机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

信息与通信工程学院 网络搜索教研中心

北京邮电大学

专题 六：支持向量机与统计学习理论

- **内容提要**
  - 引言
  - 从感知器到支持向量机
  - 统计学习理论
    - 经验风险最小化
    - 结构风险最小化
  - 从结构风险最小化到支持向量机(SVM)

思考几个问题…

- 为什么要最小化训练错误数？
  - 合理吗？其合理性需要证明
- 为什么要最大化分类间隔?
  - 这一几何直观可以找到理论保证吗？其合理性需要证明
- 在学习过程中，我们的目标是什么？
  - 获得具有良好泛化能力的学习机器
- 研究学习问题的目标是什么?
  - 寻找能够达到最好的推广性能的归纳原则，并构造算法来实现这一原则
- 如何保证推广性能? 是否有其他的归纳原理，能够达到更好的推广性能？

统计学习理论(Statistical Learning Theory)

- 什么是SLT?
  - 源于统计理论与泛函分析的机器学习理论框架
    - 从数学角度论述如何控制学习机器的推广能力这一学习过程中的基本问题，引出了成功的应用——支持向量机(SVM)
- SLT回答了什么?
  - 学习过程的一致性
    - 基于经验风险最小化的学习过程一致性之充要条件
  - 学习过程收敛速度(非渐进分析)
  - 如何控制学习过程泛化能力
  - 如何构造泛化能力可控的学习机器

内容提要

- 统计学习理论简介
  - **PAC学习**
  - **VC维**
  - 损失函数、风险泛函和经验风险泛函
  - 经验风险最小化原则
  - 经验风险最小化原则的一致性条件
  - 收敛速度
  - 学习机器的推广能力的界
  - 结构风险最小化原则
    - 结构风险最小化与SVM

PAC学习

- 概率近似正确(Probably Approximately Correct): PAC
  - L. Valiant. A theory of the learnable. Communications of the ACM, 27, 1984.
- PAC学习:
  - 给出了研究机器学习问题的基本理论框架
  - **The learner receives samples and must select a function from a certain class of possible functions. The goal is that, with high probability (the "probably" part), the selected function will have low generalization error (the "approximately correct" part).**
    - The learner is expected to find **efficient** functions and must implement an **efficient** procedure
      - » Requirements are bounded to a polynomial of the sample size

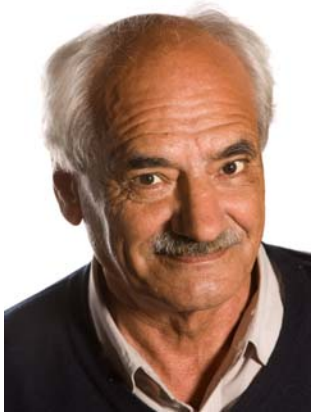
VC dimension

- [Vapnik and Chervonenkis 1968, 1971]

The VC dimension of a set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$, *is the maximum number* h *of vectors* $z_1, \ldots, z_h$ *that can be separated into two classes in all* 2^h *possible ways using functions of the set*[1] *(i.e., the maximum number of vectors that can be* shattered *by the set of functions). If for any* n *there exists a set of* n *vectors that can be shattered by the set* $Q(z, \alpha), \alpha \in \Lambda$, *then the VC dimension is equal to infinity.*

  - **VAPNIK, V. N., and A. Ya. CHERVONENKIS, 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. Theory of Probability and Its Applications, 16(2), 264–280. Translated by B. Seckler.**

➢ VC 维量度指示函数集的"消化能力".



VC维

- 物理含义
  - 对由学习机器所实现的分类函数族的容量(capacity) 的测度
- 定义
  - 二分类函数族 Ψ 的VC维是被 Ψ 所分散(shatter)的最多数据点数目
    - **其中，“二分类函数族Ψ”是指所有可能二值分类函数的集合:**

$$\Psi = \left\{ F(\mathbf{x}, \mathbf{w}), \mathbf{w} \in W, F : W \to \{0,1\} \right\}$$

