中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

统计机器学习

张文生

中国科学院自动化研究所

中科院大学人工智能学院

2019-12-12

内容简介

一、统计机器学习

二、统计学习建立

三、统计三大学派

四、统计学习方法

统计机器学习

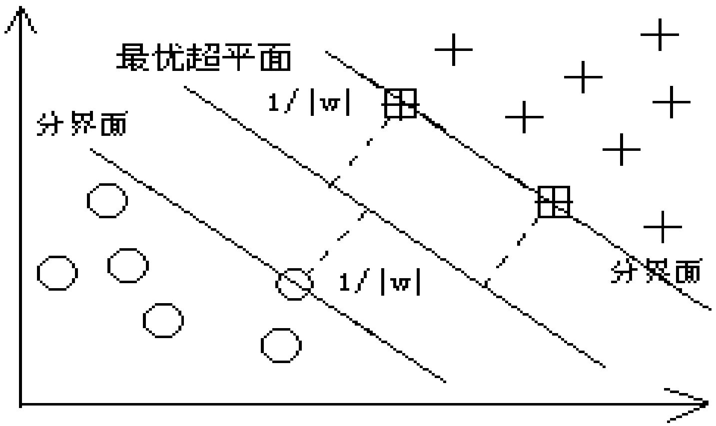
- 统计机器学习（Statistical Machine Learning, SML）

  利用有限数量的观测来寻找待求的依赖关系(Vapnik 1995)

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

- 理论风险<=经验风险+结构风险

- 最优分类超平面选取技术



支持向量机的几何解释

$$R(w) \le R_{emp}(w) + \Phi(n/h)$$

$$R(\omega) \le R_{emp}(\omega) + \tfrac{\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\omega)}{\varepsilon}}\right) \qquad \forall \omega$$

$$\min_{w} \eta(w) = \tfrac{1}{2} \| w \|^2$$

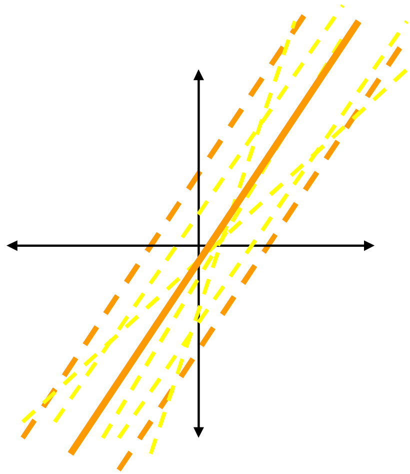
$$s.t. \quad y_i[(w \cdot x_i) + w_0] \ge 1, \quad i = 1,2,\dots,n$$

$$\max_{\alpha} \; Q(\alpha) = -\tfrac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n} \alpha_i$$

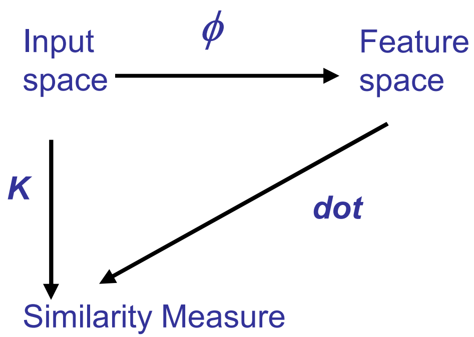
$$s.t. \quad \sum_{i=1}^{n} y_i \alpha_i = 0, \qquad \alpha_i \ge 0, \quad i = 1,2,\dots, n$$

2-class Support Vector Machines

1. Linear separation via the maximal-margin hyperplane (an principle)
2. Maximality via quadratic programming (QP)
3. Linear non-separability via penalised slacks, Wolfe dual QP
4. Non-linear separation via feature mapping: separate points in higher dimensional feature space rather than input space
5. Implicit feature mapping – the “kernel trick”



