

BigData_MachineLearning

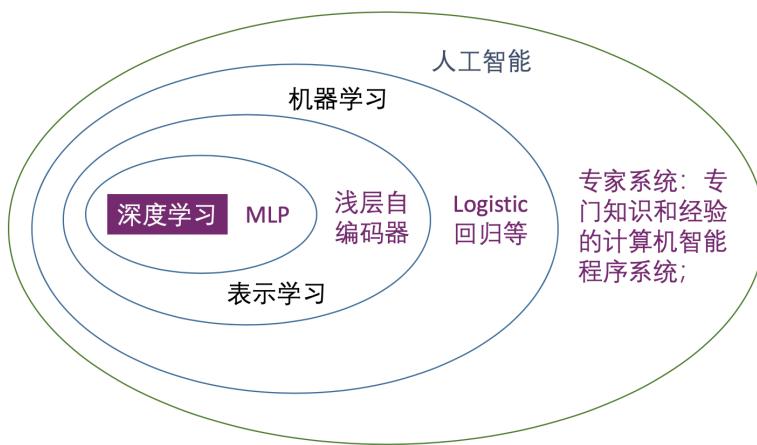
第一章 绪论

机器学习：

机器学习是近20多年兴起的一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。

人工智能/机器学习/深度学习

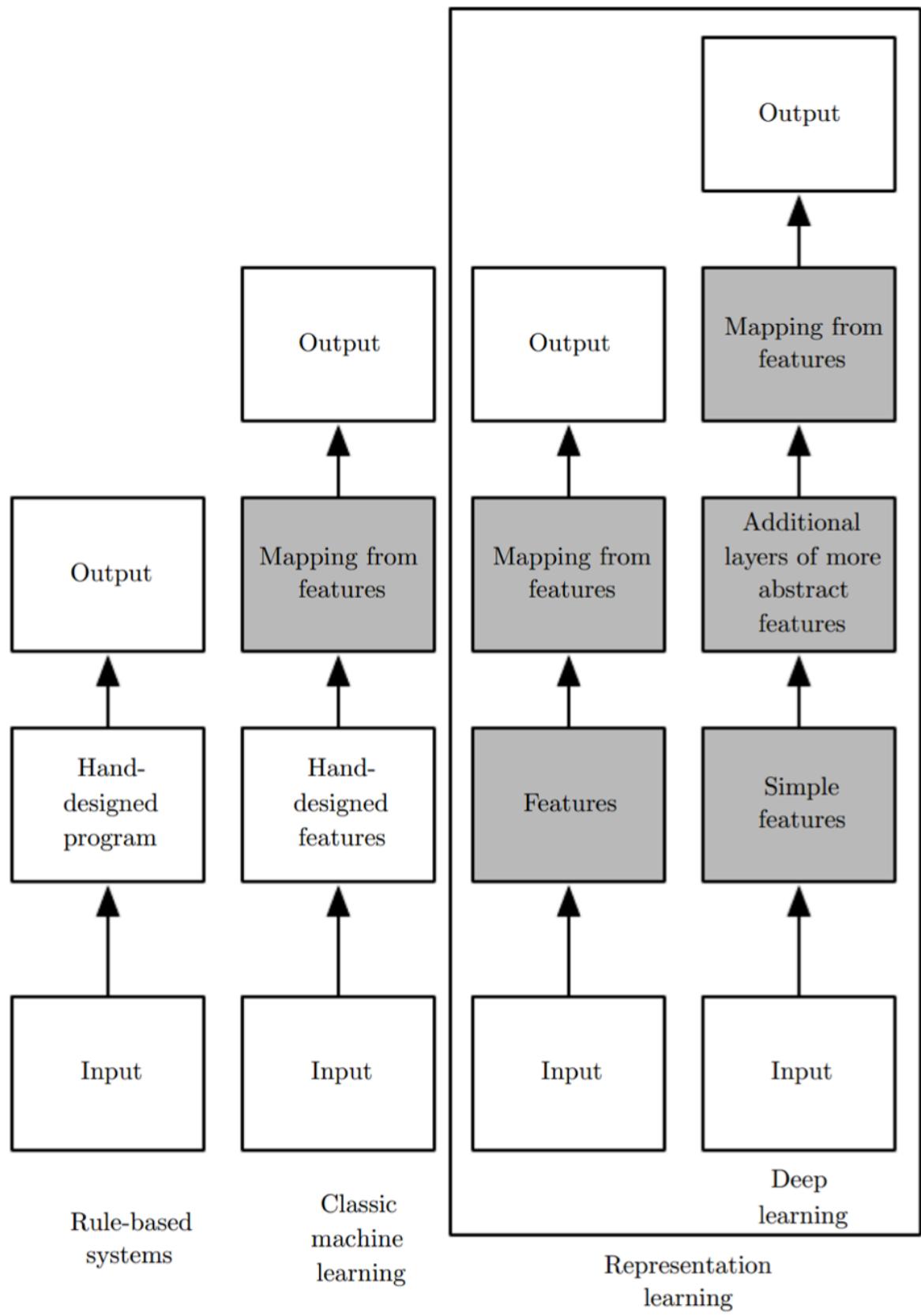
■ 人工智能/机器学习/深度学习



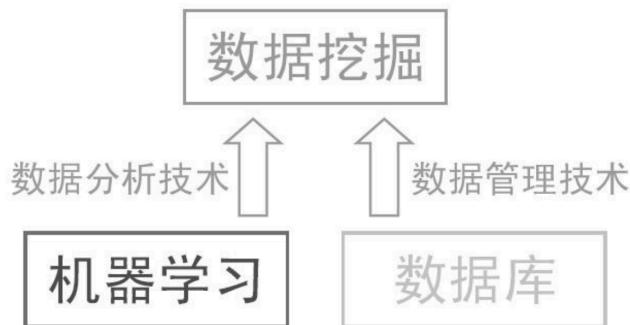
人工智能：是科学，为机器赋予视觉/听觉/触觉/推理等智能。

机器学习：人工智能的计算方法。

深度学习和人工智能其它方法



■ 机器学习和数据挖掘



计算机视觉是机器学习最重要的应用

机器学习和统计学习

- Simon Blomberg:
 - From R's fortunes package: To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'
- Andrew Gelman:
 - In that case, maybe we should get rid of checking of models and assumptions more often. Then maybe we'd be able to solve some of the problems that the machine learning people can solve but we can't

大数据机器学习的主要特征

- 与日俱增的数据量
- 实验数据量的增加
- 与日俱增的神经网络模型规模
- 与日俱增的精度、复杂度和对现实世界的冲击
- GPU (Graphic Processing Unit)
- TPU Tensor Processing Unit
- 深度学习框架
 - TensorFlow Pytorch Caffe CNTK Keras MXNet Theano Scikit-learning Spark MLlib

第二章 机器学习基本概念

基本术语

- Data set

- 形状=圆形 剥皮=难 味道=酸甜
- 形状=扁圆形 剥皮=易 味道=酸
- 形状=长圆形 剥皮=难 味道=甜
- Instance/sample
- Attribute value/feature
- Attribute/feature space
- Feature vector
- $D = x_1, x_2, \dots, x_m$ m个示例的数据集
- 是d维样本空间X的一个特征向量
- training/learning
- training data
- training sample
- Label ((形状=长圆形 剥皮=难 味道=甜), 橙子)
- example

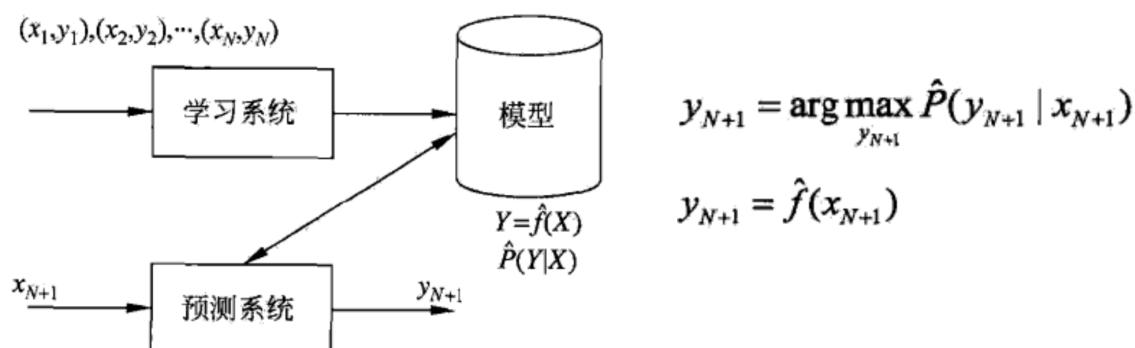
机器学习的任务

- Classification, discrete
- Regression, continuous
- Binary classification, 2-related
- Multi-class classification
- Clustering
- Multi-labeling annotation

