第1节　支持向量机基础知识

—— 统计学习理论

1、统计学习理论概述

- 有严密的数学依据，得到了严格的数学证明。它的理论基础是：
  - 概率论与数理统计
  - 泛函分析

- 传统的高维特征描述
  - 选取少量强特征，用它们的线性组合来表示。需要采用降维方法构建低维特征空间。

- 统计学习理论及SVM的高维特征描述方法
  - 映射到隐性特征空间，采用大量弱特征的线性组合逼近样本中未知的依赖关系。不关注弱特征是什么，只关心怎样将它们巧妙地组合在一起表达数据。

1.1 SVM方法与传统方法的区别

- 要较好地实现传统方法，需要人工选择(构造)少量的显性“强特征”。实际问题存在大量“弱特征”，它们巧妙的线性组合，可以更好地逼近未知的数据关系。
- “弱特征”是什么并不重要，而形成巧妙的线性组合更为重要。SVM方法则是可以构造这种巧妙的线性组合的技术和手段，而且无需进行具体隐性弱特征的提取。

SVM方法集以下模型于一身：

- 结构风险最小化（SRM）模型
- 构造复合特征的一个通用模型

  在希尔伯特空间中的内积回旋可以看作是构造特征的一种标准途径。
- 对实际数据的一种模型

  一个小的支持向量集合可能足以对不同的机器代表整个训练集。

1.2 统计学习理论中的基本概念

- 统计方法 —— 从观测自然现象或者专门安排的实验所得到的数据去推断该事务可能的规律性。

- 统计学习理论 —— 在研究小样本统计估计和预测的过程中发展起来的一种新兴理论。

- 机器学习
  - 主要研究从采集样本出发得出目前尚不能通过原理分析得到的规律，并利用这些规律对未来数据或无法观测的数据进行预测。
- 模式识别
  - 对表征事务或现象的各种形式(数值、文字及逻辑关系等)信息进行处理和分析，以对事务或现象进行描述、辨认、分类和解释的过程。
- 统计学习理论
  - 一种研究有限样本估计和预测的数学理论

1.3 统计学习理论的发展简况

- 学习过程的数学研究
  - F. Rosenblatt于1958,1962年把感知器作为一个学习机器模型
- 统计学习理论的开始
  - Novikoff(1962)证明了关于感知器的第一个定理
- Vanik和Chervonenkis(1968)提出了VC熵和VC维的概念
  - 提出了统计学习理论的核心概念
  - 得到了关于收敛速度的非渐进界的主要结论

- Vapnik和Chervonenkis(1974)提出了结构风险最小化（SRM）归纳原则。
- Vapnik和Chervonenkis(1989)发现了经验风险最小化归纳原则和最大似然方法一致性的充分必要条件,完成了对经验风险最小化归纳推理的分析。
- 90年代中期,有限样本情况下的机器学习理论研究逐渐成熟起来,形成了较完善的理论体系—统计学习理论(Statistical Learning Theory,简称SLT)

2、统计学习理论的基本内容

- 机器学习的基本问题
- 统计学习理论的核心内容

2.1 机器学习的基本问题

- 机器学习就是从给定的函数集 $f(x, \alpha)$ 中, 选择出能够最好地逼近训练器响应的分类器函数。
- 机器学习的目的可以形式化地表示为：根据n个独立同分布的观测样本$x_1, x_2, \ldots, x_n$，其对应类别标号为$y_1, y_2, \ldots, y_n$。在一组分类器$f(x, \alpha)$中求出一个最优的 $f(x, \alpha_0)$ 对训练器进行估计, 使期望风险最小：

$$R(\alpha) = \int L(y, f(x, \alpha)) dp(x, y)$$

其中 $p(x, y)$ 是未知的。

三类基本的机器学习问题

- 模式识别
- 函数逼近(回归估计)
- 概率密度估计

用有限数量信息解决问题的基本原则 —— 在解决一个给定问题时，要设法避免把解决一个更为一般的问题作为其中间步骤。

- 上述原则意味着，当解决模式识别或回归估计问题时，必须设法去“直接”寻找待求的函数，而不是首先估计密度，然后用估计的密度来构造待求的函数。

- 密度估计是统计学中的一个全能问题，即知道了密度就可以解决各种问题。一般地，估计密度是一个不适定问题(ill-posed problem)，需要大量观测才能较好地解决。

