人工智能
ARTIFICIAL INTELLIGENCE

主讲：鲍军鹏 博士
西安交通大学电信学院计算机系
电子邮箱：dr.baojp@googlemail.com

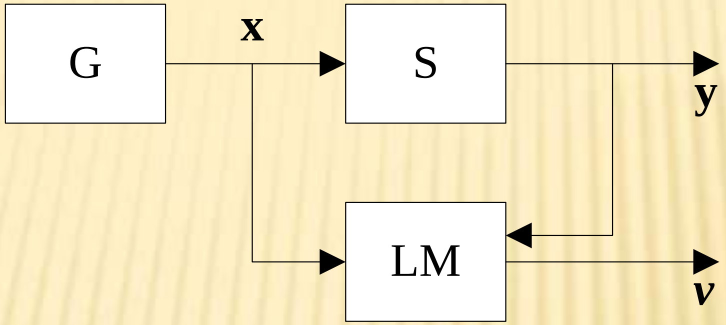
版本：2.0
2010年1月

6.4 统计学习

- 传统的统计学理论，即Fisher理论体系的前提条件
  - 已知准确的样本分布函数
  - 并且采样无穷多为
- V. Vapnik提出小样本（有限样本）统计学习理论
  - 小样本统计学习理论基于对学习错误（过学习，overfitting)和泛化能力之间关系的定量刻画，
  - 不仅避免了对样本点分布的假设和数目要求，
  - 还产生了一种新的统计推断原理——结构风险最小化原理。

6.4.1 统计学习理论

- 函数估计模型



（1）G表示产生器，用于产生输入向量x；

（2）S表示被观测的系统或者称为训练器。训练器对每个输入x产生相应的输出y，并且输入和输出遵从某个未知联合概率F(x,y)；

（3）LM表示学习机。学习机能够实现一定的函数集f(x,*a*)，*a*∈Λ，其中Λ是学习参数集合，学习参数既可能是向量也可能是函数。不同的*a*值就决定了不同的学习函数。

- 学习的问题就是从给定的函数集f(x,*a*)，*a*∈Λ中选择出能最好地逼近训练器响应的函数。

期望风险

- 损失的数学期望值就称为风险泛函（risk functional），也称为期望风险。

$$R(a)=\int L(y,f(x,a))dF(x,y)$$

- 学习的目标就是最小化风险泛函R(a)，即风险最小化问题。

经验风险

- 实际问题中，联合概率F(x,y)是未知的，所以就无法用风险泛函直接计算损失的期望值，也无法最小化。于是实践中常用算术平均代替数学期望，从而得到经验风险泛函

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

- 当N→∞时，经验风险Remp(a)才在概率意义下趋近于期望风险R(a)。传统的学习方法大多都是使经验风险最小化（Empirical risk minimization，ERM）。

小样本统计学习理论

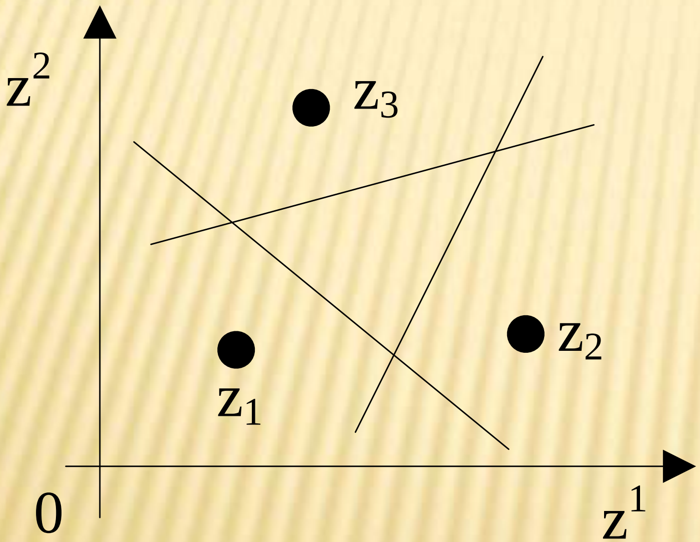
- 即使样本数目很大，也不能保证经验风险的最小值与期望风险的最小值相近。
- 所以统计学习理论就要研究在样本数目有限的情况下，经验风险与期望风险之间的关系。其核心内容包括一下4点：
  - 在什么条件下，当样本数目趋于无穷时，经验风险Remp(a)最优值趋于期望风险R(a)最优值（能够推广），其收敛速度又如何。也就是在经验风险最小化原则下的学习一致性条件。
  - 如何从经验风险估计出期望风险的上界，即关于统计学习方法推广性的界。
  - 在对期望风险界估计的基础上选择预测函数的原则，即小样本归纳推理原则。
  - 实现上述原则的具体方法。例如支持向量机（Support vector machine，SVM）就是一个具体的方法。