Maximal-margin hyperplanes



The kernel trick

内容简介

一、统计机器学习

二、统计学习建立

三、统计三大学派

四、统计学习方法

SML理论的建立过程

- **1971年[Vapnik与Chervonenkis]提出VC维**
- **1989年[Blumer]证明了VC维与Valiant的“可学习理论”(PAC)有密切联系**
- **1995年[V.N. Vapnik]出版了**统计机器学习**的本质一书, 这标志着**统计机器学习理论已经建立
- **1996年以来[Vapnik,ect]将感知机这类研究包括在SML中，他将这种模型称为支持向量机(Support Vector Machine，简称SVM)，这意味着，这个理论将走向应用**

Vapnik将SML的发展分为四个阶段

(1) Rosenblatt的感知机(60年代)

(2) 建立学习理论基础 (60-70年代)

(3) 神经网络(80年代)

(4) 返回到感知机年代(90年代)

Duda & Hart

1973年，他们出版了至今有重要影响“Pattern classification and scene analysis”，2001年，在此基础上，删除了情境分析的内容，大量增加了统计建模的内容。

尽管2001年版的内容大大丰富了，无论在理论研究结果，方法的罗列，还是参考文献的收集，都可以称为一本研究者必备的手册，但是，其理论框架的识别也比1973版困难。

统计机器学习的统计框架

Duda & Hart的模式分类理论框架=统计机器学习理论框架

Bayes理论论
后验概率：$P(\omega_j|x)=P(\omega_i)p(x_j|\omega_i)$。样本数趋于无穷大。
判决规则：对所有ω_j，最大$P(\omega_j|x)$就是x的类别。

目标：风险$R(\alpha_i|x)=\Sigma\lambda(\alpha_i|\omega_j)P(\omega_i|x)$最小，$\lambda$是损失函数。

令$g_j(x)=P(\omega_j|x)$, $g(x)=g_j(x)-g_i(x)$。判别为计算g(x)的参数。

函数$g(x)=w_0+\Sigma w^t x$，如果$\Sigma w^t x>-w_0$，x属于ω_1。

问题变为在确定的损失函数(准则函数或目标函数)意义的优化问题。
线性感知机就是如此。损失函数是平方损失。

有限样本理论

Vapnik有限样本理论：考虑两个因素，其一，有限样本，其二，算法的计算复杂性是多项式。由此，接受PAC并推出泛化界。结构风险等。

线性算法

BP算法：非线性形式$y=f_1(\theta_1 f_2(\theta_2 x))$，算法漂亮，科学上：孤立事件。

Vapnik提出核映射，将样本集合映射到线性内积的Hilbert空间，样本集合成为线性可分，直接使用感知机。

“对某个问题已经认识，是找到一个空间，这个问题可以在这个空间上线性表述”，这个在二十世纪三十年代Von Neurmman在研究量子力学数学基础时暗示的思想，其数学方法，就是Hilbert空间。

n-XOR问题：将问题映射到多项式基张成的空间，并定义空间各维度在{0, 1}上，可以证明，n-XOR线性可分维数是2^n。维数灾难！

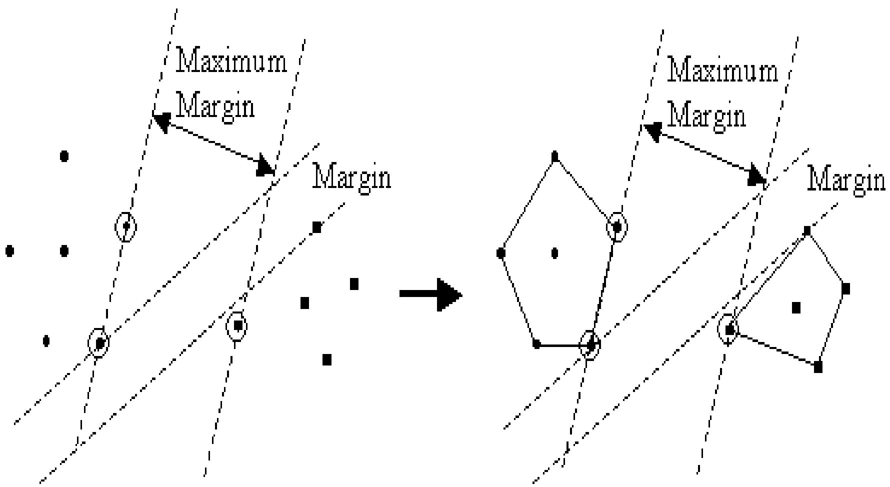
如果将空间的各维度定义在实数域上，可以线性划分这个问题的维数减低，最小的维数是什么？如果事先确定维数，代价可能就是精度。