- 实际上，需要解决的问题（如决策规则估计或回归估计）是很特殊的，通常只需要有某一合理数量的观测就可以解决。

2.2 机器学习中的经验风险最小化

2.2.1 期望风险的不足

期望风险为：

$$R(\omega)=\int L(y,f(x,\omega))\,dp(x,y)$$

其中：$L(y,f(x,\omega))$ 为损失函数；

$p(x,y)$ 为样本和类别标号的联合概率。

上式的求解依赖于联合概率 $p(x,y)$，但一般只知道样本 $(x_1,y_1),\ldots,(x_n,y_n)$ 的信息，不知道 $p(x,y)$ 的函数表达形式，所以无法直接计算和最小化期望风险 $R(\omega)$。

2.2.2 经验风险

根据概率论中的大数定理，可以考虑用算术平均代替期望风险中的积分，故而用

$$Remp(\omega)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i,\omega))$$

来逼近 $R(\omega)$ 所定义的期望风向 。由于 $Remp(\omega)$ 是用已知的经验数据(样本)定义的，故称为经验风险。

2.2.3 经验风险最小化

对参数 ω 求经验风险Remp(ω)的最小值，即是所谓经验风险最小化 (empirical risk minimization) ERM原则。

$$min\ Remp(\omega) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i, \omega))$$

2.2.4 期望风险最小化与经验风险最小化

在机器学习中，用ERM原则取代期望风险最小化并没有充分的理论根据，这只是直观上认为合理的做法。

但在实际中会发现这种做法并不时时合理。一个较典型的例子是：神经网络学习的误差小，并不总是产生好的预测结果，有时反而会使网络的泛化能力下降。

研究表明：用一个很复杂的模型去拟合有限的样本，会使学习机器在泛化能力上产生损失。

产生这种情况主要有如下两个原因：

① 虽然它们都是 ω 的函数，但是“大数定理”只证明了当样本趋于无穷多时，$Remp(\omega)$在概率意义上趋近于 $R(\omega)$，并没有保证使 $Remp(\omega)$ 最小的值 ω^*与使 $R(\omega)$ 最小的值 ω' 是同一个点，更不能保证 $Remp(\omega^*)$ 能够趋近于 $R(\omega')$。

② 即使有办法在样本无穷大时，经验风险最小化和期望风险最小化有一致性，但也无法保证在样本有限时，其结果也有效。

2.3 学习机器的复杂性与推广性

2.3.1 推广性

学习机器对未来输出进行正确预测的能力称为推广性。

2.3.2 过学习及其原因

学习结果误差过小反而导致推广能力下降，则称为过学习(overfitting)。过学习的原因如下：

(1) 学习样本不充分；

(2) 学习机器设计不合理。

例2.1： 有一组训练样本 (x, y)，$x \in R, y \in [-1, 1]$。用一个函数 $f(x, \alpha) = \sin \alpha x$ 去拟合这些样本点，总可以找到一个参数α，使学习结果误差为零。如图2.1所示，但是这并不是原来的函数模型 $y = f(x)$。