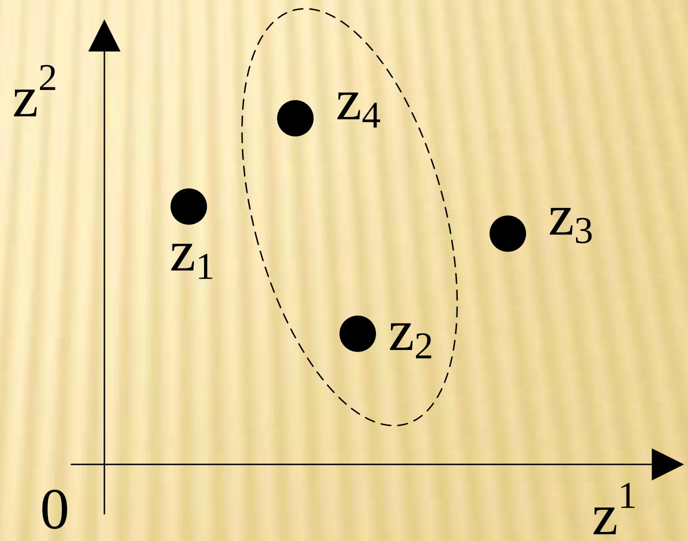
VC维

- VC维的直观定义：
  - 对一个指示函数集，如果存在h个样本能够被函数集中的函数按所有可能的2h种形式分开，则称函数集能够把h个样本打散。函数集的VC维就是它能打散的最大样本数目h。
- 所谓打散就是不管全部样本如何分布，总能在函数集中找到一个函数把所有样本正确地分为两类。
  - 若对任意数目的样本都有函数能将它们打散，则函数集的VC维是无穷大。
  - 有界实函数的VC维可以通过用一定的阈值将它转化成指示函数来定义。

实数平面的VC维

- 实际上n维超平面的VC维是n+1。





- **定理6.2** 对于R^n中的m个点集，选择任何一个点作为原点，m个点能被超平面打散当且仅当剩余点的位置向量是线性独立的。
- **推论** R^n中有向超平面集的VC维是n+1。
  - 因为总能找出n+1个点，选择其中一个作为原点，剩余n个点的位置向量是线性独立的。但无法选择n+2个这样的点，因为在Rn中没有n+2个向量是线性独立的。
- VC维反映了函数集的学习能力
  - VC维越大则学习机器越复杂，容量越大。
- 线性函数的VC维等于其自由参数的个数。
  - 但是一般来说，函数集的VC维与其自由参数的个数不相同。
- 实际上，影响学习机器推广性能的是函数集的VC维，而不是其自由参数个数。
  - 这给我们克服“维数灾难”创造了一个很好的机会：用一个包含很多参数，但却有较小VC维的函数集为基础构造学习机器会实现较好的推广性。