监督学习

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

问题的形式化



假设空间 hypothesis space

- 学习过程: 搜索所有假设空间, 与训练集匹配
 - 形状=圆形 剥皮=难 味道=酸甜 橙
 - 形状=扁圆形 剥皮=易 味道=酸 橘
 - 形状=长圆形 剥皮=难 味道=甜 橙
- 假设形状, 剥皮, 味道 分别有3, 2, 3 种可能取值, 加上取任意值*和空集, 假设空间规模 $4 \times 3 \times 4 + 1 = 49$
- Version space: 与训练集一致的假设集合
 - 形状=剥皮=难 味道= 橙
 - 形状=扁圆形 剥皮=易 味道= * 橘

学习三要素, 方法=模型+策略+算法

模型

- 当假设空间F为决策函数的集合: $F = \{f|Y = f(x)\}$
- F实质为参数向量决定的函数族: $F = \{f|Y = f_\theta(x), \theta \in R^n\}$
- 当假设空间F为条件概率的集合: $F = \{P|P(X|Y)\}$
- F实质是参数向量决定的条件概率分布族: $F = \{P|P_\theta(Y|X), \theta \in R^n\}$

策略

损失函数和风险函数

- 0-1 loss function, $L(Y, f(x)) = \begin{cases} 1, & Y \neq f(x) \\ 0, & Y = f(x) \end{cases}$
- Quadratic loss function, $L(Y, f(X)) = (Y - f(X))^2$
- Absolute loss function, $L(Y, f(X)) = |Y - f(X)|$
- Logarithmic loss function/loglikelihood loss function, $L(Y, P(Y|X)) = -\log P(Y|X)$

损失函数的期望, 风险函数risk function, 期望损失expected loss

- $R_{exp}(f) = E_p[L(Y, f(X))] = \int_{x \times y} L(y, f(x))P(x, y)dxdy$

经验风险empirical risk, 经验损失empirical loss

- $T = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$
- $R_{emp}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$

因为风险函数很难求, 一般使得经验风险最小化与结构风险最小化

- 经验风险最小化模型, $\min_{f \in F} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$
- 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over-fitting"
- 为防止过拟合提出的策略, 结构风险最小化 structure risk minimization, 等价于正则化(regularization), 加入正则化项regularizer, 或罚项 penalty term
 - $R_{emp}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$

方法

求最优模型就是求解最优化问题:

- $\min_{f \in F} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$
- 难点
 - 全剧最优
 - 高校

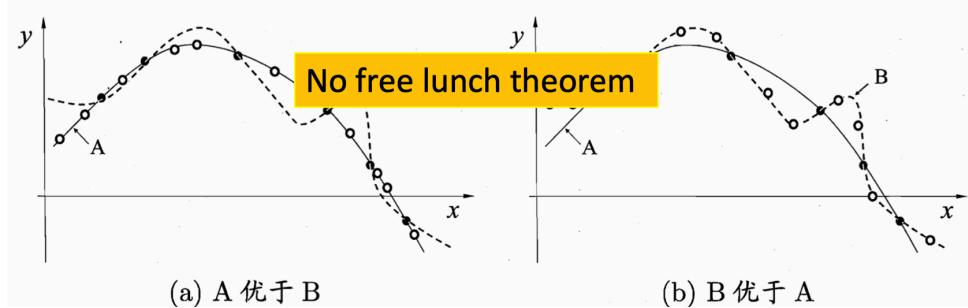
奥卡姆剃刀原理 Occam's razor

“如无必要，勿增实体”

- 疑问一：哪个更简单？

- 形状=* 剥皮=难 味道=* 橙
- 形状=长圆形 剥皮=* 味道=* 橙

- 疑问二：



No free lunch theorem

- 二分类问题：

总误差竟然与学习算法无关

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

- N F L 定理前提条件:
 - 所有“问题”出现的机会相同，或所有问题同等重要
 - 假设真实函数 f 的均匀分布。

- 形状= * 剥皮=难 味道=* 橙
- 形状=长圆形 剥皮=* 味道=* 橙

- N F L 寓意：脱离具体问题，空谈“什么方法好”毫无意义。

训练误差和测试误差

训练误差, 训练数据集的平均损失: $R_{emp}(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

测试误差, 测试训练集的平均损失: $e_{test} = \frac{1}{N} \sum_{i=1}^N L(y_i, f(\hat{x}_i))$

损失函数是0-1损失时: $e_{test} = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i \neq f(\hat{x}_i))$

测试数据集的准确率: $r_{test} = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i = f(\hat{x}_i))$

$$e_{test} + r_{test} = 1$$

过拟合

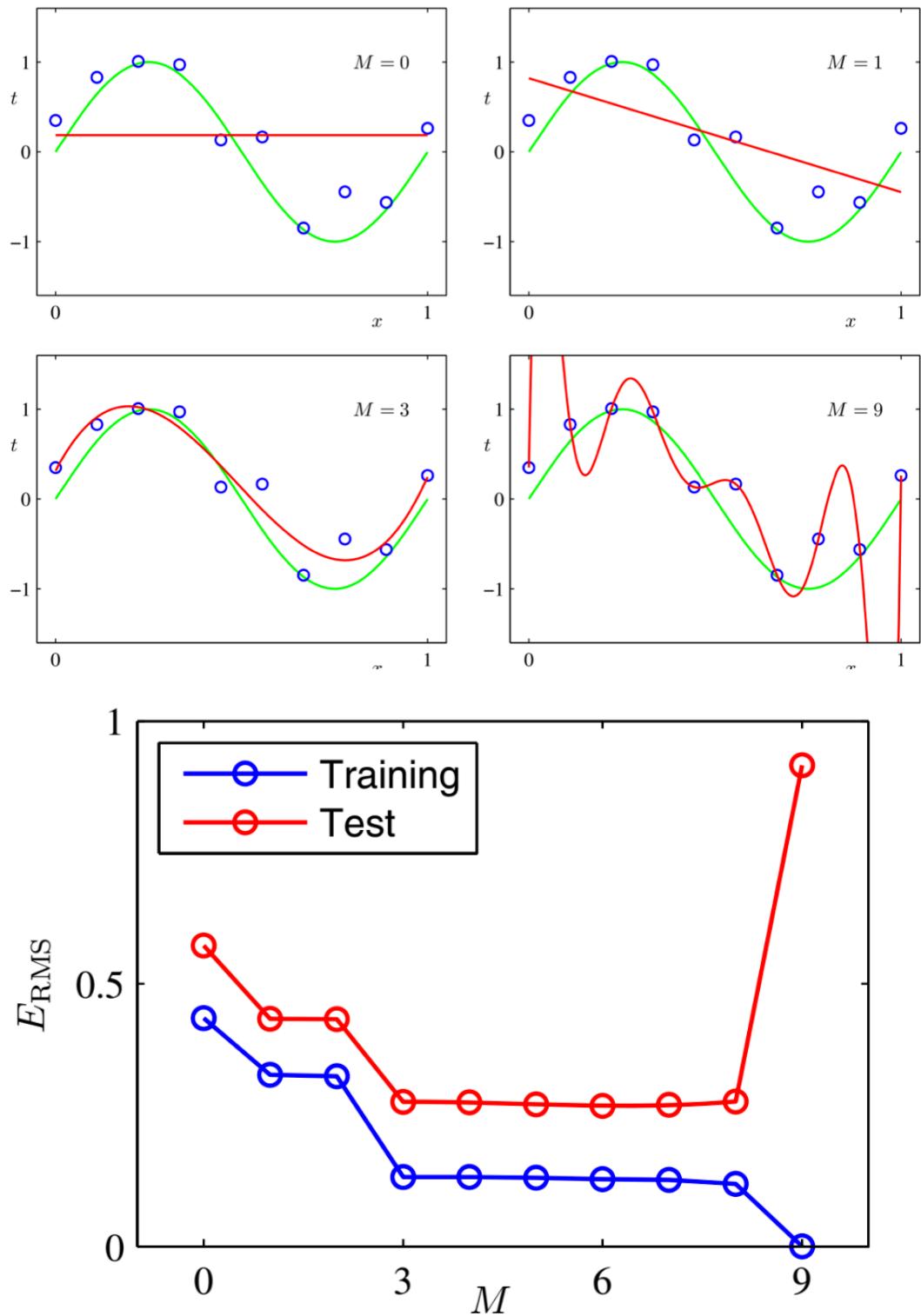
- 过拟合与模型选择-多项式曲线拟合的例子
- 假设给定训练数据集
- 假设给定训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$