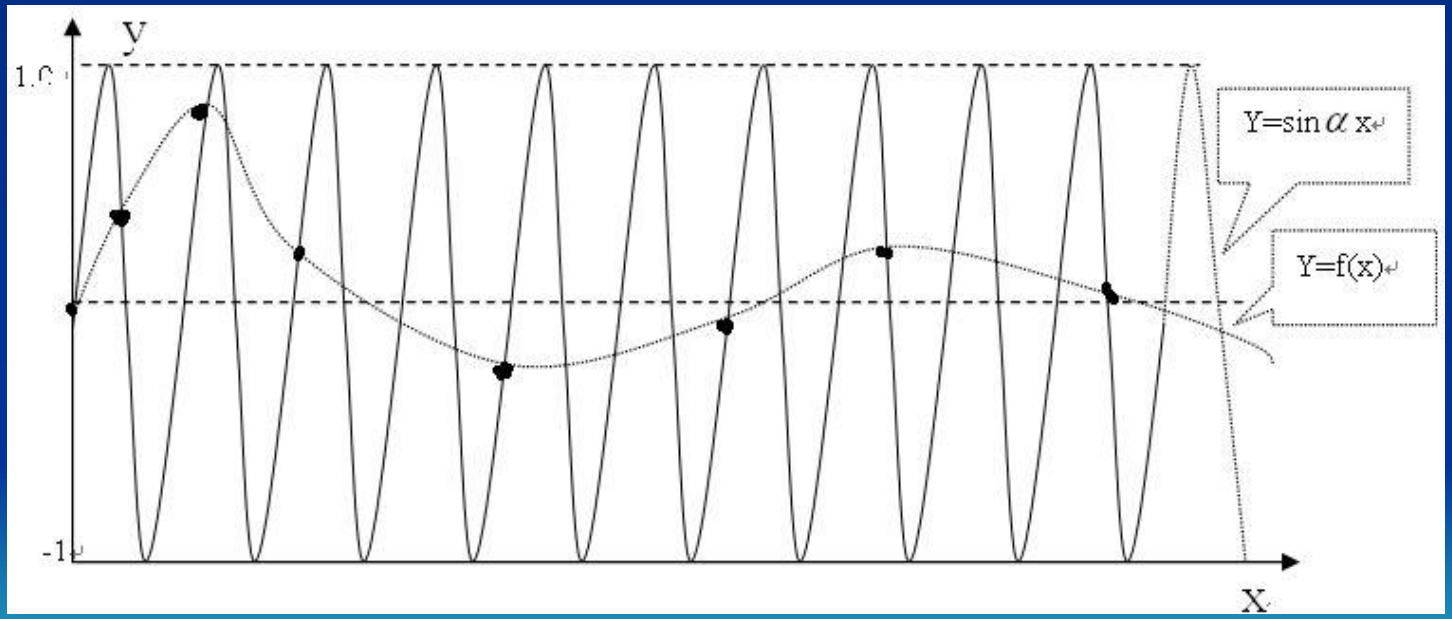

图2.1 数据的过拟合现象

出现这种现象，是因为试图中一个复杂的模型去拟合有限的样本，其推广性很差。

在实际中：对于有限的训练样本，采用复杂的预测函数去对样本进行学习，其效果通常不如相对简单的预测函数。

2.3.3 学习机器复杂性

(1) 经验风险最小不一定是期望风险最小。

(2) 学习机器的复杂性不仅和对象有关，而且和有限的学习样本相适应。

2.4 统计学习理论概述

统计学习理论是目前针对小样本进行统计估计和预测学习的最佳理论。

它包括主要四种内容：

(1) 经验风险最小化原则下学习一致性条件；

(2) 统计学习方法推广性的界；

(3) 小样本归纳推理原则；

(4) 实际应用方法和学习算法。

2.4.1 统计学习一致性条件 (Consistency)

当训练样本数目趋于无穷大时，经验风险的最优值能够收敛到真实风险的最优值 —— 称为学习过程一致性。

当下面两式成立时，称经验风险最小化学习过程与期望风险最小化学习过程是一致的：

$$\begin{cases} R(\omega'|n) \xrightarrow[n\to\infty]{} R(\omega_0) \\ Remp(\omega^*|n) \xrightarrow[n\to\infty]{} R(\omega_0) \end{cases}$$

如图2.2所示，当样本 n 趋于无穷时，经验风险和期望风险学习过程一致性的定义。

其中，$R(\omega_0)=\inf R(\omega)$ 是实际可能的最小风险.