泛化误差界

Vapnik首先推出了PAC泛化误差不等式，但是，这个研究对算法设计没有本质的指导意义

1998年，Shawe-Taylor等推出的基于边缘的泛化不等式

M是不同类别数据分界的边缘。



问题变为设计使得两个闭凸集边缘最大的算法。由于直观的几何描述受到理论和应用研究者的偏爱。Vapnik称这个时期为Margin时期。

内容简介

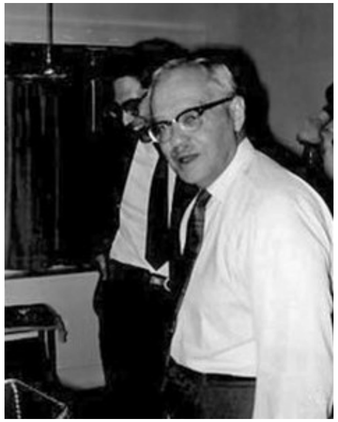
一、统计机器学习

二、统计学习建立

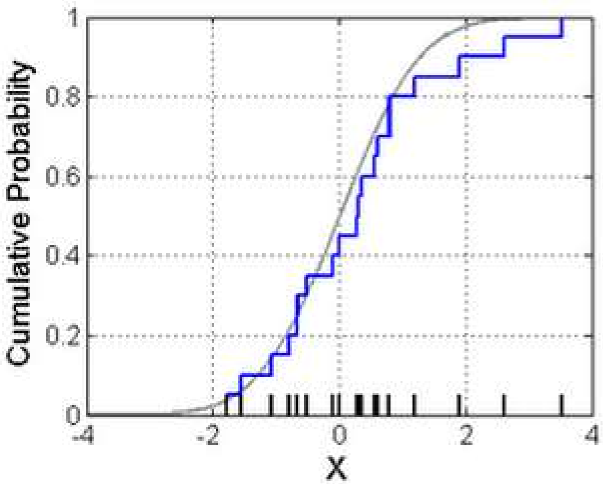
三、统计三大学派

四、统计学习方法

频率学派



Jerzy Neyman
1894-1982



经验分布函数

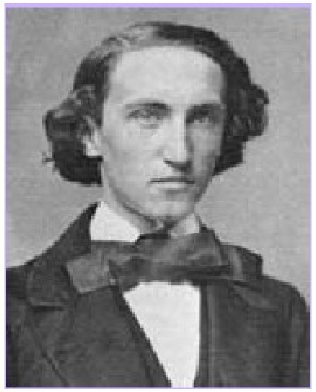
$$\hat{F}_n(t) = \frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq t},$$