结构风险

- 对于两类分类问题：
  - 指示函数集中的所有函数（包括使经验风险最小的函数），经验风险Remp(a)和期望风险R(a)之间以至少1-η的概率满足如下关系：

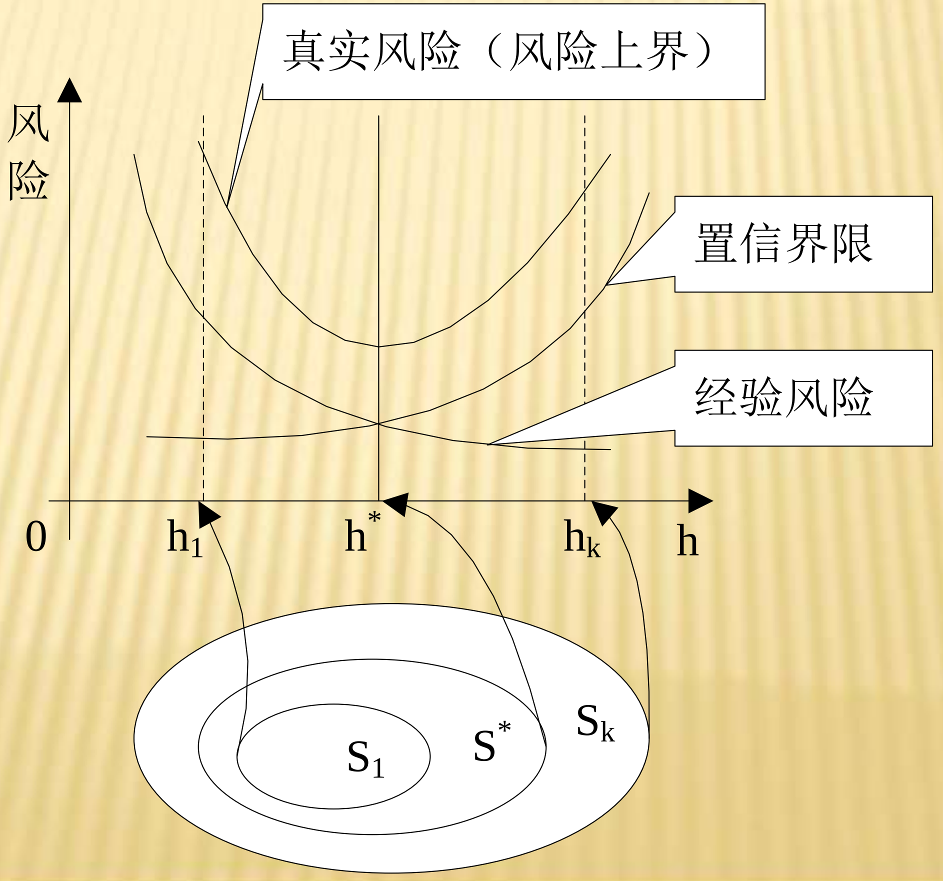
$$R(a) \le R_{emp}(a) + \sqrt{\frac{h(\ln(2N/h)+1)-\ln(\eta/4)}{N}}$$

$$R(a) \le R_{emp}(a) + \Phi(h/N)$$

- 它表明，在有限的训练样本下，学习机器的VC维越高，复杂性越高，则置信范围越大，从而导致真实风险与经验风险之间可能的差别越大。
- 由以上结论可知，ERM原则在样本有限时是不合理的

