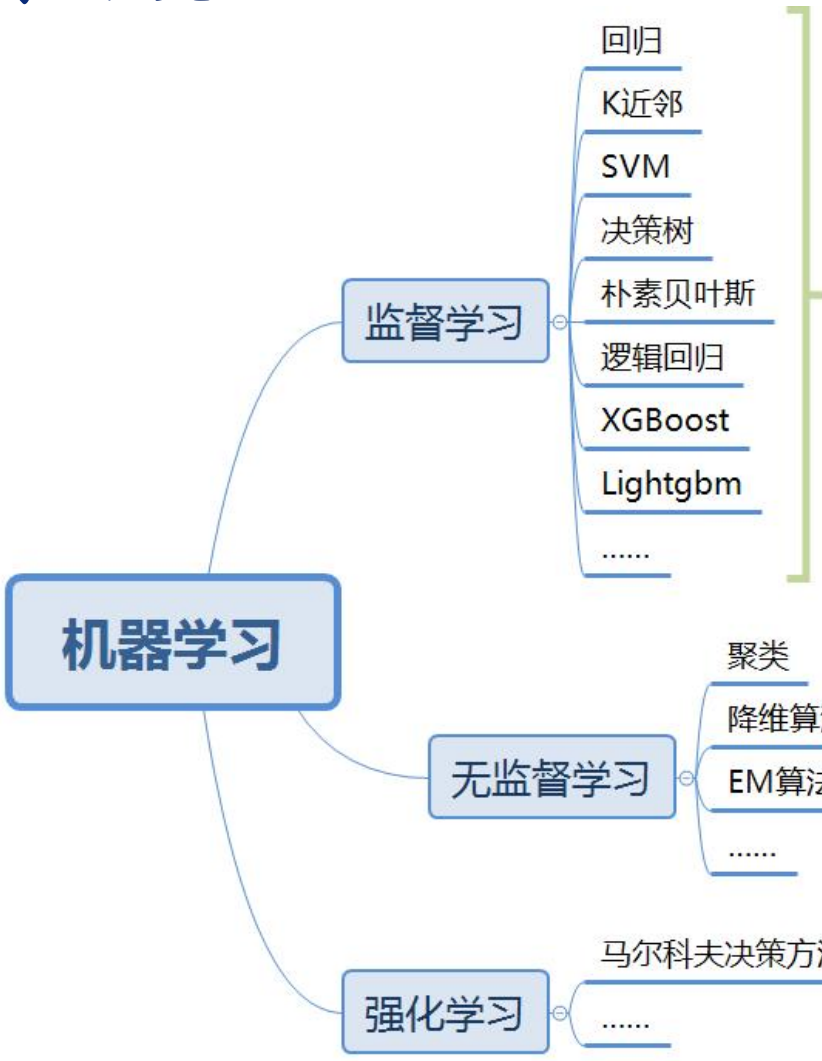
[4] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning[M]. New York: Springer,2001.

[5] CHRISTOPHER M. BISHOP. Pattern Recognition and Machine Learning[M]. New York: Springer,2006.

[6] Stephen Boyd, Lieven Vandenberghe, Convex Optimization[M]. Cambridge: Cambridge University Press, 2004.

[7] TOM M MICHELLE. Machine Learning[M]. New York: McGraw-Hill Companies,Inc,1997.

➢ 机器学习类型



- ✓ 分类（Classification）
  - ✓ 身高1.65m，体重100kg的男人肥胖吗?
  - ✓ 根据肿瘤的体积、患者的年龄来判断良性或恶性?
- ✓ 回归（Regression、Prediction）
  - ✓ 如何预测上海浦东的房价?
  - ✓ 未来的股票市场走向?
- ✓ 聚类（Clustering）
  - ✓ 如何将教室里的学生按爱好、身高划分为5类?
- ✓ 降维（ Dimensionality Reduction ）
  - ✓ 如何将将原高维空间中的数据点映射到低维度的空间中?
- ✓ 强化学习（Reinforcement Learning）
  - ✓ 用于描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题 。

➢ 监督学习

- Instance，feature vector，feature space
- 输入实例x的**特征向量**：

$$x=(x^{(1)},x^{(2)},\cdots,x^{(i)},\cdots,x^{(n)})^{\mathrm{T}}$$

- $x^{(i)}$与x_i 不同,后者表示多个输入变量中的第i个

$$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$$

- 训练集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

标签

- 输入变量和输出变量：
  - 分类问题、回归问题、标注问题

特征

标记

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否
1	青绿	蜷缩	沉闷	?

训练集

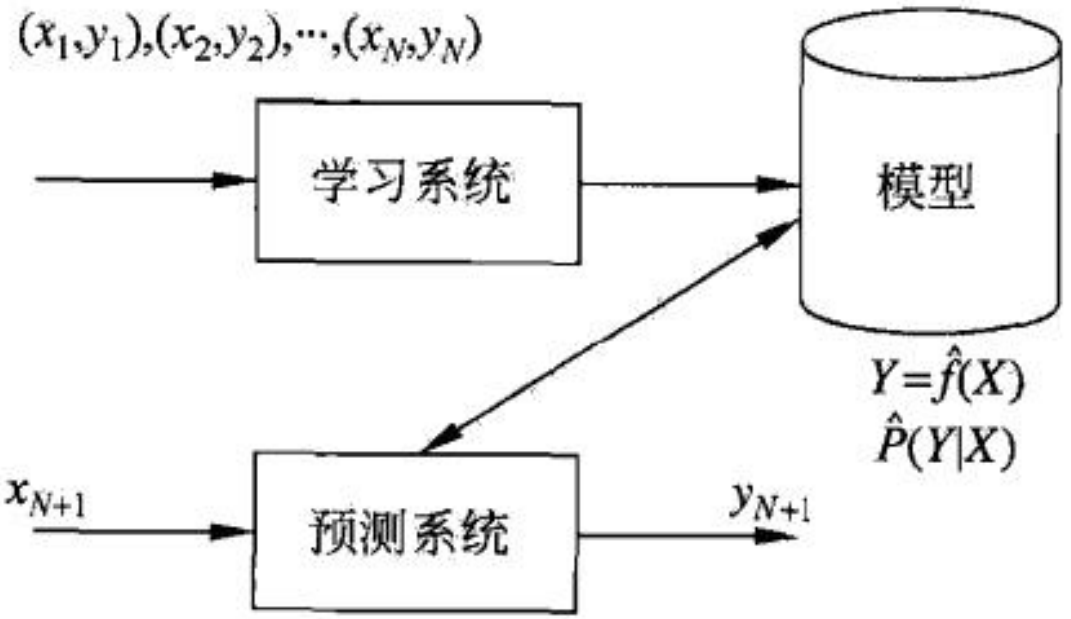
测试集

联合概率分布

假设输入与输出的随机变量X和Y遵循联合概率分布P(X,Y)，P(X,Y)为分布函数或分布密度函数。对于学习系统来说，联合概率分布是未知的，训练数据和测试数据被看作是依联合概率分布P(X,Y)独立同分布产生的。

假设空间

监督学习目的是学习一个由输入到输出的映射，称为模型模式的集合就是假设空间（hypothesis space），概率模型:条件概率分布P(Y|X), 决策函数：Y=f(X)



$$y_{N+1}=\arg\max_{y_{N+1}}\hat{P}(y_{N+1}\mid x_{N+1})$$

$$y_{N+1}=\hat{f}(x_{N+1})$$

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

1	青绿	蜷缩	沉闷	?