大数定律

$$\hat{F}_n(t) \xrightarrow{a.s.} F(t)$$

贝叶斯学派

Thomas Bayes
1701-1761



Josiah Willard Gibbs
1839–1903

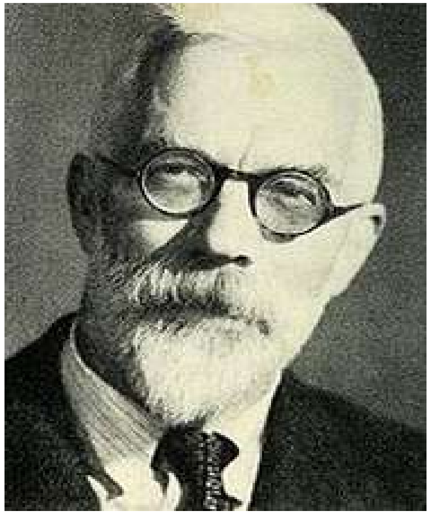
Gibbs Sampling　（贝叶斯学派典型方法）

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

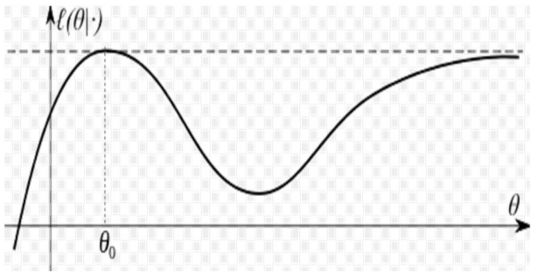
     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Fisher学派



R.A. Fisher
1890-1962



极大似然估计（MLE）

Log-likelihood: $\ln \mathcal{L}(\theta\,;\,x_1,\ldots,x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$

$$\hat{\ell} = \frac{1}{n}\ln \mathcal{L}$$

Estimator: $\{\hat{\theta}_{\text{mle}}\} \subseteq \{\underset{\theta\in\Theta}{\arg\max}\ \hat{\ell}(\theta\,;\,x_1,\ldots,x_n)\}$

Consistency: $\hat{\theta}_{\text{mle}} \xrightarrow{p} \theta_0$

内容简介

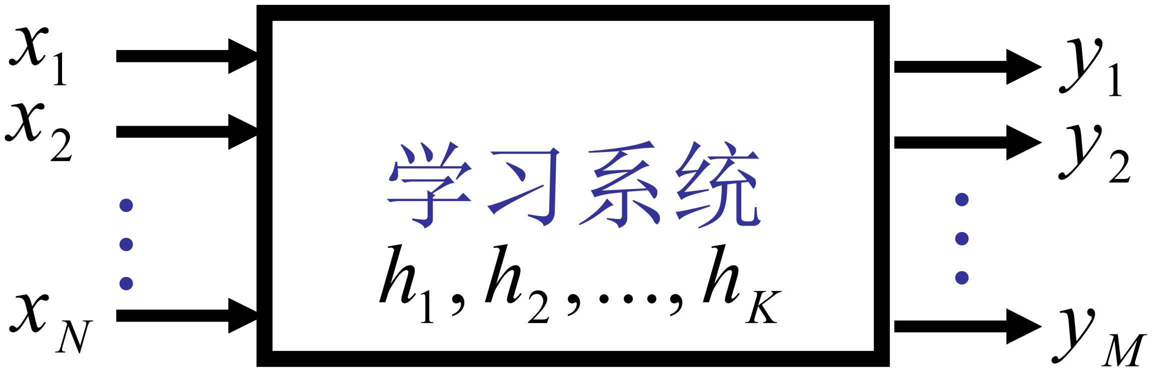
一、统计机器学习

二、学科建立历程

三、统计三大学派

四、统计学习方法

机器学习一般模型



输入变量 $\mathbf{x} = (x_1, x_2, \ldots, x_N)$

内部变量 $\mathbf{h} = (h_1, h_2, \ldots, h_K)$

观测标量 $\mathbf{y} = (y_1, y_2, \ldots, y_K)$

机器学习问题的表示

- 根据n个独立同分布观测样本确定预测函数$f(x,w)$
- 在一组函数$\{f(x,w)\}$中求一个最优的函数$f(x,w_0)$对依赖关系进行估计，使预测的期望风险最小

Performance

- Theory： generalization ability；
- Experiments： test error；

- Not only in terms of generalization;

- But also in terms of implementation, speed, understandability etc.

Theoretical Analysis

- 模型选择: estimating the performance of different models in order to choose the best one.
- 模型估计: having chosen the model, estimating the prediction error on new data.

Statistical Machine learning

- ***Try to explain the algorithms in a statistical framework.***

- Not limited to statistical **learni**ng by Vapnik.