    - **输入数据为m维空间中的样本**: $\left\{ (\mathbf{x}_i, y_i) \right\}_{i=1}^{N}$
    - **所有可能的二分：是指对数据点进行任意的标签指派(assignment)**
- 解释
  - 能被“学习机器”所“学习”的训练样本的最大数目，其中所说的“学习”是指对所有可能的二分标记可以分类无误
- 举例
  - 高阶多项式判别函数：阶数越高，VC dim 越大
  - 单参圆函数族，双参圆函数族，m+2参的圆函数族
  - m维空间中的线性判别函数族，VC dim(F) = m+1
    - **对VC维的直观认识：自由参数少的学习机器其VC一定就小么？ No.**

VC维的重要性及其估计

- VC维概念的重要性
  - 理论上
    - 是统计学习理论的一个中心角色，用于预测学习机器的泛化误差的上界(probabilistic upper bound)
  - 实践上
    - 从设计观点看，给定一个学习机器(即分类函数族)，使得该机器可靠地学习一个数据集所需样本数正比于学习机器的VC维
- VC维概念的估计
  - 在大多数实际情况下，很难通过分析的手段计算**VC**维
  - 对于神经网络的**VC**维的界，有两个结论
    - 1. 令N表示由神经元构成的任意前馈网络，激活函数为硬门限型，N的VC维为$O(W \log W)$，其中W是网络中自由参数的总数
    - 2. 令N表示一个多层前馈网络，其神经元使用一个sigmoid激活函数，N的VC维为$O(W^2)$，其中W是网络中自由参数的总数
  - 结论：多层前馈网络具有有限的**VC**维

统计学习问题描述

- **基本模型**
  - 给定观测数据及目标输出构成的训练数据集 $D=\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$，构造一个学习机器(即训练一个分类函数族)，实现输入 - 输出映射函数 $F:\mathbf{x}\to y$

- **问题实质**
  - 选择一个在统计意义下最优的函数 $F(\mathbf{x},\mathbf{w})$ 用来逼近目标输出**y**
    - 基于训练样本集，构建一个具有良好推广性能的学习机器
      - 优化过程基于**N**个独立同分布**(i.i.d.)**的训练样本来完成

损失函数

- 定义为函数$F(\mathbf{x},\mathbf{w})$与目标输出 y 之差的非负函数:

$$\ell\Big(f(x),F(\mathbf{x},\mathbf{w})\Big)=\ell\Big(y,F(\mathbf{x},\mathbf{w})\Big)$$

  - **度量由学习机器实际产生的输出$F(\mathbf{x},\mathbf{w})$与目标输出之间的差异**
    - 这里介绍的统计学习理论不严格依赖于损失函数$\ell\Big(y,F(\mathbf{x},\mathbf{w})\Big)$的具体定义形式

风险泛函与经验风险泛函

- 风险泛函
  - 定义为损失的期望 $R(\mathbf{w})=\int \ell\left(y,F(\mathbf{x},\mathbf{w})\right)dP_{X,Y}(\mathbf{x},y)$

    其中，联合累积分布函数 $P_{X,Y}(\mathbf{x},y)$通常是未知的
    - 监督学习的目标：在$\{F(\mathbf{x},\mathbf{w}):\mathbf{w}\in W\}$ 上寻找逼近函数，以最小化风险泛函$\boldsymbol{R(\mathbf{w})}$

- 经验风险泛函
  - 给定训练样本集**D**, 基于损失函数$\ell(\cdot)$ 的经验风险泛函定义为:

$$R_{emp}(\mathbf{w})=\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i,F(\mathbf{x}_i,\mathbf{w})\right)$$

➢ 监督学习中唯一可用的信息包含在训练数据 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^{N}$ 中，为克服最小化风险泛函在数学上的困难，我们采用经验风险最小化归纳原则

经验风险最小化(Empirical Risk Minimization)

- ERM的基本过程
  - 通过构建经验风险泛函来代替风险泛函，通过在训练数据集$D=\{(\mathbf{x}_i,y_i)\}_{i=1}^N$上最小化经验风险泛函$R_{emp}(\mathbf{w})$来寻找逼近函数$F(\mathbf{x},\mathbf{w})$，即寻找$F(\mathbf{x},\mathbf{w}_{emp})$，其中

$$\mathbf{w}_{emp}=\arg\min_{\mathbf{w}} R_{emp}(\mathbf{w})=\arg\min_{\mathbf{w}}\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i,F(\mathbf{x}_i,\mathbf{w})\right)$$

?