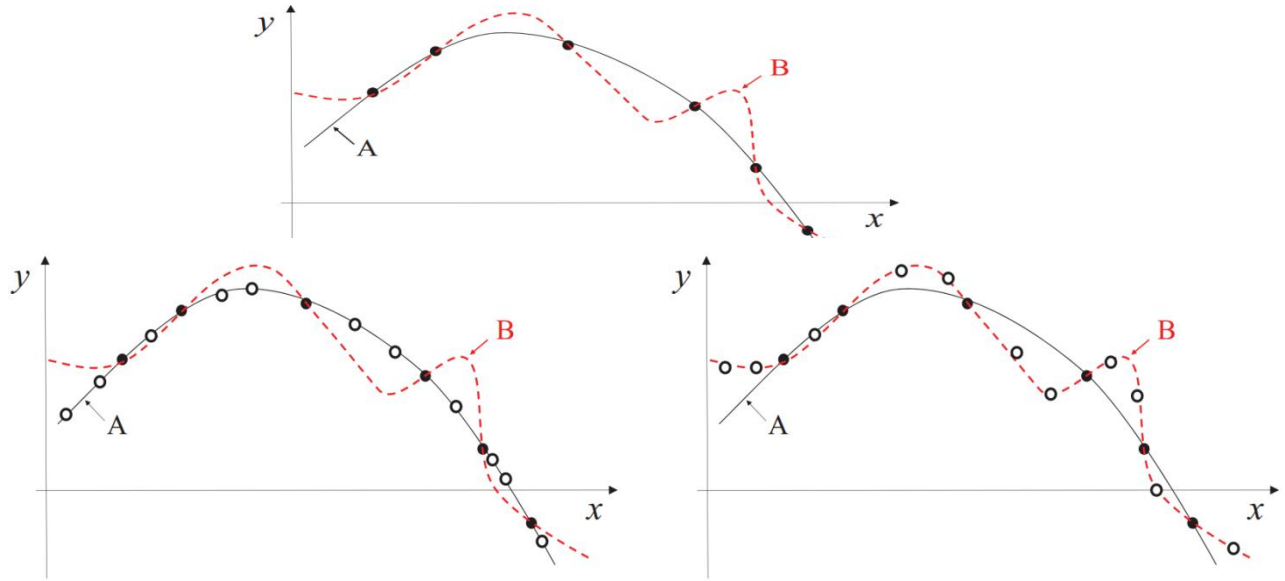
选取哪个假设作为学习模型？

(色泽=?)∧(根蒂=?)∧(敲声=?)↔好瓜

在模型空间中搜索不违背训练集的假设

假设空间大小：4*4*4+1=65

学习过程中对某种类型假设的偏好称作归纳偏好



(a) A 优于 B　　(b) B 优于 A

没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

一个算法ξ_a如果在某些问题上比另一个算法ξ_b好，必然存在另一些问题，ξ_b比ξ_a好,也即没有免费的午餐定理。

简单起见，假设样本空间$\mathcal{X}$和假设空间$\mathcal{H}$离散,令$P(h|X,\mathfrak{L}_a)$代表算法$\mathfrak{L}_a$基于训练数据X产生假设h的概率，再令f代表要学的目标函数，$\mathfrak{L}_a$在训练集之外所有样本上的总误差

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_h\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}\left(h\left(\boldsymbol{x}\right)\neq f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)$$

$\mathbb{I}(\cdot)$为指示函数，若$\cdot$为真取值1，否则取值0

考虑二分类问题，目标函数可以为任何函数$\mathcal{X}\mapsto\{0,1\}$，函数空间为$\{0,1\}^{|\mathcal{X}|}$，对所有可能$f$按均匀分布对误差求和,有：

$$\begin{aligned}\sum_f E_{ote}(\mathfrak{L}_a|X,f)&=\sum_f\sum_h\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a)\\&=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\sum_f\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\\&=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\frac{1}{2}2^{|\mathcal{X}|}\\&=\frac{1}{2}2^{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\\&=2^{|\mathcal{X}|-1}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\cdot 1\,.\end{aligned}$$