统计机器学习的形式化定义

假定空间**Z**上有一个概率测度**F(Z)**，考虑一个函数集合**Q(Z**，α)**,** α∈Λ, 在**F(Z)**未知的情况下，给定一些观测样本$z_1, z_2, \ldots, z_n$ **(**不妨假定独立同分布**)**，定义风险函数**:**

$$R(\alpha) = \int Q(Z, \alpha)\, dF(Z)$$

我们的目标是在一致性、收敛速度、泛化能力意义下寻求风险函数**R(**α)最小。

- 学习目标

$$\alpha^* = \arg\min_{\alpha \in \Lambda} R(\alpha)$$

- 损失函数

$$L:(x,y,f_\alpha) \mapsto L(y, f(x,\alpha))$$
$$Q:(z,\alpha) \mapsto L(z_y, f(z_x,\alpha))$$

- 风险函数

$$R(\alpha) = \int Q(z,\alpha)\, dF(z)$$

风险函数

- 风险函数： $$R(\omega) = \int Q(z,\omega)dF(z)$$

- 经验风险：

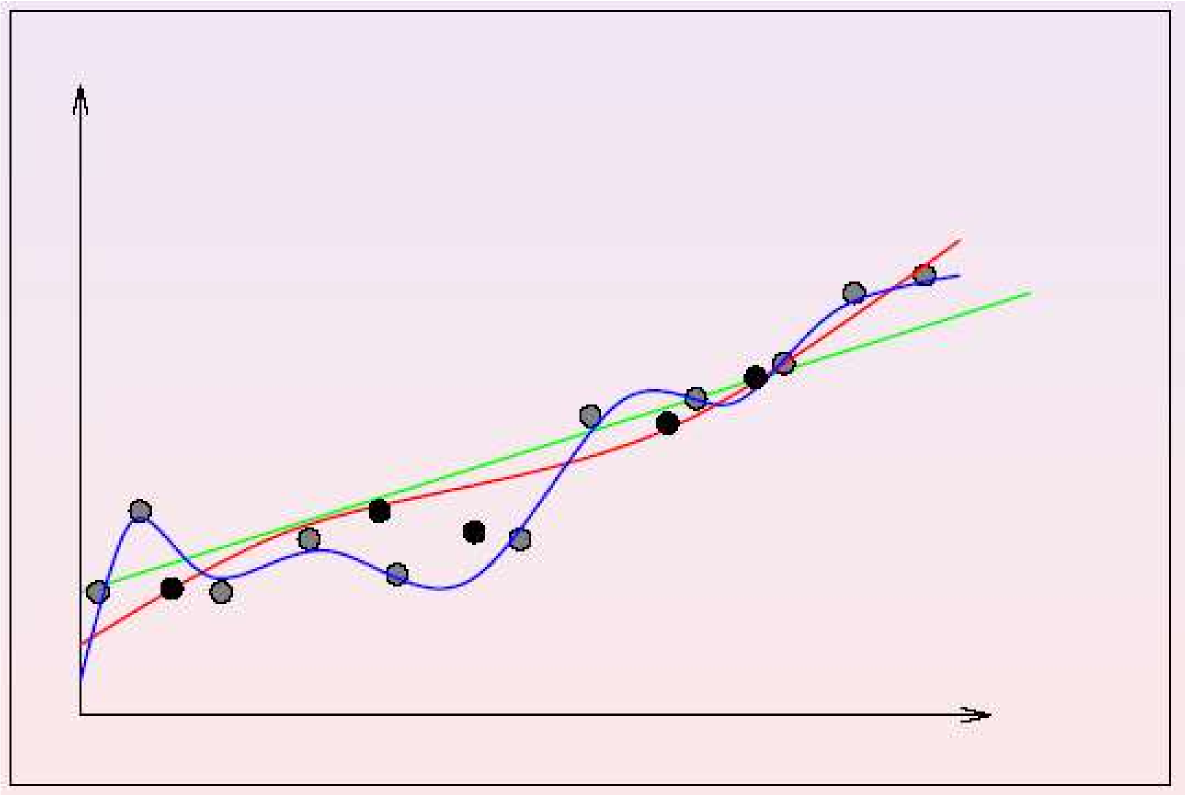
$$R_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\omega)$$