  - 监督学习的理论目标: 在$\{F(\mathbf{x},\mathbf{w}):\mathbf{w}\in W\}$上寻找逼近函数$F(\mathbf{x},\mathbf{w})$以最小化风险泛函$R(\mathbf{w})$，即寻找$F(\mathbf{x},\mathbf{w}_o)$，其中

$$\mathbf{w}_o=\arg\min_{\mathbf{w}} R(\mathbf{w})=\arg\min_{\mathbf{w}}\int\ell\left(y,F(\mathbf{x},\mathbf{w})\right)dP_{X,Y}(\mathbf{x},y)$$

经验风险泛函 vs. 风险泛函

- 经验风险泛函$R_{emp}(\mathbf{w})$与风险泛函$R(\mathbf{w})$的差异
  - 经验风险泛函$R_{emp}(\mathbf{w})$不显式地依赖未知的分布函数$P_{X,Y}(\mathbf{x},y)$，因此利用$R_{emp}(\mathbf{w})$能够对权值向量w最小化
  - 没有理由指望最小化经验风险泛函$R_{emp}(\mathbf{w})$的权值向量$R(\mathbf{w}_{\text{emp}})$同样会最小化风险泛函$R(\mathbf{w})$
    - **如果固定w=w*，则**
      - 风险泛函**R(w*)**变为下述随机变量$Z_{\mathbf{w}^*}$的数学期望，其中
      $$Z_{\mathbf{w}^*}=\ell\left(y,F(\mathbf{x},\mathbf{w}^*)\right)$$
      - 而经验风险泛函$\mathbf{R_{emp}(w^*)}$是随机变量$\mathbf{Z_{w^*}}$算术平均值

经验风险最小化原则(ERM)

- ERM原则
  - 1. 可以构建经验风险泛函 $R_{emp}(\mathbf{w})$ 代替风险泛函 $R(\mathbf{w})$, 即基于i.i.d.训练样本 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^{N}$, 定义

$$\mathbf{w}_{emp}=\arg\min_{\mathbf{w}} R_{emp}(\mathbf{w})=\arg\min_{\mathbf{w}}\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i,F(\mathbf{x}_i,\mathbf{w})\right)$$

  - 2. 只要当训练样本的数量N趋于无穷大时，经验风险泛函**一致收敛**于实际风险泛函，那么，最小化经验风险 $R_{emp}(\mathbf{w})$ 的 $\mathbf{w}_{\text{emp}}$ 所对应的实际风险 $R(\mathbf{w}_{\text{emp}})$ **依概率收敛**到实际风险 $R(\mathbf{w})$ 的最小可能值，即 $\mathrm{P}\left\{R(\mathbf{w}_{emp})-R(\mathbf{w}_o)<2\varepsilon\right\}>1-\alpha$

经验风险最小化原则的一致性条件

- ERM原则的一致性充要条件
  - 即，经验风险泛函一致收敛于风险泛函
  - 当训练样本数**N**趋于无穷时，

$$\lim_{N\to\infty} \mathrm{P}\left(\sup_{\mathbf{w}\in W} \left| R(\mathbf{w}) - R_{emp}(\mathbf{w}) \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0$$

当N有限时，如何准确刻画收敛过程?

- 推导:

$$\mathrm{P}\left(\sup_{\mathbf{w}\in W} \left| R(\mathbf{w}) - R_{emp}(\mathbf{w}) \right| > \varepsilon \right) \to 0$$

对于任意 α **>0**

$$P\left(\sup_{\mathbf{w}\in W} \left| R(\mathbf{w}) - R_{emp}(\mathbf{w}) \right| < \varepsilon \right) > 1-\alpha \quad \text{成立}$$

以**1 -** α 的概率，下面两式同时成立

$$\begin{cases} R(\mathbf{w}_{emp}) - R_{emp}(\mathbf{w}_{emp}) < \varepsilon \\ R_{emp}(\mathbf{w}_o) - R(\mathbf{w}_o) < \varepsilon \end{cases}$$

注意到不等式

$$R_{emp}(\mathbf{w}_{emp}) \le R_{emp}(\mathbf{w}_o)$$

以**1 -** α 的概率，结论成立

$$R(\mathbf{w}_{emp}) - R(\mathbf{w}_o) < 2\varepsilon$$

即 $$\mathrm{P}\left\{ R(\mathbf{w}_{emp}) - R(\mathbf{w}_o) < 2\varepsilon \right\} > 1-\alpha$$