$$f_M(x, w) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

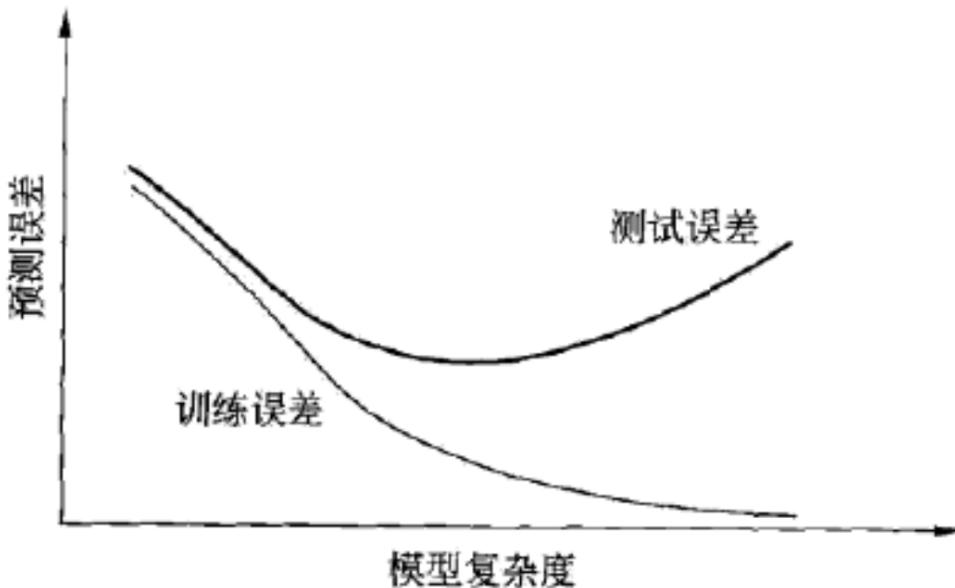
- 经验风险最小:

$$L(w) = \frac{1}{2} \sum_{i=1}^N (f(x_i, w) - y_i)^2 \quad L(w) = \frac{1}{2} \sum_{i=1}^N \left(\sum_{j=0}^M w_j x_i^j - y_i \right)^2$$

$$w_j = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^{j+1}}, \quad j = 0, 1, 2, \dots, M$$

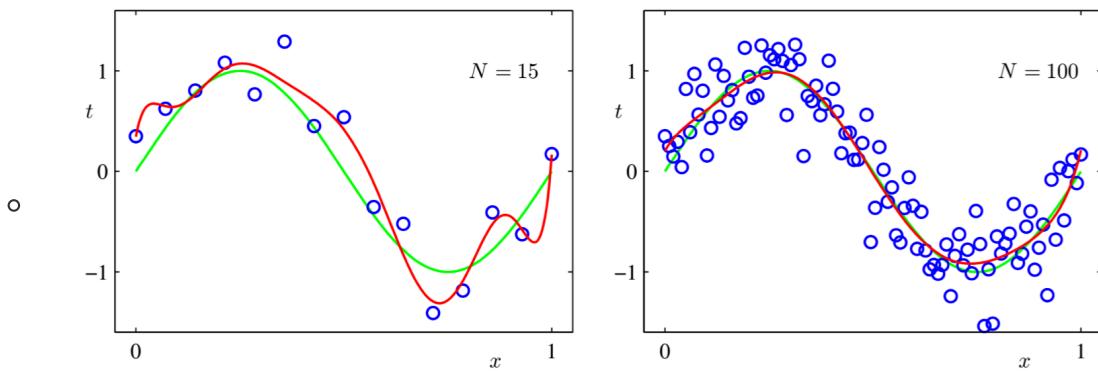


$M = 9$ 为过拟合



解决方法：

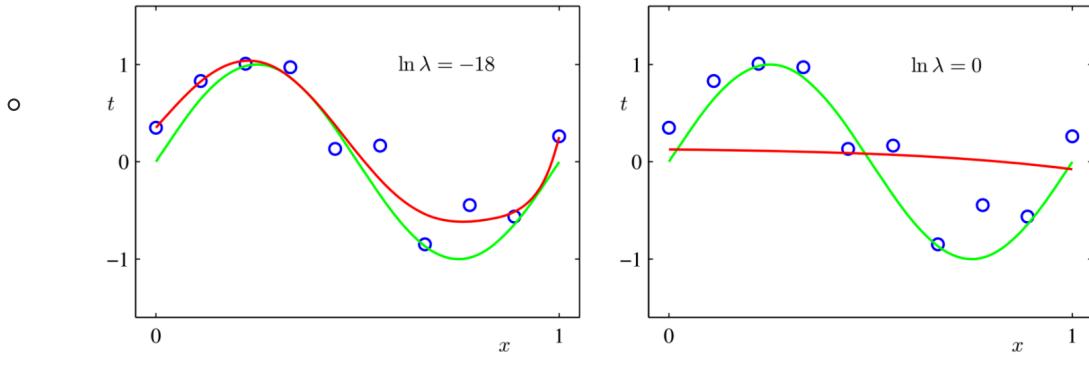
- 增大训练样本集



- 正则化

- 正则化一般形式： $\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$
- 回归问题中： $L(w) = \frac{1}{N} \sum_{i=1}^N (f(x_i; w) - y_i)^2 + \frac{\lambda}{2} \|w\|^2$
 $L(w) = \frac{1}{N} \sum_{i=1}^N (f(x_i; w) - y_i)^2 + \lambda \|w\|_1$

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$



- λ 抑制模型复杂化

泛化能力 generalization ability

- 泛化误差 generalization error

$$R_{\text{exp}}(\hat{f}) = E_p[L(Y, \hat{f}(X))] = \int_{x,y} L(y, \hat{f}(x)) P(x, y) dx dy$$

- 泛化误差上界

- 比较学习方法的泛化能力-----比较泛化误差上界
- 性质：样本容量增加，泛化误差趋于0
- 假设空间容量越大，泛化误差越大

- 二分类问题

$$X \in \mathbf{R}^n, Y \in \{-1, +1\}$$

- 期望风险和经验风险

$$R(f) = E[L(Y, f(X))]$$

- 假设空间F为有限集合

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$$

- 经验风险最小化函数：

$$f_N = \arg \min_{f \in \mathcal{F}} \hat{R}(f)$$

- 泛化能力：

$$R(f_N) = E[L(Y, f_N(X))]$$

- 定理：泛化误差上界，二分类问题，当假设空间是有限个函数的结合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 对任意一个函数f，至少以概率 $1-\delta$ ，以下不等式成立：

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N} \left(\log d + \log \frac{1}{\delta} \right)}$$

d 为假设空间

生成模型与判别模型

- 监督学习的目的就是学习一个模型:
- 决策函数: $Y = f(X)$
- 条件概率分布: $P(Y|X)$
 - 生成方法 Generative approach 对应生成模型: generative model,
 - 朴素贝叶斯法和隐马尔科夫模型
 - 判别方法 discriminative approach 对应判别模型: discriminative model
 - K近邻, 感知机, 决策树, logistic 回归等

• 二者各有优缺点

• 生成模型:

- 还原联合概率, 而判别模型不能;
- 学习收敛速度快, 当样本容量增加时, 学到的模型可以更快收敛;
- 当存在隐变量时, 可以使用生成模型, 而判别模型不行。

• 判别模型:

- 直接学习决策函数或条件概率, 学习的准确率更高;
- 可以对数据进行抽象, 定义特征和使用特征, 可以简化学习问题。

第三章 模型评估方法

模型评估方法

- 泛化误差评估:
 - 训练集 training set: 用于训练模型
 - 验证集 validation set: 用于模型选择
 - 测试集 test set: 用于模型泛化误差的近似
- 训练集和测试集的产生
 - 留出法
 - 交叉验证法
 - 自助法

留出法 Hold-out

训练集S, 测试集T, D为数据集

$$D = S \cup T$$

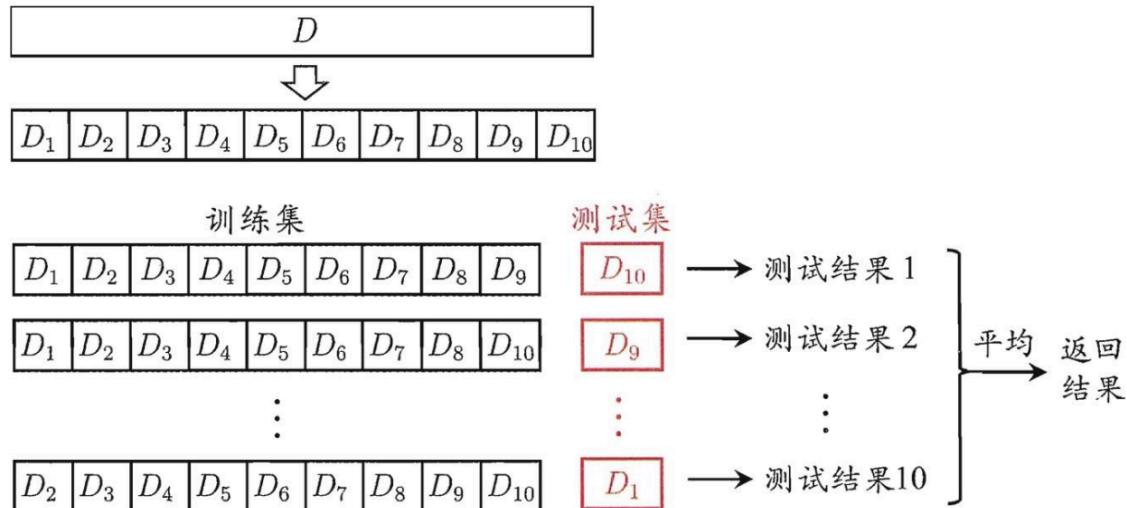
$$S \cap T = \emptyset$$

- 注意点:

- 训练/测试集的划分尽可能保持数据分布的一致性，避免引入额外偏差
- 存在多种划分方式对初始数据集进行分割，采用若干次随机划分，重复实验
- 存在问题：
 - S 大, T 小; S 小, T 大，都会带来负面影响

交叉验证法 cross validation

- $D \rightarrow k$ 个大小相等的互斥子集
- $D = D_1 \cup D_2 \cup \dots \cup D_k, D_i \cap D_j = \emptyset (i \neq j)$
- $K - 1$ 个子集并集为训练集，1个测试集



自助法 bootstrapping

- 自助采样法:
 - $\lim_{m \rightarrow \infty} (1 - \frac{1}{m})^m \rightarrow \frac{1}{e} \approx 0.368$
- 测试集: $D \setminus D'$, \ 为集合减法
- 优点
 - 适用于数据集较小，难以划分；
 - 从数据集产生不同的训练集，适用于集成学习方法；
- 缺点
 - 产生的训练集改变了初始数据集的分布，会引入估计偏差。

性能度量

- 不同任务，性能度量不同
 - 回归任务 - 均方误差:
 - $E(f; D) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$
 - 更一般:
 - $E(f; D) = \int_{x \sim D} (f(x) - y)^2 p(x) dx$
- 错误率和精度 - 分类任务

- 错误率
 - $E(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) \neq y_i)$, \mathbb{I} is the indicator function
- 精度
 - $acc(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) = y_i) = 1 - E(f; D)$
- 更一般:
 - $E(f; D) = \int_{x \sim D} \mathbb{I}(f(x) \neq y) p(x) dx$
 - $acc(f; D) = \int_{x \sim D} \mathbb{I}(f(x) = y) p(x) dx = 1 - E(f; D)$
- 查准率precision、查全率recall与F1

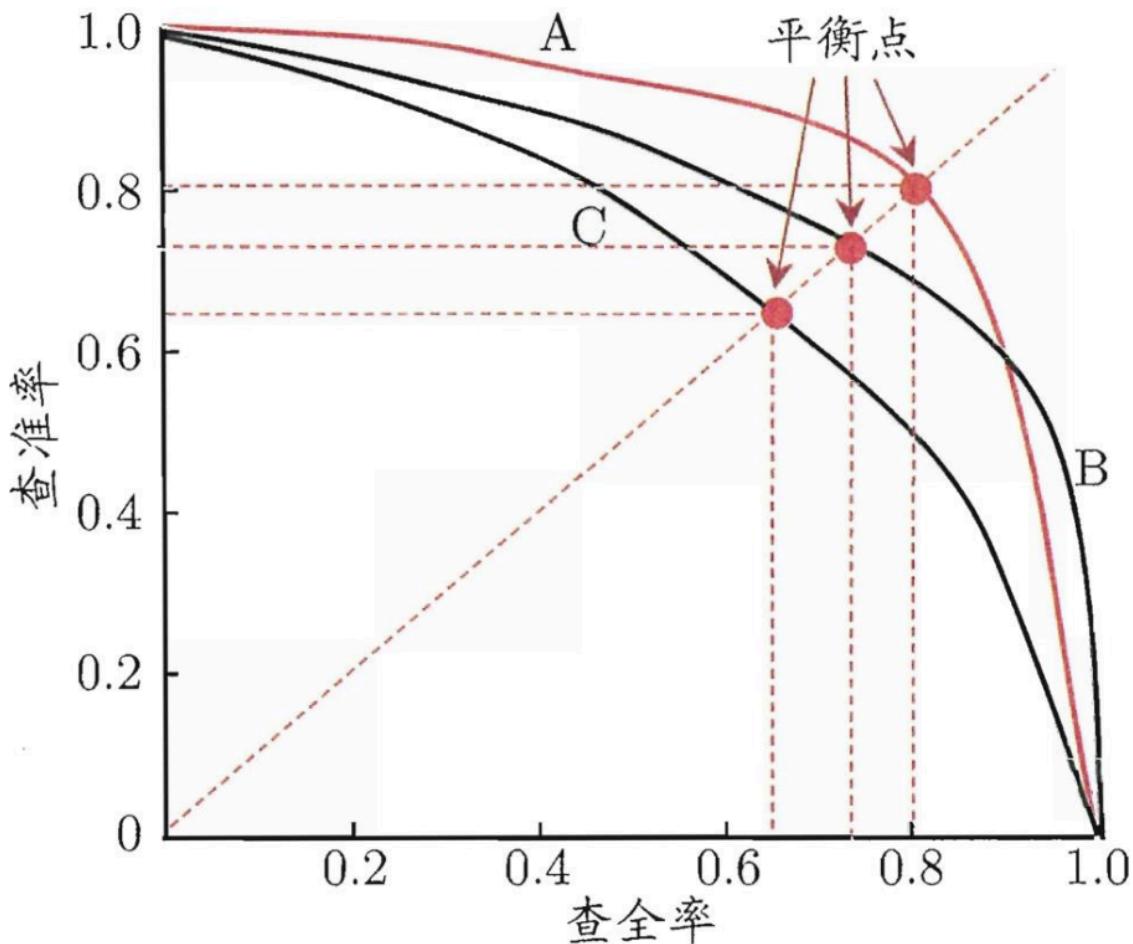
- 二分类-混淆矩阵:

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: $R = \frac{TP}{TP + FN}$

- P-R曲线

-



- 平衡点BEP
 - 查准率=查全率

- $F1$ 度量

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

- F_β 度量

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

-

- 多个二分类混淆矩阵:

- 多次训练/测试
- 多个数据集上训练/测试
- 执行多分类任务

- 宏查准率(macro-P)/宏查全率(macro-R)/宏F1

$$\text{macro-}P = \frac{1}{n} \sum_{i=1}^n P_i \quad \text{macro-}R = \frac{1}{n} \sum_{i=1}^n R_i \quad \text{macro-}F1 = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R}$$