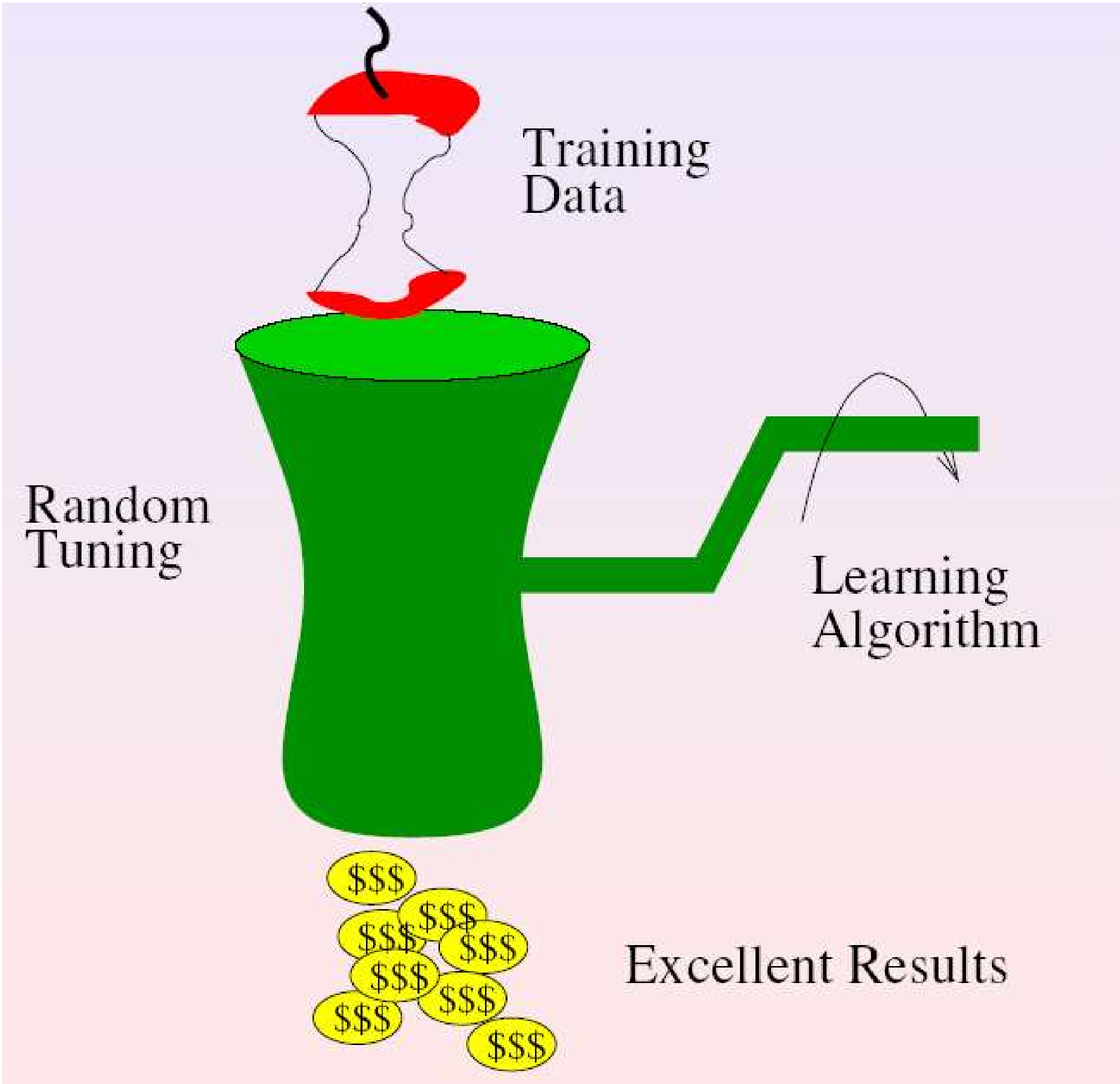
经验风险最小化准则

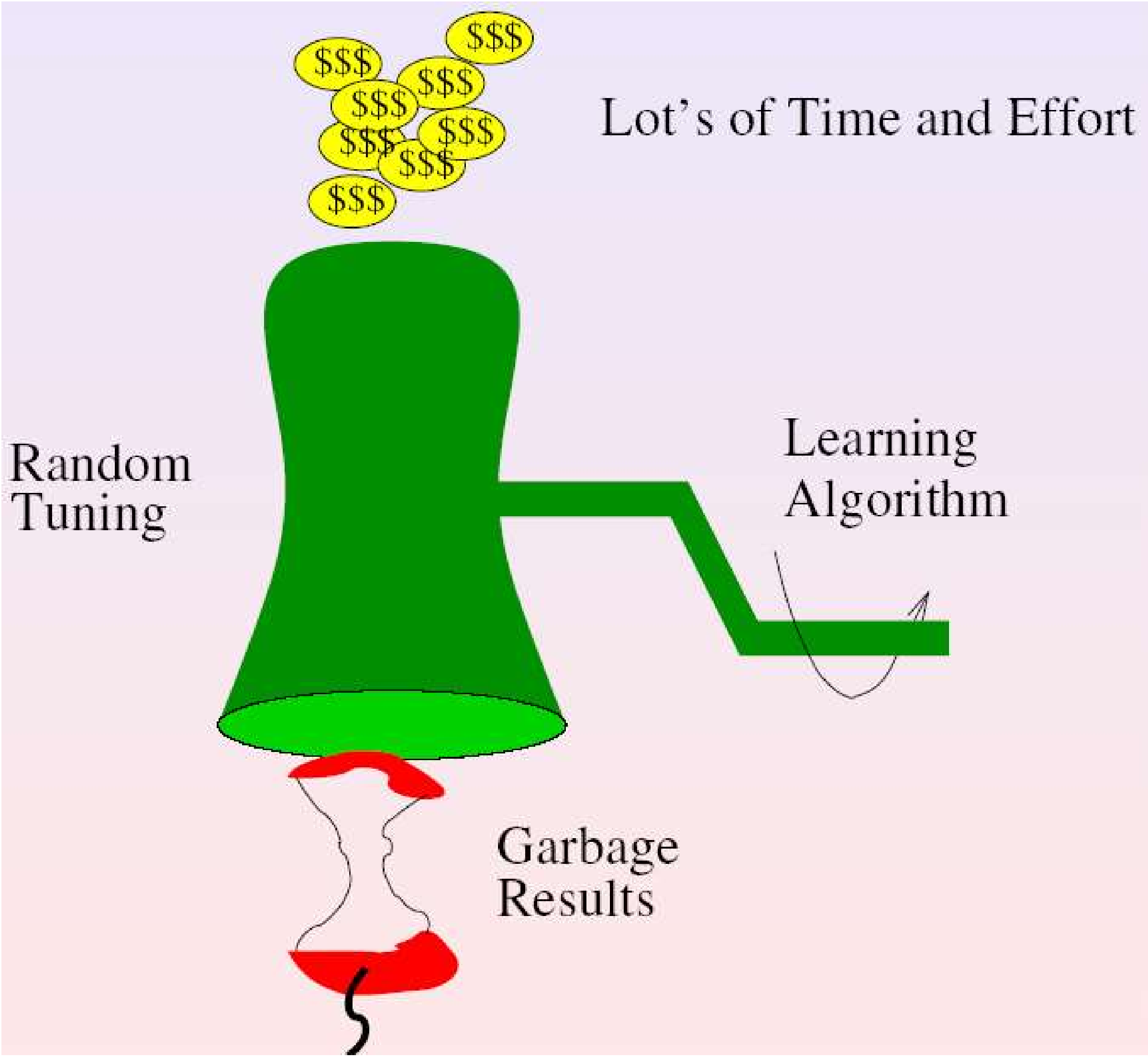
- 经验风险最小并不意谓着期望风险最小!!
  - 例子：神经网络的过学习问题。
  - 训练误差小并不总能导致好的预测效果. 若对有限的样本来说学习能力过强，足以记住每个样本，此时经验风险很快就可以收敛到很小甚至零，但却根本无法保证它对未来样本能给出好的预测.