内容提要

- 统计学习理论简介
  - **PAC学习**
  - **VC维**
  - 损失函数、风险泛函和经验风险泛函
  - 经验风险最小化原则
  - 经验风险最小化原则的一致性条件
  - 收敛速度
  - 学习机器的推广能力的界
  - 结构风险最小化原则
    - 结构风险最小化与SVM

训练错误频率 vs. 分类错误概率

- 考虑**二值分类**, 期望响应为{0,1}

  – 损失函数定义为：$\ell\left(y, F\left(\mathbf{x}, \mathbf{w}\right)\right) = \begin{cases} 0, & \text{if } F\left(\mathbf{x}, \mathbf{w}\right) = y \\ 1, & \text{else} \end{cases}$

  则风险泛函 $R(\mathbf{w})$ 即分类错误概率，记为 $\mu(\mathbf{w})$

  经验风险泛函 $R_{emp}(\mathbf{w})$ 即训练误差，即训练阶段发生错误频率，记为 $v_N(\mathbf{w})$

  - 根据大数定律，事件发生的经验频率几乎一定收敛于那一事件的实际概率，只要试验(假设是独立同分布)数目趋于无穷大: 注意到这个结论的前提是w固定

训练错误频率 vs. 分类错误概率

- 考虑**二值分类**, 期望响应为{0,1}
  - 损失函数定义为：$\ell\left(y,F\left(\mathbf{x},\mathbf{w}\right)\right)=\begin{cases}0, & \text{if } F\left(\mathbf{x},\mathbf{w}\right)=y\\ 1, & \text{else}\end{cases}$

则风险泛函 $R(\mathbf{w})$ 即分类错误概率，记为 $\mu(\mathrm{W})$

经验风险泛函 $R_{emp}(\mathbf{w})$ 即训练误差，即训练阶段发生错误频率，记为 $\nu_N(\mathrm{W})$

  - **根据大数定律，事件发生的经验频率几乎一定收敛于那一事件的实际概率，只要试验(假设是独立同分布)数目趋于无穷大: 注意到这个结论的前提是w固定**

- **对于任意$\varepsilon>0$，如果当N趋于无穷大时，有**

$$P\left(\sup_{\mathbf{w}\in W}\left|\nu_N\left(\mathbf{w}\right)-\mu\left(\mathbf{w}\right)\right|>\varepsilon\right)\to 0,$$

则称 训练误差(频率)到分类错误概率存在一致收敛

ERM原则一致性的充要条件：经验风险泛函一致收敛到实际风险泛函

使用Hoeffding不等式

- 令 $S_N = \frac{1}{N}\sum_{i=1}^{N} \ell\left(y_i, F\left(\mathbf{x}_i, \mathbf{w}\right)\right)$

$$\ell\left(y_i, F\left(\mathbf{x}_i, \mathbf{w}\right)\right) \in \{0,1\} \subset [0,1]$$

– 根据**Hoeffding**不等式，我们得到

$$\mathbf{P}\left\{\left|S_N - \mathbf{E}\left[S_N\right]\right| \geq t\right\} \leq 2\exp\left(\frac{-2t^2 N}{\frac{1}{N}\sum_{i=1}^{N}\left(b_i - a_i\right)^2}\right)$$

– 即

$$\mathbf{P}\left\{\left|\nu_N\left(\mathbf{w}\right) - \mu\left(\mathbf{w}\right)\right| \geq \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 N\right)$$

从Hoeffding不等式到一致收敛

$$\mathbf{P}\left\{\sup_{\mathbf{w}\in W}\left|v_N(\mathbf{w})-\mu(\mathbf{w})\right|\geq\varepsilon\right\}=\mathbf{P}\left\{\left|v_N(\mathbf{w}_1)-\mu(\mathbf{w}_1)\right|\geq\varepsilon\cup\ldots\cup\ldots.\left|v_N(\mathbf{w}_M)-\mu(\mathbf{w}_M)\right|\geq\varepsilon\right\}$$

$$\leq\mathbf{P}\left\{\left|v_N(\mathbf{w}_2)-\mu(\mathbf{w}_2)\right|\geq\varepsilon\right\}\ldots+\ldots$$