- 微查准率(micro-P)/微查全率"(micro-R)和“微F1

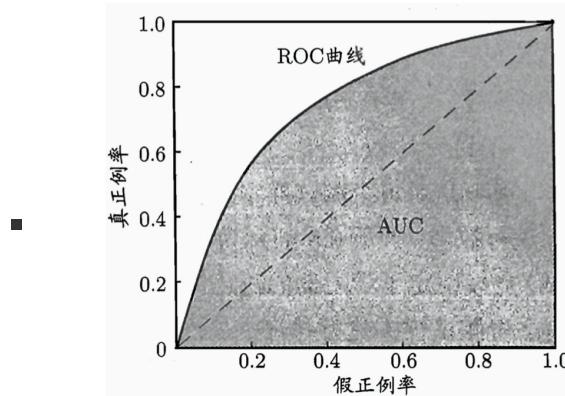
$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \quad \text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \quad \text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R}$$

- ROC (Receiver Operating Characteristic), AUC(Area Under ROC Curve)

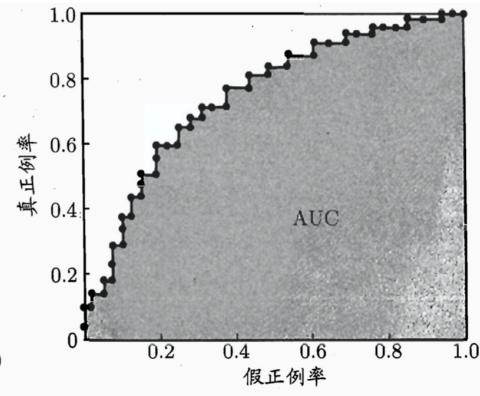
- 纵轴:“真正例率”(True Positive Rate, 简称 TPR)

- 横轴:“假正例率”(False Positive Rate, 简称 FPR)

- $$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{TN+FP}$$



(a) ROC 曲线与 AUC



(b) 基于有限样例绘制的 ROC 曲线与 AUC

- 代价敏感错误率与代价曲线

- 应用背景: 不同类型的错误所造成的后果不同

- 二分类任务: 代价矩阵(cost matrix)

-

真实类别	预测类别	
	第0类	第1类
第0类	0	$cost_{01}$
第1类	$cost_{10}$	0

- 对应代价敏感错误率

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

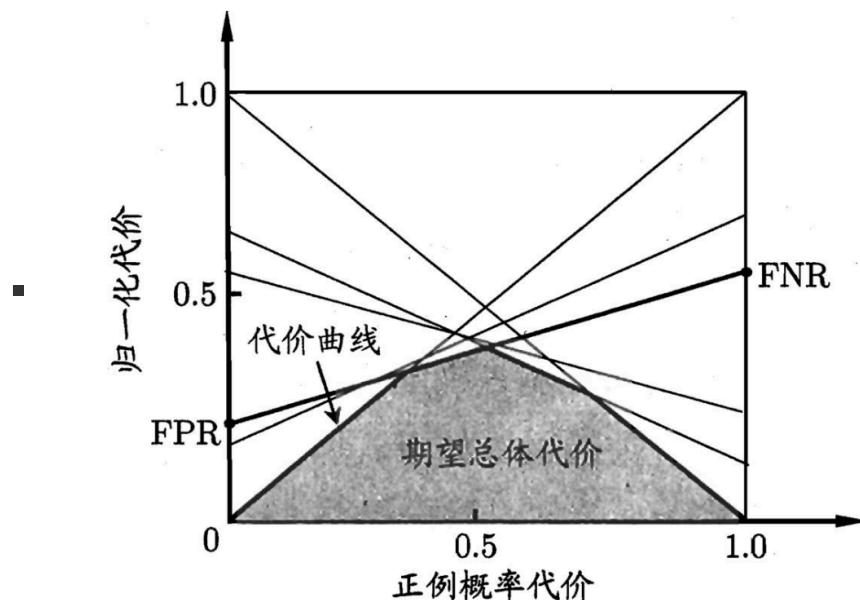
- 代价曲线cost curve: 非均等代价下ROC曲线不适用;

- 横轴: 正例概率代价: P为样例为正例的概率。

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

- 纵轴: 纵轴是取值为 [0,1] 的归一化代价

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



- 比较检验

- 问题提出: 能否直接用上述评估方法获得的性能度量"比大小"?

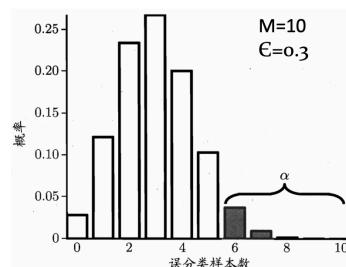
- 答案:不能

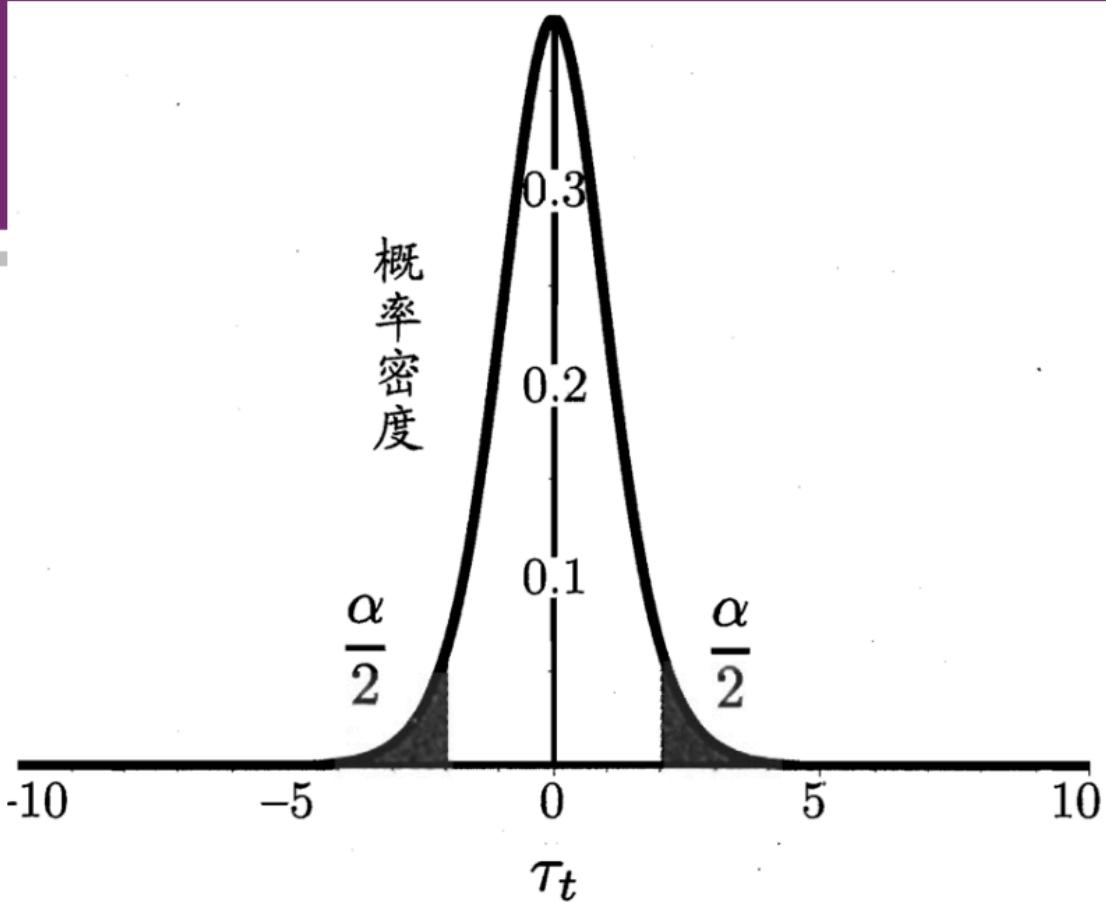
- 原因:

- 希望比较泛化性能, 实验评估的是测试集性能;

- 测试集性能和测试集的选择有关, 测试样例不同, 结果不同;