总误差与学习算法无关！

实际问题中，并非所有问题出现的可能性都相同。脱离具体问题，空谈“什么学习算法更好”毫无意义

➢ 无监督学习

- 训练集：

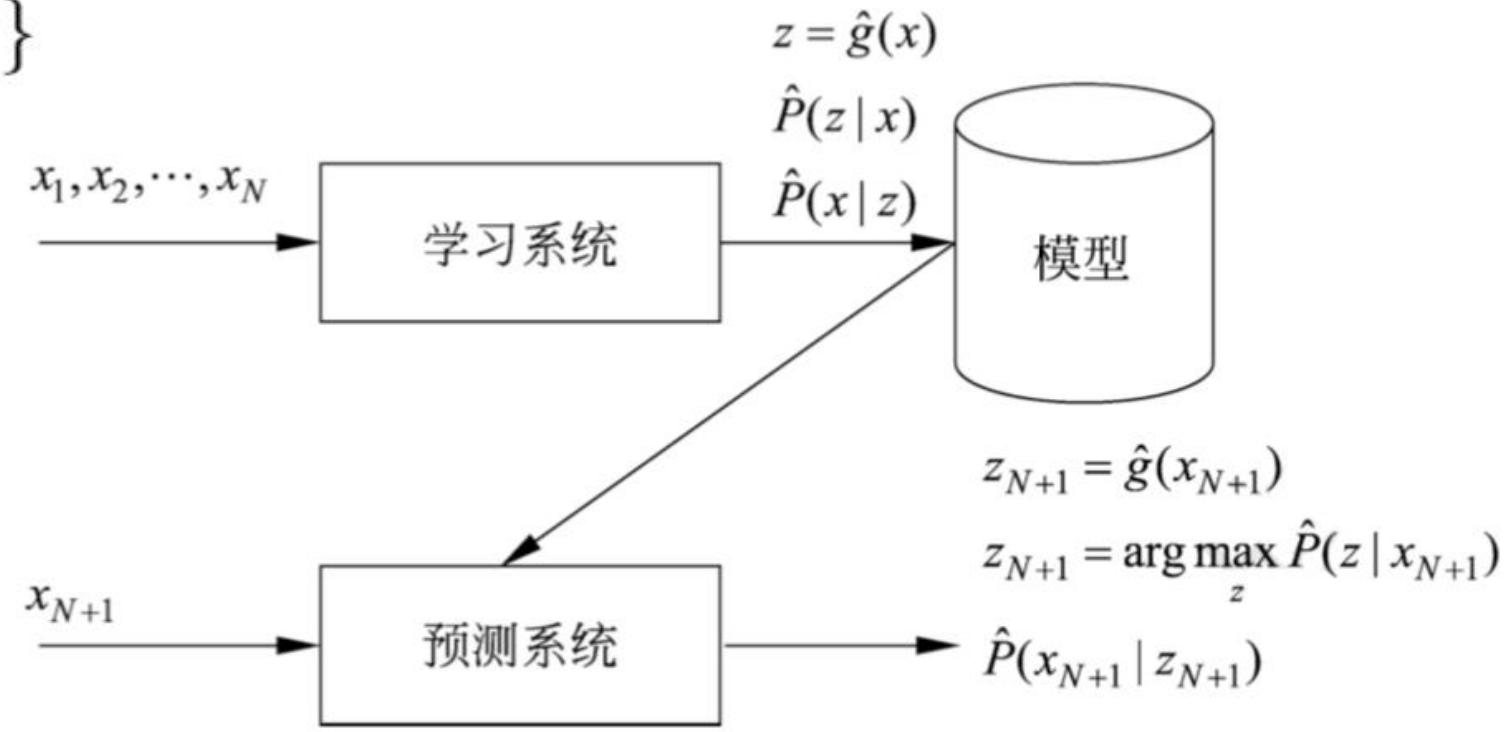
$$U = \{x_1, x_2, \cdots, x_N\}$$

- 模型函数：

$$z = g(x)$$

- 条件概率分布：

$$P(z|x)$$



➢ 机器学习的三要素

三要素：机器学习 = 模型 + 策略（**损失函数**） + 优化算法

- 机器学习模型的类别，大致有两种：一是概率模型和非概率模型。
- **监督学习**中，**概率模型**$P(y|x)$，**非概率模型**$y=f(x)$。其中，x输入，y输出。
- **无监督学习**中，**概率模型**$P(z|x)$，**非概率模型**$z=f(x)$。其中，x输入，z输出。
- 决策树、朴素贝叶斯、隐马尔科夫模型、高斯混合模型属于概率模型。感知机、支持向量机、KNN、AdaBoost、K-means以及神经网络均属于非概率模型。
- 对于非概率模型而言，按判别函数线性与否分线性模型与非线性模型。感知机、线性支持向量机、KNN、K-means是**线性模型**。核支持向量机、AdaBoost、神经网络属于**非线性模型**。

➢ 策略

- 错误率: 错分样本的占比: $E = a/m$
- 误差：样本真实输出与预测输出之间的差异（损失函数）
  - 训练(经验)误差：训练集上
  - 测试误差：测试集
  - 泛化误差：除训练集外所有样本

□ 损失函数：一次预测的好坏

□ 风险函数：平均意义下模型预测的好坏

1. 0-1损失函数 0-1 loss function $L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$

2. 平方损失函数 quadratic loss function $L(Y, f(X)) = (Y - f(X))^2$

3. 绝对损失函数 absolute loss function $L(Y, f(X)) = |Y - f(X)|$

4. 对数损失函数 logarithmic loss function 或对数似然损失函数 loglikelihood loss function

$$L(Y, P(Y \mid X)) = -\log P(Y \mid X)$$

5. 损失函数的期望

$$R_{\exp}(f) = E_P[L(Y, f(X))] = \int_{x \times y} L(y, f(x)) P(x, y) \mathrm{d}x\mathrm{d}y$$

6. 数据集

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

- 经验风险 empirical risk，经验损失 empirical loss

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

➢ 经验风险最小化与结构风险最小化

□ 经验风险最小化最优模型

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

当样本容量很小时，经验风险最小化学习的效果未必很好，会产生“过拟合”

□ 结构风险最小化 structure risk minimization，为防止过拟合提出的策略，等价于正则化（regularization），加入正则化项regularizer，或罚项penalty term：

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

算法：

如果最优化问题有显式的解析式，算法比较简单
但通常解析式不存在，就需要数值计算的方法

↳ 求最优模型就是求解最优化问题：

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

- 过拟合:

学习器把训练样本学习的“太好”，将训练样本本身的特点当做所有样本的一般性质，导致泛化性能下降

  - 优化目标加正则项
  - early stop

- 欠拟合:

对训练样本的一般性质尚未学好

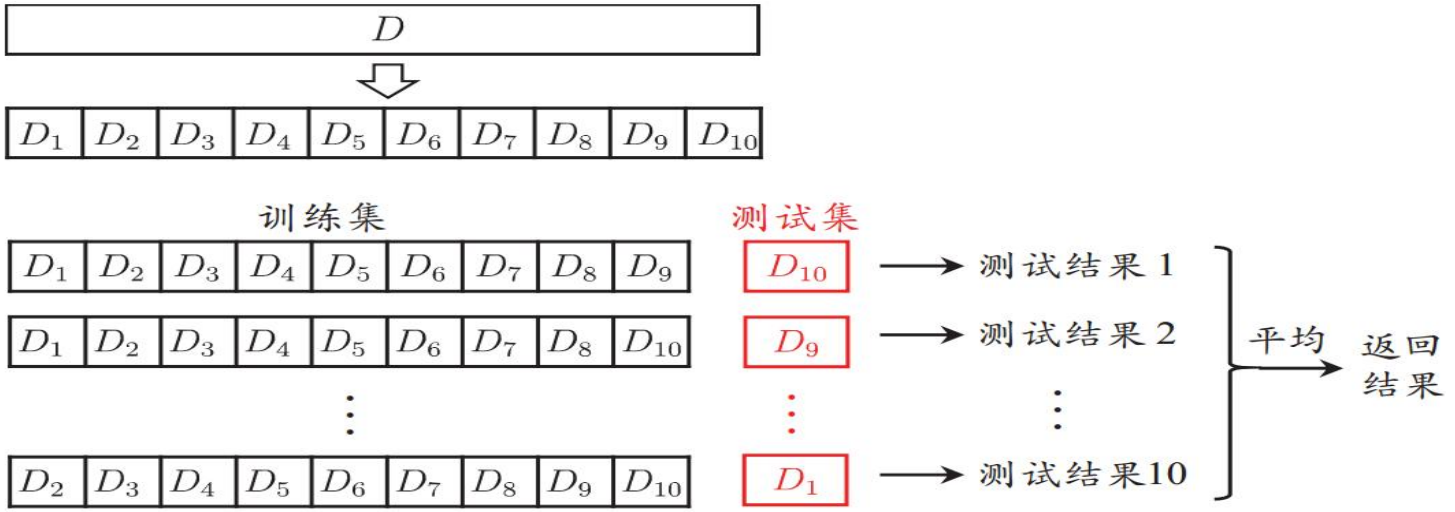
  - 决策树:拓展分支
  - 神经网络:增加训练轮数

通常将包含个m样本的数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
拆分成训练集S和测试集T

- 留出法：
  - 直接将数据集划分为两个互斥集合
  - 训练/测试集划分要尽可能保持数据分布的一致性
  - 一般若干次随机划分、重复实验取平均值
  - 训练/测试样本比例通常为2:1~4:1

- 交叉验证法：

将数据集分层采样划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值，k最常用的取值是10.



10 折交叉验证示意图

与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值，例如常见的“10次10折交叉验证”

- 自助法：

以自助采样法为基础，对数据集D有放回采样m次得到训练集D'，用$D \setminus D'$做测试集。

- 实际模型与预期模型都使用m个训练样本
- 约有1/3的样本没在训练集中出现
- 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处
- 自助法在数据集较小、难以有效划分训练/测试集时很有用；由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用。

性能度量是衡量模型泛化能力的评价标准，反映了任务需求；使用不同的性能度量往往会导致不同的评判结果

在预测任务中，给定样例集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$

评估学习器的性能 f 也即把预测结果 $f(\boldsymbol{x})$和真实标记比较.

回归任务最常用的性能度量是“均方误差”：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(\boldsymbol{x}_i)-y_i)^2$$

对于分类任务,错误率和精度是最常用的两种性能度量:

- 错误率：分错样本占样本总数的比例
- 精度：分对样本占样本总数的比例

分类错误率

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$$

精度

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)=y_i)\\ &= 1-E(f;D)\ .\end{aligned}$$

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比例，此时查准率和查全率比错误率和精度更适合。

统计真实标记和预测结果的组合可以得到“混淆矩阵”

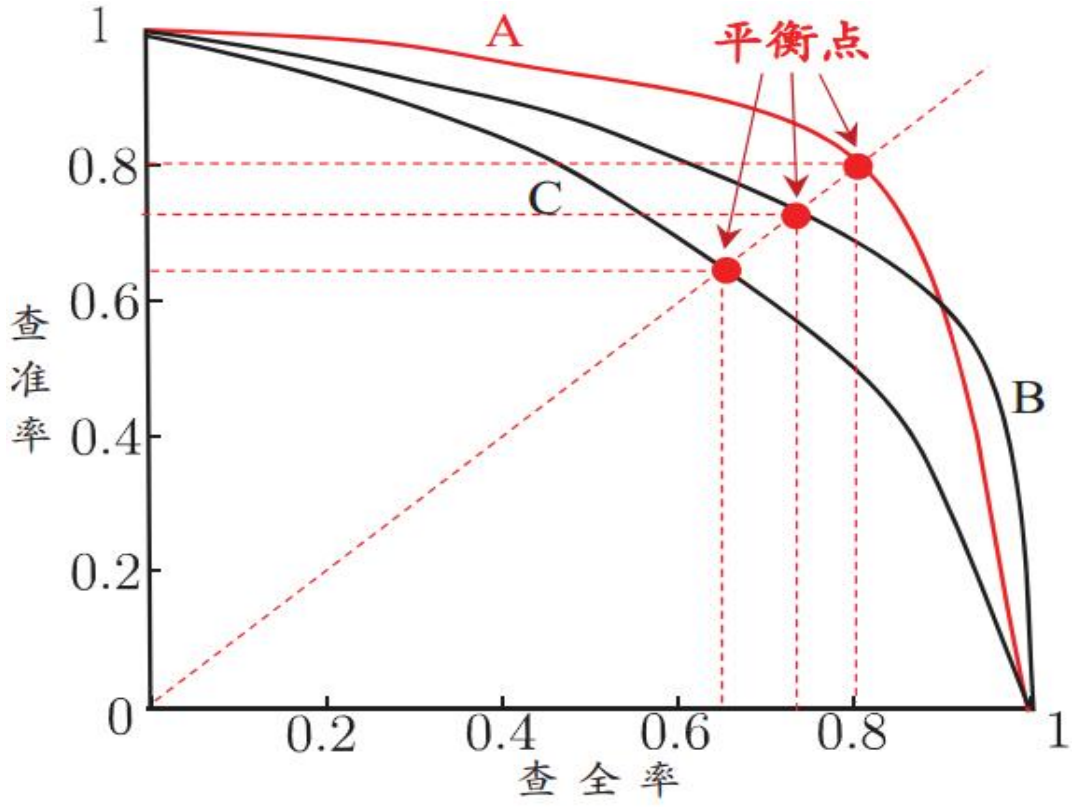
分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

查准率 $P = \frac{TP}{TP + FP}$

查全率 $R = \frac{TP}{TP + FN}$

根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”



P-R曲线与平衡点示意图

平衡点是曲线上“查准率=查全率”时的取值，可用来用于度量P-R曲线有交叉的分类器性能高低

比P-R曲线平衡点更用常用的是F1度量:

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{样例总数 + TP - TN}$$

比F1更一般的形式 F_β,

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$\beta = 1$: 标准F1

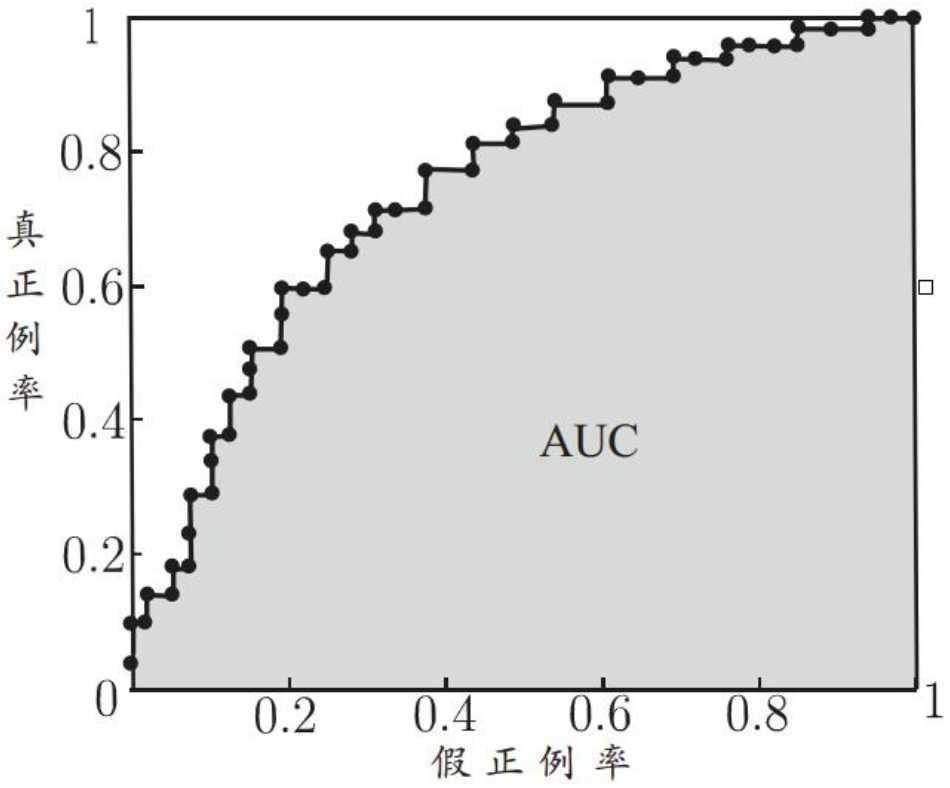
$\beta > 1$: 偏重查全率(逃犯信息检索)

$\beta < 1$: 偏重查准率(商品推荐系统)

类似P-R曲线，根据学习器的预测结果对样例排序，并逐个作为正例进行预测，以“假正例率”为横轴，“真正例率”为纵轴可得到ROC曲线,全称“受试者工作特征”.

ROC图的绘制：给定m^+个正例和m^-个负例，根据学习器预测结果对样例进行排序，将分类阈值设为每个样例的预测值，当前标记点坐标为(x, y),当前若为真正例，则对应标记点的坐标为$(x, y + \frac{1}{m^+})$;当前若为假正例，则对应标记点的坐标为$(x + \frac{1}{m^-}, y)$,然后用线段连接相邻点.

若某个学习器的ROC曲线被另一个学习器的曲线“包住”，则后者性能优于前者；否则如果曲线交叉，可以根据ROC曲线下面积大小进行比较，也即AUC值.



基于有限样例绘制的 ROC 曲线与 AUC

假设ROC曲线由$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$的点按序连接而形成$(x_1 = 0, x_m = 1)$，则AUC可估算为

$$\mathrm{AUC} = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1})$$

AUC衡量了样本预测的排序质量。

现实任务中不同类型的错误所造成的后果很可能不同，为了权衡不同类型错误所造成的不同损失，可为错误赋予“非均等代价”。

以二分类为例，可根据领域知识设定“代价矩阵”，如下表所示，其中$cost_{ij}$表示将第i类样本预测为第j类样本的代价。损失程度越大，$cost_{01}$与$cost_{10}$值的差别越大。

在非均等代价下，不再最小化错误次数，而是最小化“总体代价”，则“代价敏感”错误率相应的为

$$E(f;D;cost)=\frac{1}{m}\left(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{01}+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{10}\right)$$

在非均等代价下，ROC曲线不能直接反映出学习器的期望总体代价，而“代价曲线”可以。

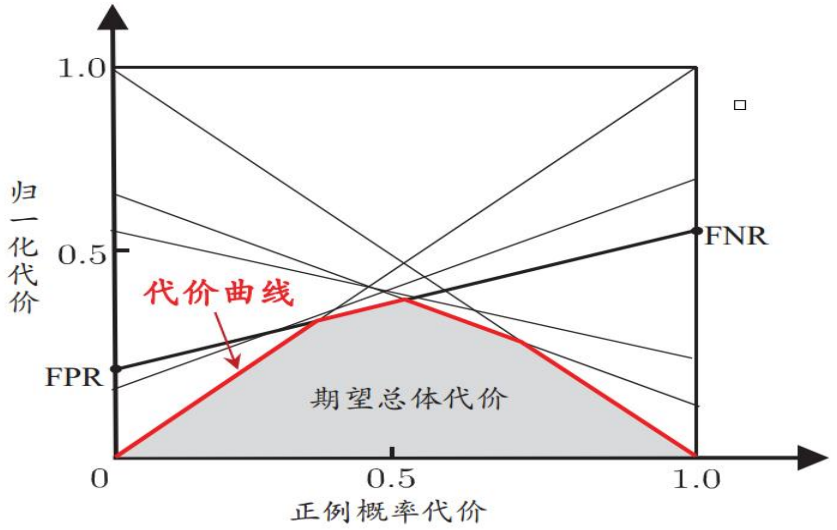
代价曲线的横轴是取值为[0,1]的正例概率代价

$$P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

纵轴是取值为[0,1]的归一化代价

$$cost_{norm} = \frac{\mathrm{FNR} \times p \times cost_{01} + \mathrm{FPR} \times (1-p) \times cost_{10}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

代价曲线图的绘制：ROC曲线上每个点对应了代价曲线上的一条线段，设ROC曲线上点的坐标为(FPR,TPR),则可相应计算出FNR,然后在代价平面上绘制一条从(0,FPR)到(1,FNR)的线段，线段下的面积即表示了该条件下的期望总体代价；如此将ROC曲线上的每个点转化为代价平面上的一条线段，然后取所有线段的下界，围成的面积即为所有条件下学习器的期望总体代价。



代价曲线与期望总体代价

通过实验可以估计学习算法的泛化性能，而“偏差-方差分解”可以用来帮助解释泛化性能。偏差-方差分解试图对学习算法期望的泛华错误率进行拆解。

对测试样本x,令y_D为x在数据集中的标记,y为x的真实标记,$f(\boldsymbol{x};D)$为训练集D上学得模型f在x上的预测输出。以回归任务为例：学习算法的期望预期为：

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x};D)]$$

使用样本数目相同的不同训练集产生的方差为

$$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right]$$

噪声为

$$\varepsilon^2 = \mathbb{E}_D\left[\left(y_D - y\right)^2\right]$$

期望输出与真实标记的差别称为偏差，即 $bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$ 为便与讨论，假定噪声期望为0，也即 $\mathbb{E}_D[y_D - y] = 0$，对泛化误差分解

$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&\quad + \mathbb{E}_D\left[2\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y + y - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)^2\right] + \mathbb{E}_D\left[\left(y - y_D\right)^2\right] \\
&\quad + 2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)\left(y - y_D\right)\right]
\end{aligned}
$$