$$+\mathbf{P}\left\{\left|v_N(\mathbf{w}_M)-\mu(\mathbf{w}_M)\right|\geq\varepsilon\right\}$$

$$\leq 2M\exp\left(-2\varepsilon^2N\right)$$

函数集合中不同函数的数目

如果M是有限的，则当N趋于无穷大，则右侧收敛于0

– 即 对于任意 $\varepsilon>0$，如果当N趋于无穷大时，有

$$\mathbf{P}\left\{\sup_{\mathbf{w}\in W}\left|v_N(\mathbf{w})-\mu(\mathbf{w})\right|>\varepsilon\right\}\to 0,$$ 等价于

$$\mathbf{P}\left\{\sup_{\mathbf{w}\in W}\left|R(\mathbf{w})-R_{emp}(\mathbf{w})\right|>\varepsilon\right\}\to 0$$

经验风险泛函一致收敛到实际风险泛函 （ERM原则一致性的充要条件）

增长函数 S

- 分类函数族 $\Psi = \{F(\mathbf{x},\mathbf{w}), \mathbf{w} \in W, F: W \to \{0,1\}\}$ 的增长函数 S_N 定义为:

$$\max_{\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\} \subseteq X} \left| \{F(\mathbf{x}_1,\mathbf{w}), F(\mathbf{x}_2,\mathbf{w}), ..., F(\mathbf{x}_N,\mathbf{w}) : \mathbf{w} \in W\} \right|$$

  - 增长函数S_N表征学习机器(即分类函数族)对N个样本所能赋予二分标签的最大可能结果数

一致收敛的条件

- Sauer's 推论:
  - 增长函数 **S** (也称为 打散系数**Shatter Coefficient**)是样本数**N**的多项式

$$S_N \le (N+1)^h, \qquad S_N \le \left(\frac{eN}{h}\right)^h$$

    - **其中h为学习机器的VC维(即分类函数集的容量)，N是训练样本数目，e为自然对数的底**
- VC不等式: $\mathbf{P}\left(\sup_{\mathbf{w}\in W}\left|v_N(\mathbf{w})-\mu(\mathbf{w})\right|\ge\varepsilon\right)\le 8S_N\exp\left(-\frac{N\varepsilon^2}{32}\right)$
- **借助VC不等式和Sauer推论**

$$\mathbf{P}\left(\sup_{\mathbf{w}\in W}\left|v_N(\mathbf{w})-\mu(\mathbf{w})\right|\ge\varepsilon\right)\le 8(N+1)^h\exp\left(-\frac{N\varepsilon^2}{32}\right)$$

$$\mathbf{P}\left(\sup_{\mathbf{w}\in W}\left|v_N(\mathbf{w})-\mu(\mathbf{w})\right|\ge\varepsilon\right)\le 8\left(\frac{eN}{h}\right)^h\exp\left(-\frac{N\varepsilon^2}{32}\right)$$

只要h是有限的，则右侧会随N趋于无穷大而趋于零！

学习机器推广能力的界*

- 令α表示事件 $\sup_{\mathbf{w}\in W}\left|v_N(\mathbf{w})-\mu(\mathbf{w})\right|\geq\varepsilon$ 的发生概率，那么以概率1−α，
  对于所有权值向量，有： $\mu(\mathbf{w})<v_N(\mathbf{w})+\varepsilon$
  - 结合上面描述的界和概率**α**的定义，可以令 $\alpha=8\left(\frac{eN}{h}\right)^h\exp\left(-\frac{N\varepsilon^2}{32}\right)$
  - 由此，可以得出置信区间： $\varepsilon(N,h,\alpha)=\sqrt{\frac{32}{N}\left(\ln 8+h\ln\left(\frac{eN}{h}\right)-\ln\alpha\right)}$
  - 其中 $\varepsilon(N,h,\alpha)$ 即为一致收敛速度的界

- **对于分类错误概率(风险泛函，泛化误差) μ(w) ，存在界**

$$\mu(\mathbf{w})\leq v_N(\mathbf{w})+\varepsilon(N,h,\alpha)$$

内容提要

- 统计学习理论简介
  - **PAC学习**
  - **VC维**
  - 损失函数、风险泛函和经验风险泛函
  - 经验风险最小化原则
  - 经验风险最小化原则的一致性条件
  - 收敛速度
  - 学习机器的推广能力的界
  - 结构风险最小化原则
    - 结构风险最小化与SVM