- 机器学习算法本身有一定的随机性，相同的参数，相同的数据集，结果也会不同。
- 方案：统计假设检验(hypothesis test)
 - 在测试集上观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握有多大
- 假设检验
 - 对单个学习器泛化性能的假设进行检验
 - "二项检验" (binomial test)
 - t 检验 (t-test)
 - 对不同学习器的性能进行比较
 - "成对t 检验" (paired t-tests)
- 二项检验
 - 假设检验：“假设”是对学习器泛化错误率分布的某种判断或猜想，如 ϵ
 - 现实任务中我们只能获知测试错误率 $\hat{\epsilon}$
 - 那么：泛化错误率为 ϵ 的学习器将其中 m' 个样本误分类的概率：
 - $P(\hat{\epsilon}; \epsilon) = \binom{m}{\hat{\epsilon} \times m} \epsilon^{\hat{\epsilon} \times m} (1 - \epsilon)^{m - \hat{\epsilon} \times m}$
 - 使用二项检验对泛化误差 $\epsilon \leq 0.3$ 的假设进行检验
 - $1 - \alpha$ 的概率内所能观测到的最大错误率：
 - $\bar{\epsilon} = \max \epsilon \text{ s.t. } \sum_{i=\epsilon_0 \times m+1}^m \binom{m}{i} \epsilon^i (1 - \epsilon)^{m-i} < \alpha$
 - 二项检验
 - 假设检验：“假设”是对学习器泛化错误率分布的某种判断或猜想，如 ϵ
 - 现实任务中我们只能获知测试错误率
 - 那么：泛化错误率为 ϵ 的学习器将其中 $\hat{\epsilon}_n$ 个样本误分类的概率：
 - $P(\hat{\epsilon}; \epsilon) = \binom{m}{\hat{\epsilon} \times m} \epsilon^{\hat{\epsilon} \times m} (1 - \epsilon)^{m - \hat{\epsilon} \times m}$
 - 使用二项检验对泛化误差 $\epsilon \leq 0.3$ 的假设进行检验；
 - $1 - \alpha$ 的概率内所能观测到的最大错误率：
- t检验
 - 多次重复训练/测试，得到多个测试错误率
 - K个测试错误率， $\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_k$
 - 均值 $\mu = \frac{1}{k} \sum_{i=1}^k \hat{\epsilon}_i$
 - 方差 $\sigma^2 = \frac{1}{k-1} \sum_{i=1}^k (\hat{\epsilon}_i - \mu)^2$
 - 考虑到K个测试错误率可看作泛化错误率 ϵ_0 的独立采样，
 - 则变量 $\tau_t = \frac{\sqrt{k}(\mu - \epsilon_0)}{\sigma}$ 服从自由度为K-1的t分布
 - 对假设 $\mu = \epsilon_0$ 和显著度 α ，可计算当测试错误率均值为 ϵ_0 ，在 $1 - \alpha$ 概率内能观测到的最大错误率，即临界值，如果 $|\mu - \epsilon_0|$ 位于临界值内，则假设成立





- 交叉验证t检验

- 学习器A,B, 得到:
 $\epsilon_1^A, \epsilon_2^A, \dots, \epsilon_k^A$ 和 $\epsilon_1^B, \epsilon_2^B, \dots, \epsilon_k^B$
- 成对t检验: 假设 $\epsilon_i^A = \epsilon_i^B$
- 计算: $\Delta_i = \epsilon_i^A - \epsilon_i^B$
- 计算均值和方差
- 在显著度 α 下, 若 $\tau_t = \left| \frac{\sqrt{k}\mu}{\sigma} \right|$ 小于临界值, 则假设不能被拒绝。

偏差与方差

- 偏差-方差分解
- 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
- $f(X; D)$ 为训练集 D 上学得模型 f 在 x 上的预测输出
- 回归方法的期望预测:
 - $\bar{f}(x) = E_D[f(x; D)]$
 - $var(x) = E_D[(f(x; D) - \bar{f}(x))^2]$

- 噪声为

- $\epsilon^2 = E_D[(y_D - y)^2]$
- 期望输出与真实标记的差别称为偏差
 - $bias^2(x) = (\bar{f}(x) - y)^2$

◦

- 假定噪声期望为0，对算法的期望泛化误差进行分解：

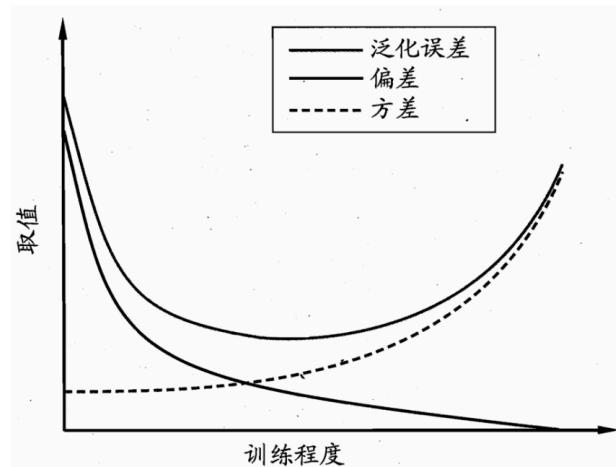
$$\begin{aligned} E(f; D) &= \mathbb{E}_D [(f(\mathbf{x}; D) - y_D)^2] \\ &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - y_D)^2] \\ &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2] + (\bar{f}(\mathbf{x}) - y)^2 + \mathbb{E}_D [(y_D - y)^2] \end{aligned}$$

$$E(f; D) = bias^2(\mathbf{x}) + var(\mathbf{x}) + \epsilon^2$$

- 泛化误差可分解为偏差、方差与噪声之和.

◦

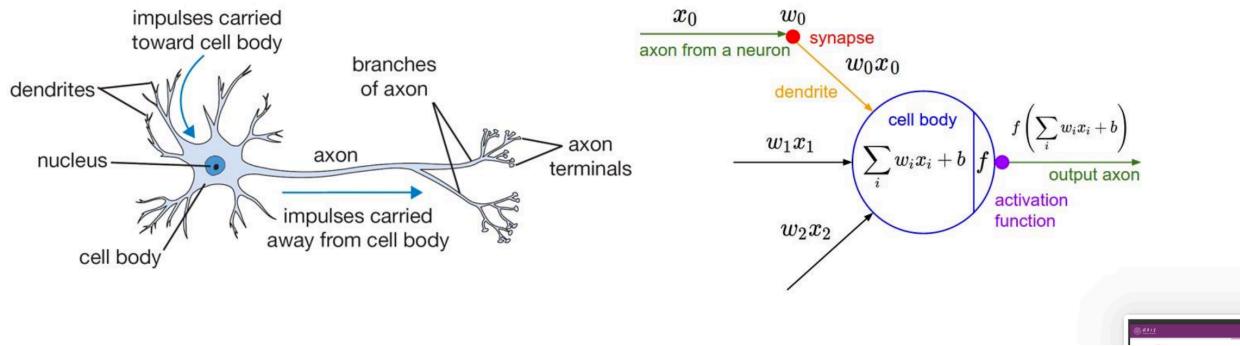
- **偏差-方差窘境** (bias-variance dilemma)



第四章 感知机模型

感知机模型

- 神经网络、支持向量机的基础 (线性可分性和对偶性)



- 感知机 (Perceptron)