又由假设中噪声期望为0，可得

$$E(f;D)=\mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right)-\bar{f}\left(\boldsymbol{x}\right)\right)^2\right]+\left(\bar{f}\left(\boldsymbol{x}\right)-y\right)^2+\mathbb{E}_D\left[\left(y_D-y\right)^2\right]$$

于是：

$$E(f;D)=bias^2\left(\boldsymbol{x}\right)+var\left(\boldsymbol{x}\right)+\varepsilon^2$$

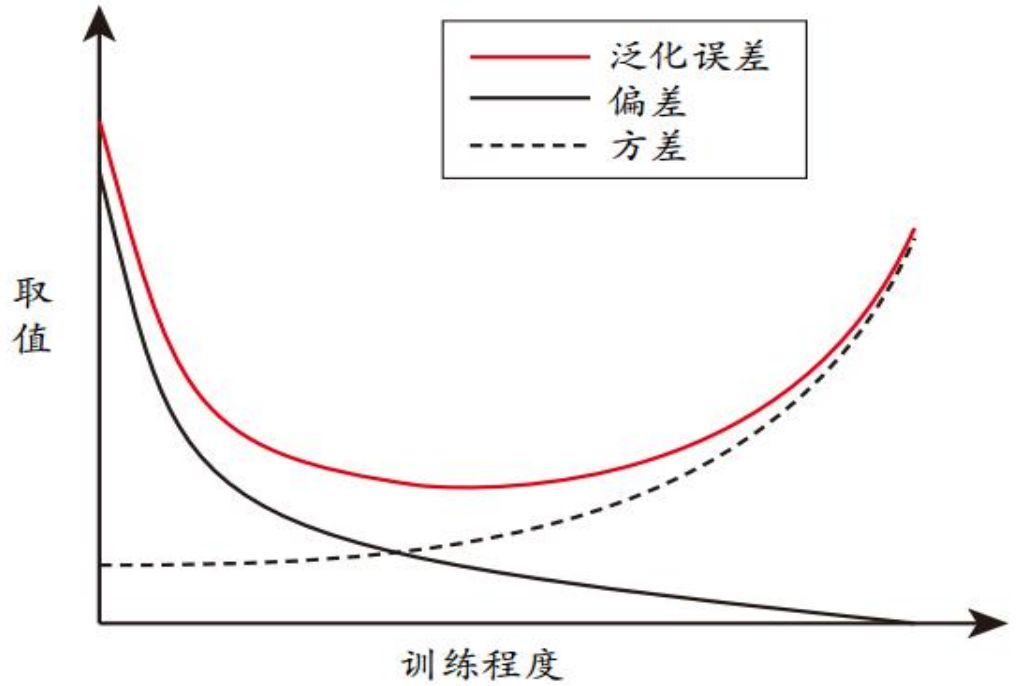
也即泛化误差可分解为偏差、方差与噪声之和。

- 偏差度量了学习算法期望预测与真实结果的偏离程度；即刻画了学习算法本身的拟合能力；
- 方差度量了同样大小训练集的变动所导致的学习性能的变化；即刻画了数据扰动所造成的影响；
- 噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界；即刻画了学习问题本身的难度。

泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度所共同决定的。给定学习任务为了取得好的泛化性能，需要使偏差小(充分拟合数据)而且方差较小(减少数据扰动产生的影响)。

一般来说，偏差与方差是有冲突的，称为偏差-方差窘境。如右图所示，假如我们能控制算法的训练程度：

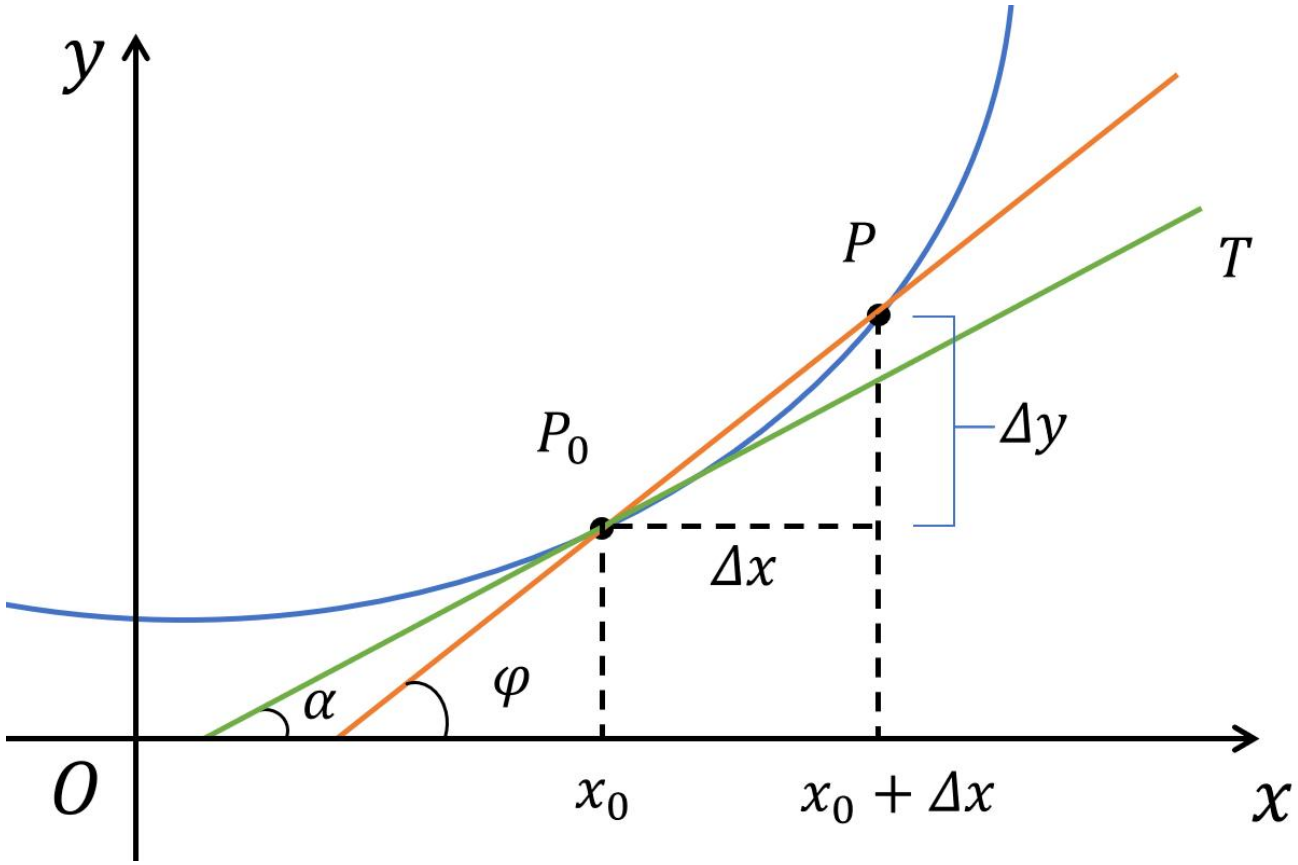
- 在训练不足时，学习器拟合能力不强，训练数据的扰动不足以使学习器的拟合能力产生显著变化，此时偏差主导泛化错误率；
- 随着训练程度加深，学习器拟合能力逐渐增强，方差逐渐主导泛化错误率；
- 训练充足后，学习器的拟合能力非常强，训练数据的轻微扰动都会导致学习器的显著变化，若训练数据自身非全局特性被学到则会发生过拟合。