如何控制推广能力

- 保证风险 $v_{guarant}(\mathbf{w})$
  - 定义为两个竞争项的和，即 $v_{guarant}(\mathbf{w}) = v_N(\mathbf{w}) + \varepsilon_1(N, h, \alpha, v_{train})$

    训练误差：学习机器在训练阶段出现错误的频率 $v_N(\mathbf{w})$

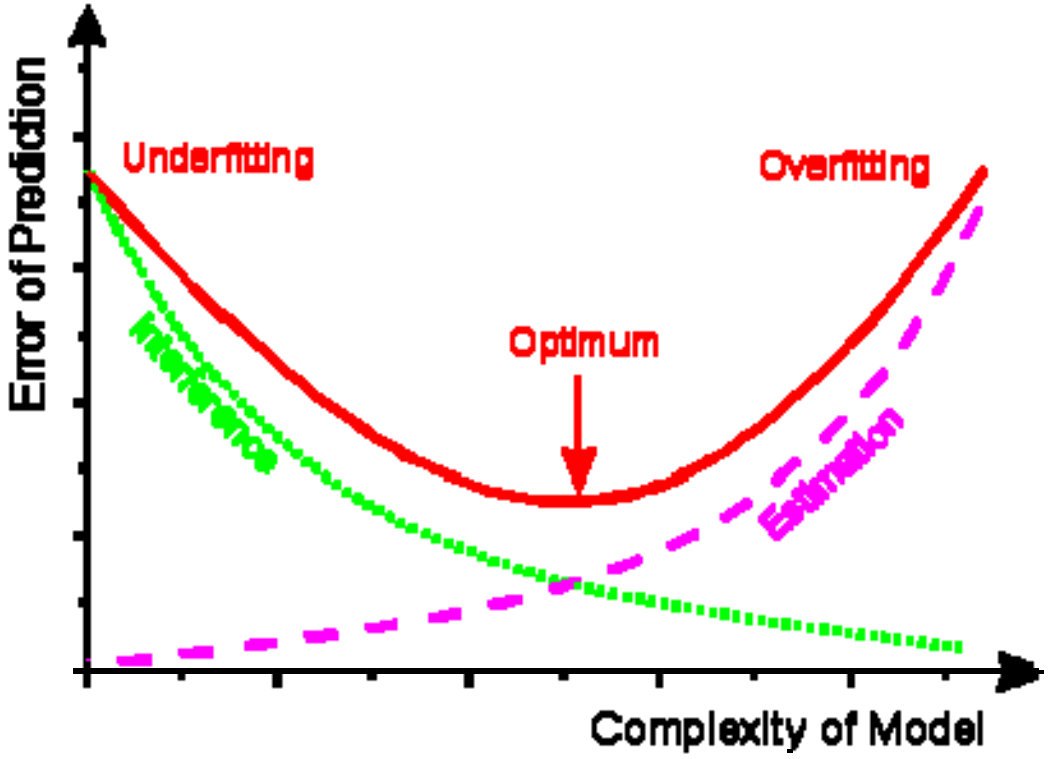
    泛化误差：学习机器在未见样本上测试时所犯错误的概率 $\mu(\mathbf{w})$
  - 令h为分类函数族 $\{F(\mathbf{x}, \mathbf{w}); \mathbf{w} \in \mathbb{W}\}$ 的VC维，根据一致收敛速度理论，可知：
    - **以概率1 - α，对于训练样本数量N > h，对于所有分类函数F(x,w)，泛化误差 $\mu(\mathbf{w})$ 比保证风险 $v_{guarant}(\mathbf{w})$ 小**
  - 因此，通过最小化保证风险来设计学习机器，能够获得良好泛化能力

- 几点说明：
  - 上述结论是在统计意义下给出的,α是显著性水平,ε_1是“置信区间”
  - 上述结论是一般情况下，一致收敛速度的界的另一种说法

$$\mu(\mathbf{w}) \le v_N(\mathbf{w}) + \varepsilon_1(N, h, \alpha, v_N)$$

训练误差 vs. 模型复杂度

– 对于固定训练样本数N，训练误差是h的单调减函数，置信区间则单调递增，因此保证风险和泛化误差都经历最小值



结构风险最小化(SRM)