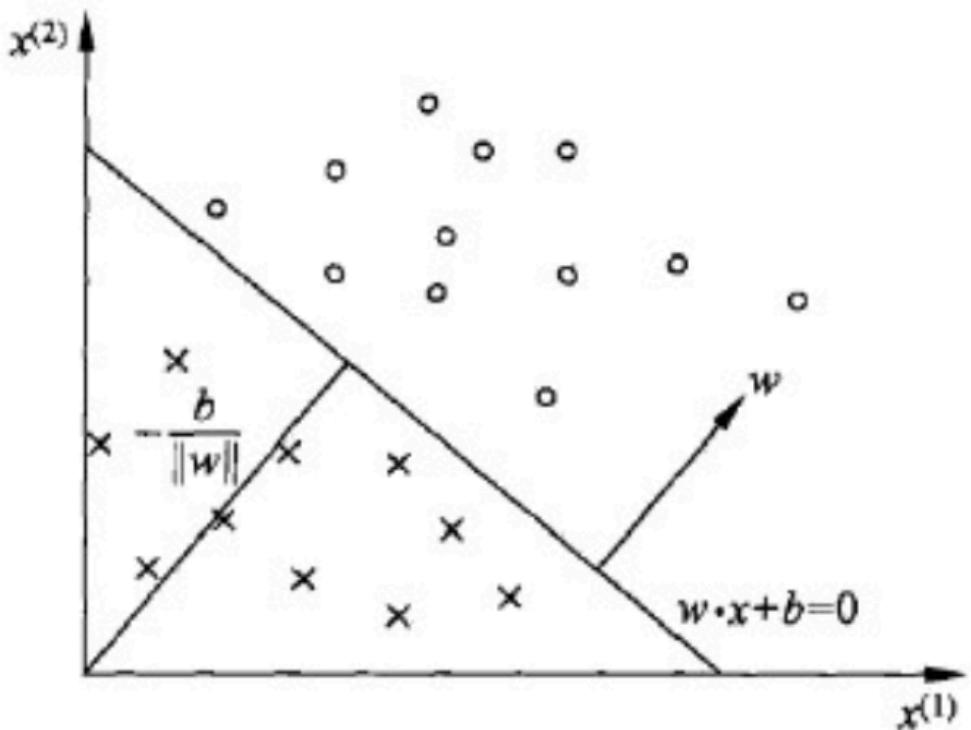
- 针对: 二分类问题
- 实质: 分离超平面, 判别模型
- 策略: 基于误分类的损失函数
- 方法: 利用梯度下降法对损失函数进行极小化
- 特点: 感知机学习算法具有简单而易于实现的优点
- 分类: 分为原始形式和对偶形式

- 定义

- 假设输入空间 (特征空间) 是 $X \subseteq \mathbb{R}^n$, 输出空间是 $Y = \{+1, -1\}$
- 输入 $x \in X$ 表示实例的特征向量, 对应于输入空间 (特征空间) 的点, 输出表示实例的类别, 由输入空间到输出空间的函数:
 - $f(x) = \text{sign}(\vec{w} \cdot \vec{x} + b)$ 称之为感知机
 - 模型参数: w , x , 内积, 权值向量, 偏置
 - 符号函数
 - $\text{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$

- 感知机的几何解释

- 线性方程: $\vec{w} \cdot \vec{x} + b = 0$
- 对应于超平面 S , \vec{w} 为法向量, b 截距, 分离正负类
- 分离超平面



- 证明 \vec{w} 是法向量
 - 超平面为 $\vec{w} \cdot \vec{x} + b = 0$, 取平面内任意两点 x_1, x_2 , 有

- $$\begin{cases} \vec{w} \cdot \vec{x}_1 + b = 0 & (1) \\ \vec{w} \cdot \vec{x}_2 + b = 0 & (2) \end{cases}$$
- $(1) - (2) = \vec{w} \cdot (\vec{x}_1 - \vec{x}_2) = 0$ 且 $\vec{x}_1 - \vec{x}_2 = \vec{x}_2 - \vec{x}_1$
- 因此 \vec{w} 垂直此平面

- 感知机是线性的, 不能处理异或分类问题

感知机学习策略

- 定义损失函数, 并将其极小化
- 点到直线的距离
 - $Ax + By + C = 0$
 - $d = \left| \frac{Ax_0 + By_0 + C}{\sqrt{A^2 + B^2}} \right|$
- 如何定义损失函数
 - 自然选择: 误分类点的数目, 但损失函数不是 w, b 连续可导, 不宜优化
 - 另一选择: 误分类点到超平面的总距离:
 - 距离: $\frac{1}{\|\vec{w}\|} |\vec{w} \cdot \vec{x}_0 + b|$
 - 误分类点:
 - $-y_i(\vec{w} \cdot \vec{x} + b) > 0$
 - 误分类点的距离: $-\frac{1}{\|\vec{w}\|} y_i |\vec{w} \cdot \vec{x}_i + b|$
 - 总距离: $-\frac{1}{\|\vec{w}\|} \sum_{x_i \in M} y_i |\vec{w} \cdot \vec{x}_i + b|$
 - 损失函数, 不考虑范数

- $L(\vec{w}, b) = -\sum_{x_i \in M} y_i (\vec{w} \cdot \vec{x}_i + b)$
- M 为误分类点的数目

感知机的学习算法

- 求解最优化问题:
 - $\min_{w,b} L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$
 - 随机梯度下降法
 - 首先任意选择一个超平面, w, b , 然后不断极小化目标函数, 损失函数 L 的梯度
 - 选取误分类点更新
 - $\nabla_w L(w, b) = -\sum_{x_i \in M} y_i x_i, w \leftarrow w + \eta y_i x_i$
 - $\nabla_b L(w, b) = -\sum_{x_i \in M} y_i, b \leftarrow b + \eta y_i$
 - η : 学习步长, 学习率
 -

感知机学习算法的原始形式:

输入: 训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$,
 其中 $x_i \in \mathcal{X} = \mathbf{R}^n$, $y_i \in \mathcal{Y} = \{-1, +1\}$, $i = 1, 2, \dots, N$
 学习率 η ($0 < \eta \leq 1$);
 输出: w, b ; 感知机模型 $f(x) = \text{sign}(w \cdot x + b)$
 (1) 选取初值 w_0, b_0
 (2) 在训练集中选取数据 (x_i, y_i)
 (3) 如果 $y_i(w \cdot x_i + b) \leq 0$
 $w \leftarrow w + \eta y_i x_i$
 $b \leftarrow b + \eta y_i$
 (4) 转至 (2), 直至训练集中没有误分类点

- 例子

清华大学
Tsinghua University

感知机学习算法

- 例: 正例: $x_1 = (3, 3)^T, x_2 = (4, 3)^T$
- 负例: $x_3 = (1, 1)^T$

$x^{(2)}$

$2x^{(1)}+x^{(2)}-5=0$

$x^{(1)}$

x_1 x_2

x_3

$x^{(1)}+x^{(2)}-5=0$

■ 感知机学习算法

- 解：构建优化问题： $\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i (w \cdot x + b)$
- 求解： $w, b, \eta = 1$
 - 取初值 $w_0 = 0, b_0 = 0$
 - 对 $x_1 = (3, 3)^T, y_1 (w_0 \cdot x_1 + b_0) = 0$, 未能被正确分类, 更新 w, b
 $w_1 = w_0 + y_1 x_1 = (3, 3)^T, b_1 = b_0 + y_1 = 1$
- 得线性模型： $w_1 \cdot x + b_1 = 3x^{(1)} + 3x^{(2)} + 1$
 - x_2 , 显然, $y_2 (w_1 \cdot x_2 + b_1) > 0$, 被正确分类,
 - 对 $x_3 = (1, 1)^T, y_3 (w_1 \cdot x_3 + b_1) < 0$, 被误分类,
 $w_2 = w_1 + y_3 x_3 = (2, 2)^T, b_2 = b_1 + y_3 = 0$

■ 感知机学习算法

- 得到线性模型： $w_2 \cdot x + b_2 = 2x^{(1)} + 2x^{(2)}$
- 如此继续下去： $w_7 = (1, 1)^T, b_7 = -3 \quad w_7 \cdot x + b_7 = x^{(1)} + x^{(2)} - 3$
- 分离超平面： $x^{(1)} + x^{(2)} - 3 = 0$
- 感知机模型： $f(x) = \text{sign}(x^{(1)} + x^{(2)} - 3)$

迭代次数	误分类点	w	b	w · x + b
0		0	0	0
1	x_1	$(3, 3)^T$	1	$3x^{(1)} + 3x^{(2)} + 1$
2	x_2	$(2, 2)^T$	0	$2x^{(1)} + 2x^{(2)}$
3	x_1	$(1, 1)^T$	-1	$x^{(1)} + x^{(2)} - 1$
4	x_1	$(0, 0)^T$	-2	-2
5	x_1	$(3, 3)^T$	-1	$3x^{(1)} + 3x^{(2)} - 1$
6	x_2	$(2, 2)^T$	-2	$2x^{(1)} + 2x^{(2)} - 2$
7	x_1	$(1, 1)^T$	-3	$x^{(1)} + x^{(2)} - 3$
8	0	$(1, 1)^T$	-3	$x^{(1)} + x^{(2)} - 3$