泛化误差与偏差、方差的关系示意图

➢ 机器学习中的数学基础-微积分

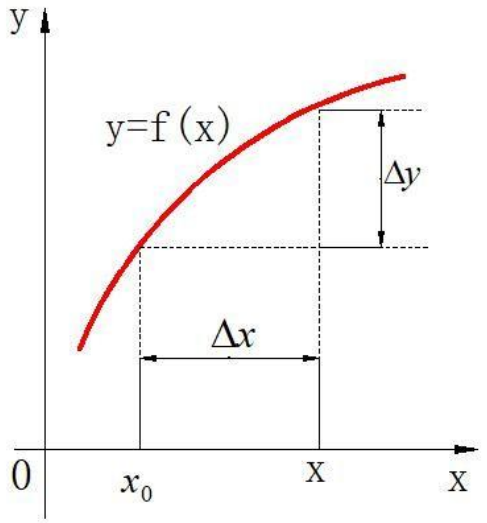
导数(Derivative)，也叫导函数值。又名微商，是微积分中的重要基础概念。当函数$y = f(x)$的自变量x在一点x_0上产生一个增量Δx时，函数输出值的增量Δy与自变量增量Δx的比值在Δx趋于0时的极限a如果存在，a即为在x_0处的导数，记作$f'(x_0)$。



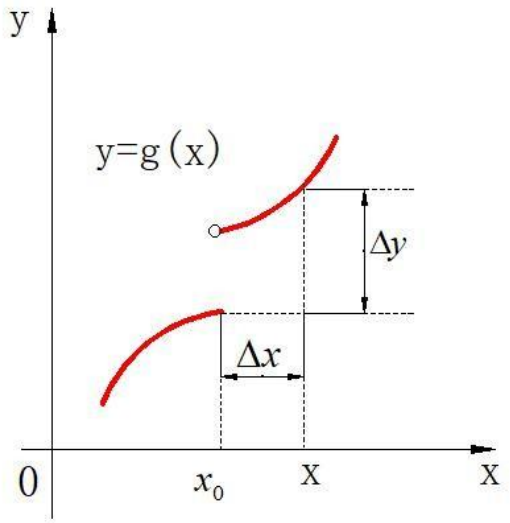
➢ 机器学习中的数学基础-微积分

设函数 $y=f(x)$ 在点 x_0 的某邻域内有定义，如果当自变量的改变量 Δx 趋近于零时，相应函数的改变量 Δy 也趋近于零，则称 $y=f(x)$ 在点 x_0 处连续。

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



当 $\Delta x\to 0$ 时，$\Delta y\to 0$；



当 $\Delta x\to 0^+$ 时，Δy 不能趋近于 **0**

机器学习中的数学基础-微积分

函数$f(x)$在点x_0处连续，需要满足的条件：

1. 函数在该点处有定义
2. 函数在该点处极限 $\lim_{x\to x_0} f(x)$ 存在
3. 极限值等于函数值 $f(x_0)$

如果平均变化率的极限存在，$\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$

则称此极限为函数$y=f(x)$ 在点x_0 处的导数，$f'(x_0)$

$y'|_{x=x_0}, \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_0}$ 或 $\left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=x_0}$

机器学习中的数学基础-微积分

(1) $y=c$（常数） 则： $y'=0$

(2) $y=x^{\alpha}$(α为实数) 则： $y'=\alpha x^{\alpha-1}$

(3) $y=a^{x}$ 则： $y'=a^{x}\ln a$ 特例: $(e^{x})'=e^{x}$

(4) $y=\log_a x$ 则： $y'=\frac{1}{x\ln a}$ ， 特例 $(\ln x)'=\frac{1}{x}$

(5) $y=\sin x$ 则： $y'=\cos x$

(6) $y=\cos x$ 则： $y'=-\sin x$

(7) $y=\tan x$ 则： $y'=\frac{1}{\cos^2 x}=\sec^2 x$

(8) $y=\cot x$ 则： $y'=-\frac{1}{\sin^2 x}=-\csc^2 x$

(9) $y=\sec x$ 则： $y'=\sec x\tan x$

(10) $y=\csc x$ 则： $y'=-\csc x\cot x$

(11) $y=\arcsin x$ 则： $y'=\frac{1}{\sqrt{1-x^2}}$

(12) $y=\arccos x$ 则： $y'=-\frac{1}{\sqrt{1-x^2}}$

(13) $y=\arctan x$ 则： $y'=\frac{1}{1+x^2}$

(14) $y=\operatorname{arccot} x$ 则： $y'=-\frac{1}{1+x^2}$

(15) $y=shx$ 则： $y'=chx$， (16) $y=chx$ 则： $y'=shx$

➢ 机器学习中的数学基础-微积分

四则运算法则

设函数$u=u(x)$，$v=v(x)$在点x可导，则：

(1) $(u\pm v)'=u'\pm v'$

(2) $(uv)'=uv'+vu'$　　$\mathrm{d}(uv)=u\mathrm{d}v+v\mathrm{d}u$

(3) $(\frac{u}{v})'=\frac{vu'-uv'}{v^2}(v\neq 0)$　　$\mathrm{d}(\frac{u}{v})=\frac{v\mathrm{d}u-u\mathrm{d}v}{v^2}$

➢ 机器学习中的数学基础-微积分

$$\frac{\mathrm{d}x^T}{\mathrm{d}x}=I \qquad \frac{\mathrm{d}x}{\mathrm{d}x^T}=I \qquad \frac{\mathrm{d}x^{\mathrm{T}}A}{\mathrm{d}x}=A$$