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

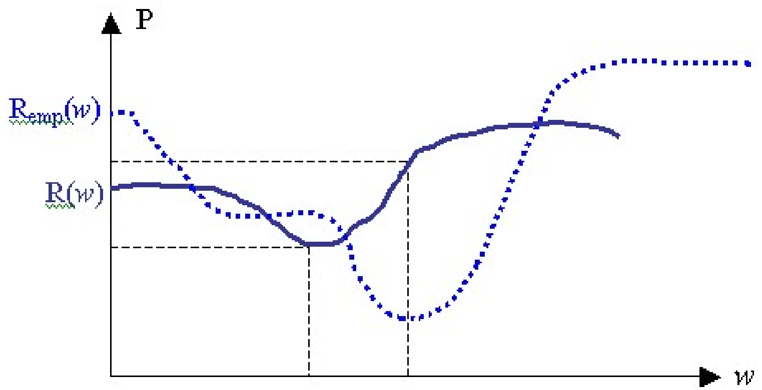
为什么统计机器学习如此困难?




Training
Data
Random
Tuning
Learning
Algorithm
$$$
$$$
$$$
$$$
$$$
$$$
$$$
$$$
Excellent Results


$$$
$$$
$$$
$$$
$$$
$$$
$$$
$$$
Lot's of Time and Effort
Random
Tuning
Learning
Algorithm
Garbage
Results

- 需要建立在小样本情况下有效的学习方法
  - 小样本条件下的统计学习理论
  - 支持向量机(**SVM**)



统计机器学习的问题

传统的统计学所研究的主要是渐进理论，而机器学的训练样本数目通常有限。

【人们过去一直采用样本数目无穷为假设条件推导各种算法，然后将算法用于样本较小的情况，希望能有较好的效果，然而，算法往往不令人满意】

由此，人们提出了学习的推广能力的重要问题。过去多数工作集中在对大样本统计学习方法的改进和修改，或利用启发式方法设计特殊算法。

Definition of Classifications

- Assumption: (x_i, y_i) i.i.d.
- Hypothesis space: H
- Loss function: $$c(y, f(x)) = \begin{cases} 0 & if \quad f(x) = y \\ 1 & if \quad f(x) \neq y \end{cases}$$
- Objective function: $$R(f) = \int c(y, f(x)) P(x, y) dx$$

Definition of regression

- Assumption: $(x_i, \ y_i)$ i.i.d.
- Hypothesis space: H
- Loss function:

$$c(y, f(x)) = \|y - f(x)\|^2$$

- Objective function:

$$R(f) = \int c(y, f(x)) P(x, y) dx$$

Several well-known algorithms

- K-Nearest Neighbor;
- LMS (Least Mean Square);
- Ridge regression;
- Fisher Discriminant Analysis;
- Neural Networks ;
- Support Vector Machines and boosting.

统计机器学习的三个核心问题

- 界的问题： 包括收敛界、泛化界、VC维
- 结构风险最小化： 包括SRM、MDL
- 实现统计机器学习的方法： 如SVM

- SML的难点——问题复杂性的估计

  经验风险最小归纳原理的一致性条件

  基于一致性条件的学习机泛化能力的界

  基于该界的小样本归纳推理原理

  实现上述归纳原理的构造方法

统计机器学习的关键技术

- ***SVM是通用的构造学习过程***

（**1**）基于统计学习理论**(Vapnik,1995)**

（**2**）概括了过去大量的学习类型**(**描述**,** 表述**)**

神经网络**,**径向基函数**,**样条**,**多项式估计

（**3**）提供一种新的函数**(**作为**)**参数的形式

（**3**）提供一种有意义的函数复杂性特征刻画， 该函数复杂性与问题的维数无关

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

感谢同学们光临！

敬请交流和指正