结构风险最小化原则

- 在同一子集中置信界限相同；在每一个子集中寻找最小经验风险；最后在不同子集间综合考虑经验风险和置信界限，使得真实风险最小。



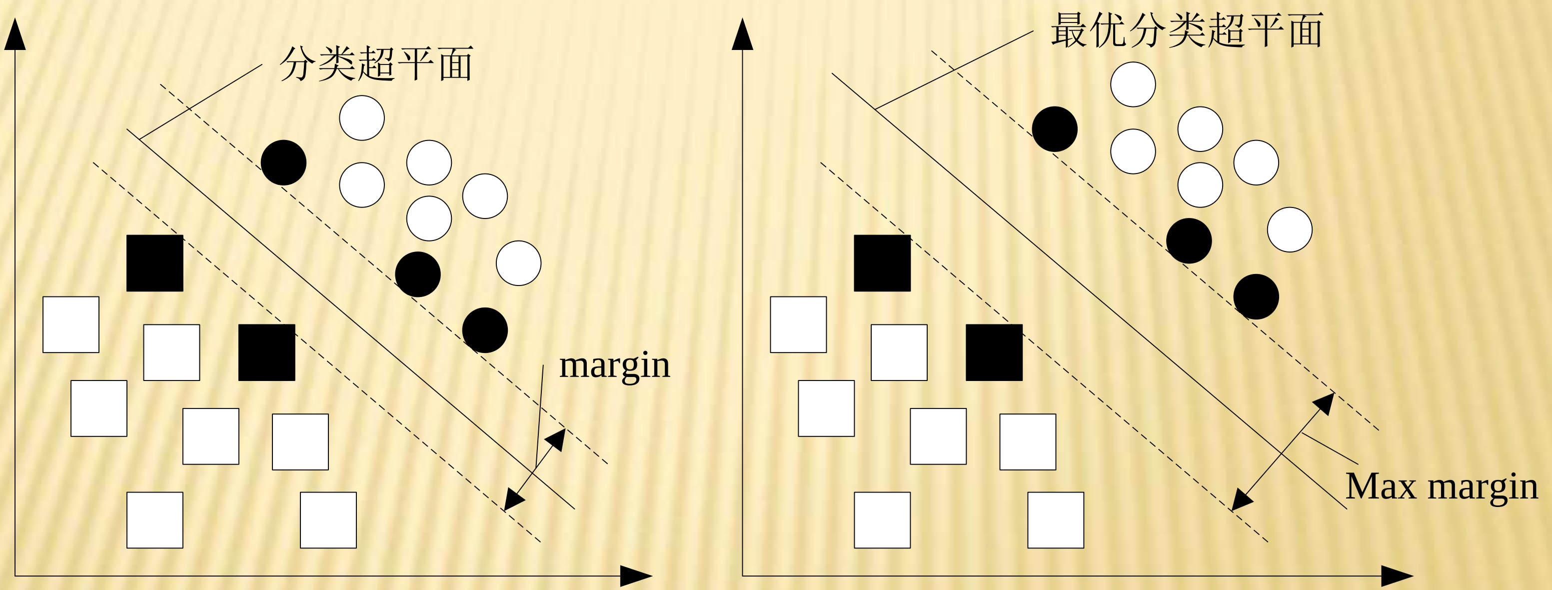
6.4.2 支持向量机

- 采用了保持经验风险值固定而最小化置信界限的策略。
- **1．线性可分数据的最优分类超平面**

$$(\boldsymbol{w} \cdot \boldsymbol{x}) - b = 0$$

- 最优分类超平面
  - 训练数据可以被无错误地划分
  - 并且每一类数据与超平面距离最近的向量距超平面之间的距离最大
- 两类数据之间最近的距离称为分类边距（Margin）
  - 对于上式分类边距等于2/||w||
  - 最优超平面就是使分类边距最大的分类超平面

最优分类面



- 在线性可分情况下，求解最优超平面，需要求解下面的二次规划问题（最小化泛函）

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

  - 约束条件为不等式

$$y_i[(\mathbf{w}\cdot\mathbf{x}_i)-b]-1\geq 0,\quad i=1,2,\ldots,N$$

- 这个优化问题的解由下面拉格朗日函数的鞍点给出：

$$L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}a_i\{y_i[(w\bullet x_i)-b]-1\}$$

  - 其中$a_i \geq 0$为拉格朗日系数。L的极值点为鞍点，L求导可得**w***和a*：

$$\frac{\partial L(w,b,a)}{\partial b}=0 \quad\Rightarrow\quad \sum_{i=1}^{N}a_i^* y_i=0$$

$$\frac{\partial L(w,b,a)}{\partial w}=0 \quad\Rightarrow\quad w^*=\sum_{i=1}^{N}y_i a_i^* x_i$$

- 此时原目标函数的对偶问题（最大化泛函）为

$$W(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_j(x_i\bullet x_j)$$

  - 其约束条件为

$$a_i\geq 0,\quad i=1,2,\ldots,N$$

$$\sum_{i=1}^{N}a_iy_i=0$$

- 这是一个不等式约束下的二次函数极值问题，且存在唯一解。根据Karush-Kuhn-Tucker (KKT)条件，这个优化问题的解必须满足：

  $a_i(y_i[(\mathbf{w} \cdot \mathbf{x}_i)-b]-1)=0$，$i=1,2,\ldots,N$

- 由于多数样本所对应的a_i将为0，这些样本对于分类超平面根本没有作用。
- 只有当a_i不为0时才对分类超平面有用，这些不为0的a_i所对应的样本就是支持向量。
- 也就是说最优分类超平面只用支持向量就决定了，即

$$w^* = \sum_{SV} y_i a_i^* x_i$$

- a*通过训练算法可显式求得。用支持向量样本又可以求得b*（阈值）：

$$b^* = \frac{1}{2}[(w^* \bullet x_{+1}^*) + (w^* \bullet x_{-1}^*)]$$

  - 其中，x^*_{+1}表示属于第一类的某个（任意一个）支持向量，x^*_{-1}表示属于另一类的任意一个支持向量。
- 最后基于最优超平面的分类规则就是下面的指示函数。

$$f(x) = \operatorname{sgn}\left((w^* \bullet x) - b^*\right) = \operatorname{sgn}\left(\sum_{SV} y_i a_i^* (x_i \bullet x) - b^*\right)$$