A为$n\times n$的矩阵，x为$n\times 1$的列向量

$$\frac{\mathrm{d}Ax}{\mathrm{d}x^{\mathrm{T}}}=A \qquad \frac{\mathrm{d}Ax}{\mathrm{d}x}=A^{\mathrm{T}} \qquad \frac{\mathrm{d}xA}{\mathrm{d}x}=A^{\mathrm{T}}$$

$$\frac{\partial u}{\partial x^{\mathrm{T}}}=\left(\frac{\partial u^{\mathrm{T}}}{\partial x}\right)^{\mathrm{T}} \qquad \frac{\partial u^{\mathrm{T}}v}{\partial x}=\frac{\partial u^{\mathrm{T}}}{\partial x}v+\frac{\partial v^{\mathrm{T}}}{\partial x}u^{\mathrm{T}} \qquad \frac{\partial uv^{\mathrm{T}}}{\partial x}=\frac{\partial u}{\partial x}v^{\mathrm{T}}+u\frac{\partial v^{\mathrm{T}}}{\partial x}$$

$$\frac{\mathrm{d}x^{\mathrm{T}}x}{\mathrm{d}x}=2x \qquad \frac{\mathrm{d}x^{\mathrm{T}}Ax}{\mathrm{d}x}=(A+A^{\mathrm{T}})x \qquad \frac{\mathrm{d}x^{\mathrm{T}}Ax}{\mathrm{d}x}=2Ax \ (\text{如果}A\text{为对称阵})$$

$$\frac{\partial AB}{\partial x}=\frac{\partial A}{\partial x}B+A\frac{\partial B}{\partial x} \qquad \frac{\partial u^{\mathrm{T}}Xv}{\partial X}=uv^{\mathrm{T}}$$

$$\frac{\partial u^{\mathrm{T}}X^{\mathrm{T}}Xu}{\partial X}=2Xuu^{\mathrm{T}} \qquad \frac{\partial[(Xu-v)^{\mathrm{T}}(Xu-v)]}{\partial X}=2(Xu-v)u^{\mathrm{T}}$$

➢ 机器学习中的数学基础-概率论与数理统计基础

(1) Bayes公式: $P(B_j|A)=\frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n}P(A|B_i)P(B_i)},j=1,2,\cdots,n$

(2) 乘法公式: $P(A_1A_2)=P(A_1)P(A_2|A_1)=P(A_2)P(A_1|A_2)$

$P(A_1A_2\cdots A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$

(1) 0-1分布:$P(X=k)=p^k(1-p)^{1-k},k=0,1$

(2) 二项分布:$B(n,p)$: $P(X=k)=C_n^kp^k(1-p)^{n-k},k=0,1,\cdots,n$

(3) Poisson分布:$p(\lambda)$: $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\lambda>0,k=0,1,2\cdots$

Poisson分布的期望和方差都等于参数λ

➢ 机器学习中的数学基础-概率论与数理统计基础

(4) 均匀分布$U(a,b)$: $f(x)=\begin{cases}\frac{1}{b-a}, a<x<b\\ 0,\end{cases}$

(5) 正态分布:$N(\mu,\sigma^2)$: $\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \sigma>0, -\infty<x<+\infty$

(6)指数分布:$E(\lambda)$: $f(x)=\begin{cases}\lambda e^{-\lambda x}, x>0, \lambda>0\\ 0,\end{cases}$

数学期望

离散型: $P\{X=x_i\}=p_i, E(X)=\sum_i x_i p_i$　　连续型: $X\sim f(x), E(X)=\int_{-\infty}^{+\infty} xf(x)dx$

➢ 机器学习中的数学基础-概率论与数理统计基础

求极大似然估计的一般步骤

(1) 构造似然函数$L(\theta)$

若总体X是离散型随机变量，其分布律为 $P(X=x)=p(x,\theta)$ 其中θ为未知参数

设$X_1,X_2,\cdots,X_n$是来自总体X的一个样本，而$x_1, x_2,\cdots, x_n$为$X_1,X_2,\cdots,X_n$的一个样本值，那么称

$$L(\theta)=L(x_1,\cdots,x_n,\theta)=P(X_1=x_1,\cdots,X_n=x_n)=\prod_{i=1}^{n}p(x_i,\theta)$$

为参数θ的似然函数。

➢ 机器学习中的数学基础-概率论与数理统计基础

若总体X是连续型随机变量，其概率密度为$f(x,\theta)$，$x_1, x_2,\cdots, x_n$为$X_1,X_2,\cdots,X_n$的一个样本值，则参数θ的似然函数为

$$L(\theta)=L(x_1,\cdots,x_n,\theta)=\prod_{i=1}^{n} f(x_i,\theta)$$

(2) 求似然函数$L(\theta)$的最大值点　挑选使$L(\theta)$达到最大的参数$\hat{\theta}$，作为θ的估计

$$L(x_1,\cdots,x_n,\hat{\theta})=\max L(x_1,\cdots,x_n,\theta)$$

一般地，参数θ的极大似然估计值可由下式求得

$$\frac{dL(\theta)}{d\theta}=0 \text{ 或者 } \frac{d\ln L(\theta)}{d\theta}=0$$

似然方程

似然方程组

$$\frac{\partial L(\theta_1,\cdots,\theta_k)}{\partial\theta_i}=0 \text{ 或者 } \frac{\partial \ln L(\theta_1,\cdots,\theta_k)}{\partial\theta_i}=0 \quad (i=1,\cdots,k)$$

当未知参数可以不止一个时，例如$\theta_1,\cdots,\theta_k$，那么可由下述方程组求得

➢ 机器学习中的数学基础-线性代数与优化

线性代数

- 矩阵的运算与秩
- 矩阵分解：LU分解、QR正交分解、奇异值分解、谱分解、非负矩阵分解

优化

- 优化模型、凸规划、范数、梯度、最优性条件、对偶理论
- 梯度下降算法、SGD、动量加速、AdaGrad、RMSProp、Adam
- 牛顿算法与拟牛顿算法
- 分块坐标下降法、ADMM
- 近邻点算法、投影算法

作业

1. 试述真正例率(TPR) 、假正例率(FPR)与查准率(P) 、查全率(R)之间的联系.

2. 试述min-max 规范化、z-score 规范化的优缺点.

3. 数据集包含1000 个样本，其中500 个正例、500 个反例，将其划分为包含70% 样本的训练集和30% 样本的测试集用于留出法评估，试估算共有多少种划分方式.

4. 本章 1.4 节在论述“没有免费的午餐”定理时, 默认使用了“分类错误率”作为性能度量来对分类器进行评估. 若换用其他性能度量 ℓ, 则式(1.1)将改为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\ell\left(h\left(\boldsymbol{x}\right),f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)\ ,$$

试证明“没有免费的午餐定理”仍成立.

感 谢 观 看

统 计 机 器 学 习

主讲人：彭振华

数学与计算机学院

2026年