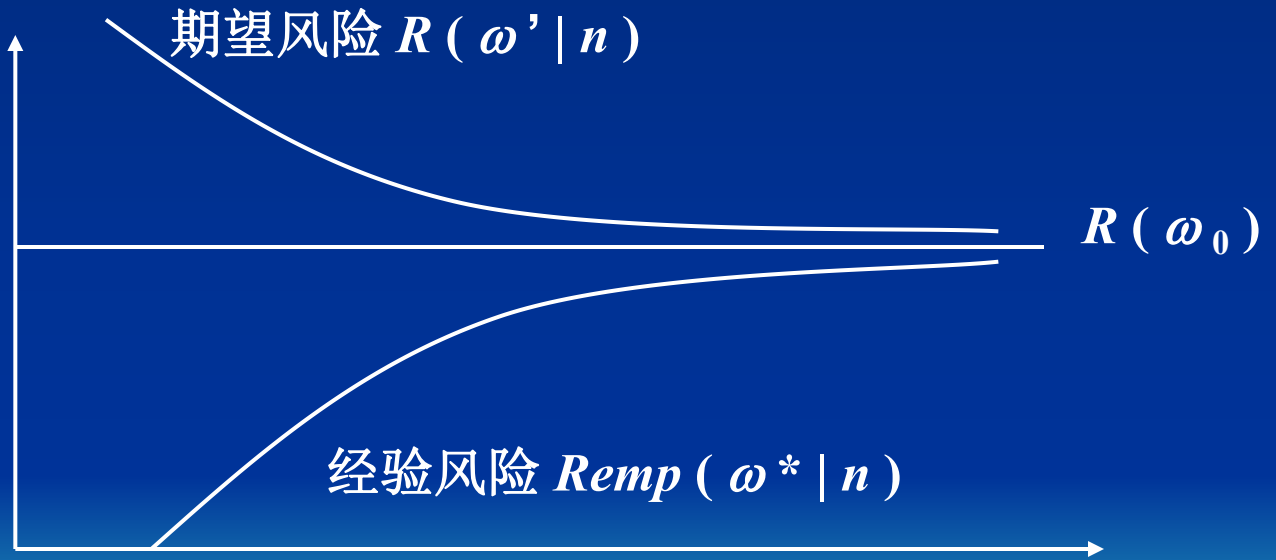

图2.2 经验风险和期望风险学习一致性

2.4.2 VC维

学习理论关键定理给出了经验风险最小化原则成立的充分必要条件；但并没有给出什么样的学习方法能满足这些条件。

为了寻求满足充分条件的学习方法，统计学习理论定义了一些指标来衡量函数集的性能，其中最重要的是VC维（Vapnik-Chervonenkiv Dimension）。

VC维的定义

- VC维：对于一个指示函数(即只有0和1两种取值的分类器) 集，如果存在 h 个样本能够被函数集里的分类器按照所有可能的 2^h 种形式分开，则称该分类器能够把 h 个样本打散，该分类器的VC维就是能够打散的样本数目 h 。