- 解决监督学习问题的泛化能力问题的基本思想
  - 使学习机器的容量与训练数据的有效数量相匹配

- 解决监督学习问题的泛化能力问题的方法
  - 结构风险最小化方法
  - 通过使学习机器的VC维成为控制变量来提供一个归纳过程，以达到使得学习机器的容量与训练数据的有效数量相匹配的目标

- **结构风险最小化方法的实质：**
  - 以训练误差最小可能增加为代价，换取VC维的降低

结构风险最小化(SRM)

- 考虑模式分类器的集合 $\Psi = \{F(\mathbf{x}, \mathbf{w}); \mathbf{w} \in \Theta\}$ 定义n个学习机器的嵌套结构

$$\Psi_t = \{F(\mathbf{x}, \mathbf{w}); \mathbf{w} \in \Theta_t\}, t = 1, ..., n$$

使得 $\Psi_1 \subset \Psi_2 \subset \cdots \subset \Psi_n$，各个学习机器的VC维满足递增条件 $h_1 \le h_2 \le \cdots \le h_n$

那么，结构风险最小化方法可以按如下方式进行：

1. **对于每个学习机器 Ψ_t，最小化其经验风险 (即最小化训练误差)**
2. **确定具有最小保证风险的学习机器 (即提供训练误差与学习机器复杂度之间的最好折衷)**

结构风险最小化方法的应用: 1

- 对于神经网络设计问题
  - 神经网络的**VC**维的界
    - 1. 令N表示由神经元构成的任意前馈网络，激活函数为硬门限型，N的VC维为$O(W \log W)$，其中W是网络中自由参数的总数
    - 2. 令N表示一个多层前馈网络，其神经元使用一个sigmoid激活函数，N的VC维为$O(W^2)$，其中W是网络中自由参数的总数
  - 可以看出：
    - 通过改变隐藏层神经元的个数可以改变VC维
    - 通过改变神经网络中连接的数目可以改变VC维
  - 让隐藏层神经元数量以单调方式增加，从而构成**VC**维满足递增条件的学习机器(即分类函数族)的嵌套结构

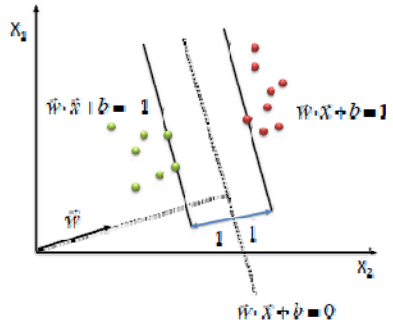
结构风险最小化方法的应用: 2

- 支持向量机
  - 构造VC维可变的分离超平面集合，使经验风险(即训练分类误差)和VC维同时最小化

- 最大间隔分离超平面的VC维定理:
  - 令D表示包含所有输入向量$\{x_1, x_2, \ldots, x_N\}$的最小球的直径，则由方程 $\mathbf{w}_o^T \mathbf{x} + b_o = 0$ 所描述的最优超平面集合，有一个VC维数h的上界

$$h \le \min\left\{m, \left\lceil \frac{D^2}{\rho^2} \right\rceil\right\} + 1$$

其中，m为输入空间的维数，$\rho = \frac{2}{\|\mathbf{w}_o\|_2}$



  - 通过控制超平面的分离间隔，可以控制最优超平面的VC维，使其与输入空间的维数m无关！

结构风险最小化方法的应用: 2

- 支持向量机
  - 构造VC维可变的分离超平面集合，使经验风险(即训练分类误差)和VC维同时最小化
  - 定义分离超平面的嵌套结构如下:

$$S_t = \left\{ \mathbf{w}^T \mathbf{x} + b : \|\mathbf{w}\|_2^2 \le c_t \right\}, t = 1, 2, ..., n.$$ 满足 $h_1 \le h_2 \le \cdots \le h_n$

- 具有最大分离间隔的最优超平面很自然地构成VC维递增的学习机器(即分类函数族)的嵌套结构，从而满足结构风险最小化要求

Nothing is more practical than a good theory.

Vapnik, 1992， 1995, 1998

Q / A

- Any Question? …