页码: 15/30

● 算法的收敛性

- 算法的收敛性: 证明经过有限次迭代可以得到一个将训练数据集完全正确划分的分离超平面及感知机模型。
- 将b并入权重向量w, 记作: $\hat{w} = (w^T, b)^T$
 $\hat{x} = (x^T, 1)^T$
 $\hat{x} \in \mathbf{R}^{n+1}, \hat{w} \in \mathbf{R}^{n+1}$
 $\hat{w} \cdot \hat{x} = w \cdot x + b$

○

- 定理：设训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 是线性可分的，其中 $x_i \in \mathcal{X} = \mathbf{R}^n$, $y_i \in \mathcal{Y} = \{-1, +1\}$, $i = 1, 2, \dots, N$,

- 则：(1) 存在满足条件 $\|\hat{w}_{opt}\| = 1$ 的超平面 $\hat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x + b_{opt} = 0$ ；且存在 $\gamma > 0$, 对所有 $i = 1, 2, \dots, N$

$$y_i(\hat{w}_{opt} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

◦

- 证明：(1)
- 由线性可分，存在超平面： $\hat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x + b_{opt} = 0$
- 使 $\|\hat{w}_{opt}\| = 1$ 有限的点，均有：

$$y_i(\hat{w}_{opt} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) > 0$$

- 存在

$$\gamma = \min_i \{y_i(w_{opt} \cdot x_i + b_{opt})\}$$

- 使：

$$y_i(\hat{w}_{opt} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

◦

- (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$, 感知机算法在训练集的误分类次数k满足不等式， $k \leq \left(\frac{R}{\gamma}\right)^2$
- 证明：令 \hat{w}_{k-1} 是第k个误分类实例之前的扩充权值向量，即： $\hat{w}_{k-1} = (w_{k-1}^T, b_{k-1})^T$
- 第k个误分类实例的条件是： $y_i(\hat{w}_{k-1} \cdot \hat{x}_i) = y_i(w_{k-1} \cdot x_i + b_{k-1}) \leq 0$
- 则w和b的更新： $w_k \leftarrow w_{k-1} + \eta y_i x_i$ 即： $\hat{w}_k = \hat{w}_{k-1} + \eta y_i \hat{x}_i$
 $b_k \leftarrow b_{k-1} + \eta y_i$

◦

- (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$, 感知机算法在训练集的误分类次数k满足不等式, $k \leq \left(\frac{R}{\gamma}\right)^2$

• 推导两个不等式:

$$\text{(1)} \quad \hat{w}_k \cdot \hat{w}_{\text{opt}} \geq k\eta\gamma$$

• 由:

$$\begin{aligned} \hat{w}_k \cdot \hat{w}_{\text{opt}} &= \hat{w}_{k-1} \cdot \hat{w}_{\text{opt}} + \eta y_i \hat{w}_{\text{opt}} \cdot \hat{x}_i \\ &\geq \hat{w}_{k-1} \cdot \hat{w}_{\text{opt}} + \eta\gamma \end{aligned}$$

$$\text{• 得: } \hat{w}_k \cdot \hat{w}_{\text{opt}} \geq \hat{w}_{k-1} \cdot \hat{w}_{\text{opt}} + \eta\gamma \geq \hat{w}_{k-2} \cdot \hat{w}_{\text{opt}} + 2\eta\gamma \geq \dots \geq k\eta\gamma$$

◦

- (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$, 感知机算法在训练集的误分类次数k满足不等式, $k \leq \left(\frac{R}{\gamma}\right)^2$

$$(2) \quad \|\hat{w}_k\|^2 \leq k\eta^2 R^2$$

• 则:

$$\begin{aligned} \|\hat{w}_k\|^2 &= \|\hat{w}_{k-1}\|^2 + 2\eta y_i \hat{w}_{k-1} \cdot \hat{x}_i + \eta^2 \|\hat{x}_i\|^2 \\ &\leq \|\hat{w}_{k-1}\|^2 + \eta^2 \|\hat{x}_i\|^2 \\ &\leq \|\hat{w}_{k-1}\|^2 + \eta^2 R^2 \\ &\leq \|\hat{w}_{k-2}\|^2 + 2\eta^2 R^2 \leq \dots \\ &\leq k\eta^2 R^2 \end{aligned}$$

◦

- (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$, 感知机算法在训练集的误分类次数k满足不等式, $k \leq \left(\frac{R}{\gamma}\right)^2$

$$\begin{aligned} \text{结合两个不等式: } k\eta\gamma &\leq \hat{w}_k \cdot \hat{w}_{\text{opt}} \leq \|\hat{w}_k\| \|\hat{w}_{\text{opt}}\| \leq \sqrt{k}\eta R \\ k^2\gamma^2 &\leq kR^2 \end{aligned}$$

$$\text{得: } k \leq \left(\frac{R}{\gamma}\right)^2$$

◦ 定理表明:

- 误分类的次数k是有上界的, 当训练数据集线性可分时, 感知机学习算法原始形式迭代是收敛的。
- 感知机算法存在许多解, 既依赖于初值, 也依赖迭代过程中误分类点的选择顺序。
- 为得到唯一分离超平面, 需要增加约束, 如SVM。
- 线性不可分数据集, 迭代震荡。

感知机算法的对偶形式, 类似SVM的对偶形式

- 基本想法：
- 将w和b表示为实例xi和标记yi的线性组合的形式，通过求解其系数而求得w和b，对误分类点：

$$\begin{array}{l} w \leftarrow w + \eta y_i x_i \\ b \leftarrow b + \eta y_i \end{array} \quad \xrightarrow{\alpha_i = \eta_i \eta} \quad \begin{array}{l} w = \sum_{i=1}^N \alpha_i y_i x_i \\ b = \sum_{i=1}^N \alpha_i y_i \end{array}$$

最后学习到的 w, b
 $\alpha_i \geq 0, i = 1, 2, \dots, N$

- 实例点更新次数越多，意味着该点离分离超平面？

- 不是

- 感知机学习算法的对偶形式：

输入：训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，

其中 $x_i \in \mathcal{X} = \mathbf{R}^n$, $y_i \in \mathcal{Y} = \{-1, +1\}$, $i = 1, 2, \dots, N$

学习率 η ($0 < \eta \leq 1$)；

输出： α, b ：感知机模型 $f(x) = \text{sign}\left(\sum_{j=1}^N \alpha_j y_j x_j \cdot x + b\right)$.

其中 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$

(1) $\alpha \leftarrow 0, b \leftarrow 0$

(2) 在训练集中选取数据 (x_i, y_i)

(3) 如果 $y_i \left(\sum_{j=1}^N \alpha_j y_j x_j \cdot x_i + b \right) \leq 0$

$$\alpha_i \leftarrow \alpha_i + \eta$$

$$b \leftarrow b + \eta y_i$$

(4) 转至 (2) 直到没有误分类数据。

Gram 矩阵 $G = [x_i \cdot x_j]_{N \times N}$

- 例：正样本点是 $x_1 = (3, 3)^T$, $x_2 = (4, 3)^T$, 负样本点是 $x_3 = (1, 1)^T$

解 按照算法 2.2,

(1) 取 $\alpha_i = 0, i = 1, 2, 3, b = 0, \eta = 1$

(3) 误分条件

(2) 计算 Gram 矩阵

$$y_i \left(\sum_{j=1}^N \alpha_j y_j x_j \cdot x_i + b \right) \leq 0$$

$$G = \begin{bmatrix} 18 & 21 & 6 \\ 21 & 25 & 7 \\ 6 & 7 & 2 \end{bmatrix}$$

参数更新

$$\alpha_i \leftarrow \alpha_i + 1, b \leftarrow b + y_i$$

- 例：正样本点是 $x_1 = (3, 3)^T$, $x_2 = (4, 3)^T$, 负样本点是 $x_3 = (1, 1)^T$

(4) 迭代. 过程从略, 结果列于表 2.2.

k	0	1	2	3	4	5	6	7
		x_1	x_3	x_3	x_3	x_1	x_3	x_3
α_1	0	1	1	1	2	2	2	2
α_2	0	0	0	0	0	0	0	0
α_3	0	0	1	2	2	3	4	5
b	0	1	0	-1	0	-1	-2	-3

(5) $w = 2x_1 + 0x_2 - 5x_3 = (1, 1)^T$ 分离超平面 $x^{(1)} + x^{(2)} - 3 = 0$
 $b = -3$

感知机模型 $f(x) = \text{sign}(x^{(1)} + x^{(2)} - 3)$