线性不可分数据

- 2．线性不可分数据的最优分类超平面
- 引入非负松弛变量$\xi_i \geq 0$。
- 线性约束条件转化为

$$y_i[(w \cdot x_i) - b] \geq 1 - \xi_i, \quad i=1,2,\ldots,N$$

- 二次规划问题就变成

$$\min_{w,b}\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i\right\}$$

  - 其中C被称为惩罚因子。通过改变惩罚因子可以在最大分类间隔和误分率之间进行折衷。
- 求解这个二次优化问题的方法与在可分情况下几乎相同，只是约束条件有一点小变化

$$0 \leq a_i \leq C, \quad i = 1,2,\ldots,N$$

$$\sum_{i=1}^{N} a_i y_i = 0$$

非线性数据

- 3．非线性数据的最优分类超平面
- 非线性问题，SVM通过非线性变换把非线性数据映射到另一个高维空间（特征空间）。
- 即对于线性不可分的样本$x \in R^d$，作非线性变换$\Phi: R^d \rightarrow H$，使得$\Phi(x) \in H$在特征空间H中是线性可分的。
- 下面的问题就转化成在高维空间H中求广义最优分类超平面的问题，也就是用最大边距法解决高维空间中的线性可分问题。

SVM解决思路

- 直接寻求非线性变换Φ往往很复杂，一般很难实现。
- 但是SVM巧妙地通过核函数（Kernel function）避开了这种分线性变换。用特征向量Φ(x)代替输入向量x。

$$W(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_j\left(\Phi(x_i)\bullet\Phi(x_j)\right)$$

$$f(x)=\operatorname{sgn}\left((w^*\bullet\Phi(x))-b^*\right)=\operatorname{sgn}\left(\sum_{SV}y_ia_i^*\left(\Phi(x_i)\bullet\Phi(x)\right)-b^*\right)$$

- 令K(x_i，x_j)=Φ(x_i) Φ(x_j)，K被称为核函数。
- 根据泛函有关理论，只要一种核函数K(x_i，x_j)满足Mercer条件，那么它就对应某一变换空间中的内积。

支持向量机与核函数

- 在最优分类超平面中采用适当的核函数就可以实现某一非线性变换后的线性分类，而计算复杂度却没有增加。

$$W(a)=\sum_{i=1}^{N} a_i-\frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i,x_j)$$

$$f(x)=\operatorname{sgn}\left(\sum_{SV} y_i a_i^* K(x_i,x)-b^*\right)$$

6.4.3 核函数

- 思想：
  - 将样本空间的内积替换成了核函数，而运算实际上是在样本空间中进行的，并未在特征空间中计算高维向量内积。
- 条件
  - 满足Mercer条件的函数$K(x,y)$必定是核函数，也就是肯定存在着一个映射Φ使得$K(x,y)=\Phi(x)\cdot\Phi(y)$。

MERCER条件

- 定理6.3（Mercer条件）
  - 函数K(x,y)描述了某个空间中一个内积的充分必要条件是，对于任意给定的函数g(x)，当

$$\int g^2(x)dx < \infty$$

时，有

$$\iint K(x,y)g(x)g(y)dxdy \geq 0$$

常用的核函数

- 多项式核函数（Polynomial kernel function）

$$K(x, x_i) = [(x \bullet x_i) + 1]^q, \quad q = 1,2,3,\ldots$$

- 径向基核函数（Radial basis function，RBF）

$$K(x, x_i) = \exp\left\{-\frac{\| x - x_i \|^2}{\sigma^2}\right\}$$

- Sigmoid核函数

$$K(x, x_i) = \tanh(\gamma(x \bullet x_i) + c]$$

- 并非任意的γ、c参数值都使Sigmoid函数满足Mercer条件。
- 多项式核和径向基核总是满足Mercer条件的。
- 核函数的线性组合仍然是核函数。