- VC维反映了函数(分类器)的学习能力。

例如：如图2.3三个样本点，使用直线分类器分类，有$2^3=8$种分法，因此直线分类器的VC 维为 3。

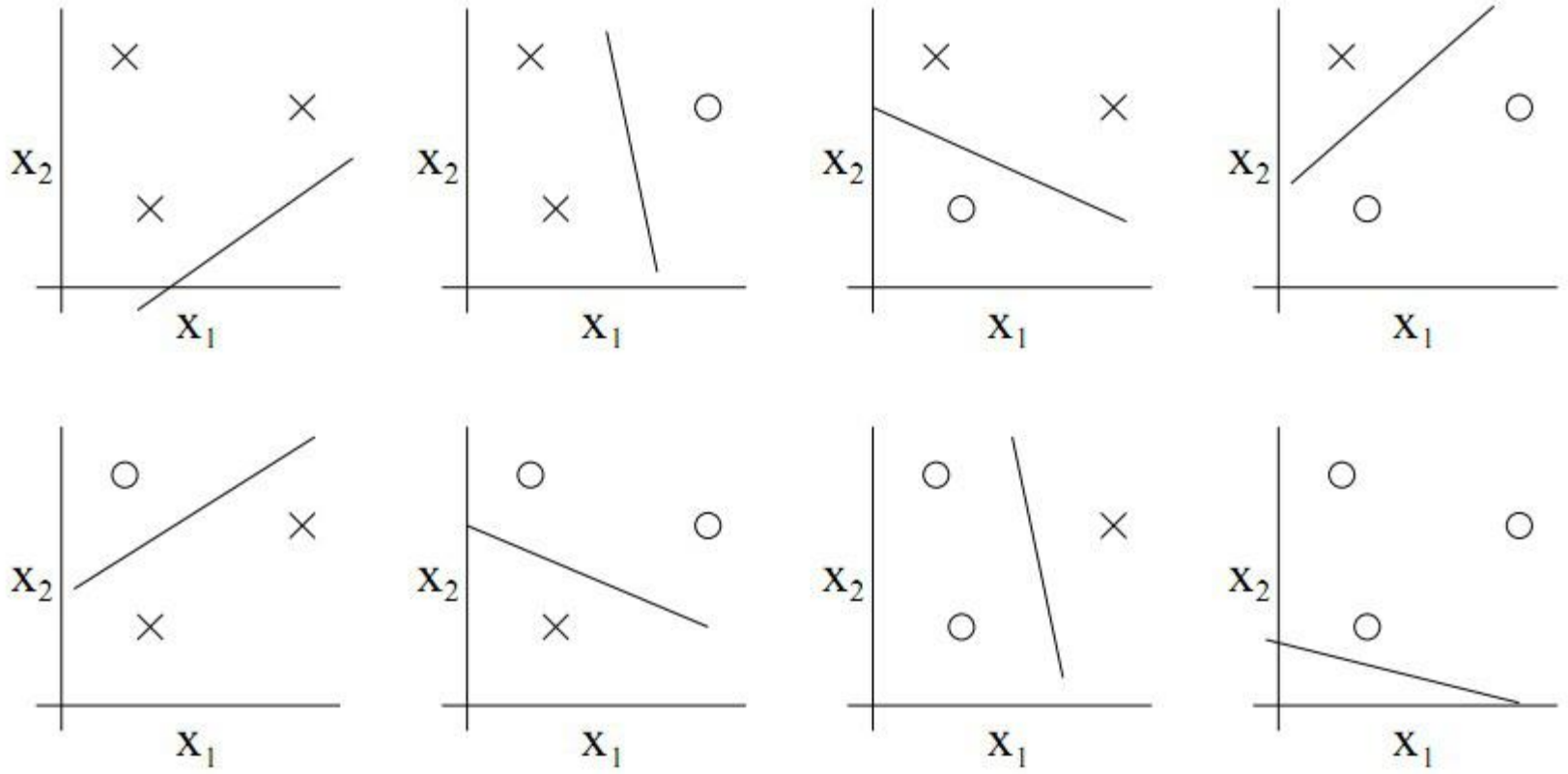

图2.3 不共线的三个样本点使用直线分类器

而对于如图2.4所示的4个样本点：

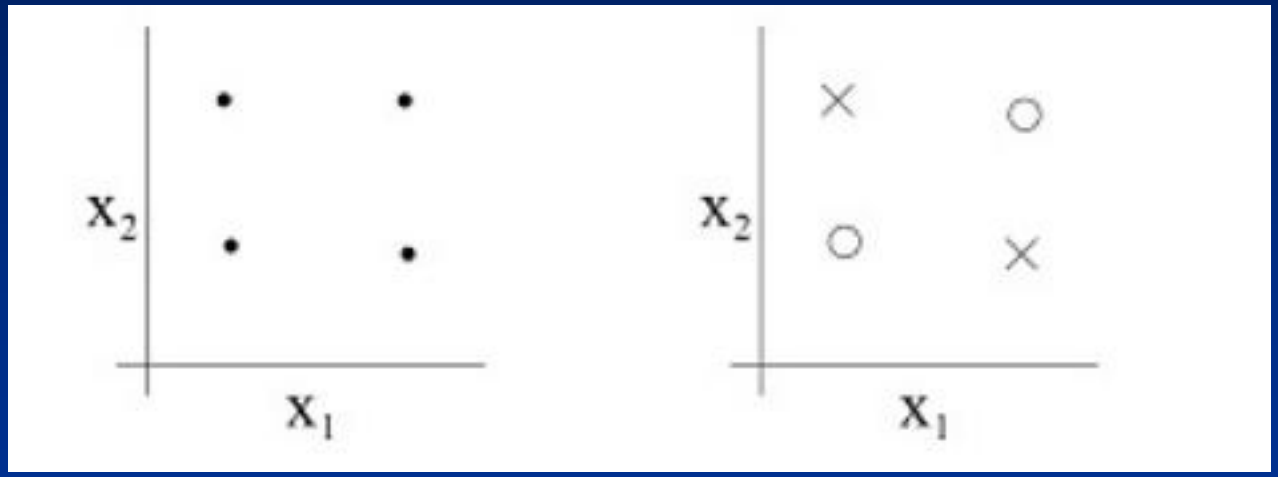

图2.4 不共线的四个样本点不能用直线分类器

- 则无论这4个点在一条直线上还是不在一条直线上(任意位置)，直线分类器都不能进行2^4=16种类型的分类，因此它不能满足4个样本点的分类要求 。

VC维的定义(续)

- 一般而言，VC维越大，学习能力就越强，但学习机器也越复杂。
- 目前还没有通用的关于计算任意分类器的VC维的理论，只有对一些特殊函数集(分类器)的VC维可以准确知道。
- N维实数空间中线性分类器和线性实函数的VC维是$N+1$。
- $\sin(x)$的VC维为无穷大。
- 计算函数集(分类器)的VC维是当前统计学习理论研究中有待解决的一个难点问题。

2.4.3 推广性的界

- 统计学习理论系统地研究了经验风险和实际风险之间的关系，也即推广性的界。

对于指示函数集中所有的函数，经验风险 $Remp(\omega)$ 和实际风险 $R(\omega)$ 之间至少以概率 $1-\eta$ 满足如下关系：

$$R(\omega) \le Remp(\omega) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}}$$

其中，h 是函数集的VC维，n是样本数。

- 学习机器的实际风险由两部分组成:
  - 训练样本的经验风险 ***Remp*** **(ω);**
  - 置信范围 **(置信水平 1-η 有关)，以及学习机器的VC维和训练样本数有关。**

$$R(\omega) \leq Remp(\omega) + \Phi\left(\frac{n}{h}\right)$$

  - 在训练样本有限的情况下，学习机器的VC维越高，则置信范围就越大，导致实际风险与经验风险之间可能的差就越大。

基于推广性的界的分类器选择

- 在设计分类器时，不但要使经验风险最小化，还要使VC维尽量小，从而缩小置信范围，使期望风险最小。
- 寻找反映学习机器的能力的更好参数，从而得到更好的界是统计学习理论今后的重要研究方向之一。

2.4.4 统计学习几个性能指标：

(1) 指示函数：分类函数 $f(x,\alpha)$，只取0, 1两个值，称其为类型标号指示函数。

(2) 指示函数集的熵：设有一个指示函数集 $f(x,\alpha)$ 和一组 n 个训练样本的样本集

$$Z_n=\{Z_i=(x_i,y_i),\ \ i=1,2,...,n\}$$

Z_n 可能有 2^n 种分类方法，假设一种分类方法就生成一种样本分类集合，则 $N(Z_n)$ 为正确分类的数目。

图2.3上不共线的三个样本点，其分类方式有$2^3=8$种，每一种分类方式都能被直线分类器正确分类，因此这3个样本点组成集合的 $N(Z_3)=8$。

而共线的3个特征点，其分类方式也有$2^3=8$种，但只有其中六种方式能够被直线分类器正确分类，因此它的 $N(Z_3)=6$，如图2.5 所示。

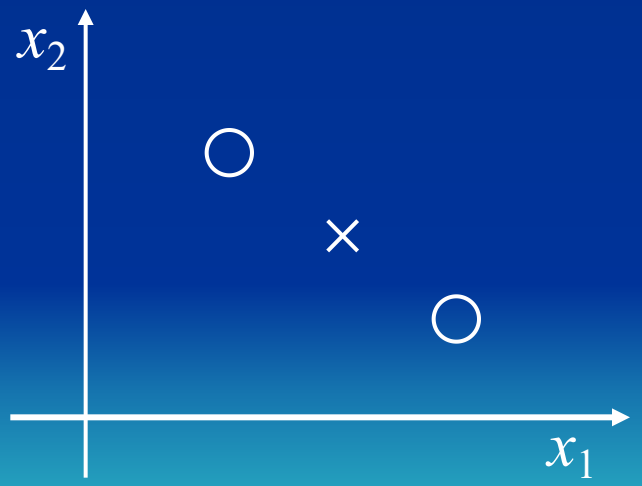

图2.5 共线的三个样本点只有六类被正确分类

(3) VC随机熵：分类器对样本集实现正确分类数目的对数，称为函数集的随机熵，记为$H(Z_n)=\ln N(Z_n)$

例如：不共线三点直线分类器的 $H(Z_3)=3\ln 2$；

共线三点直线分类器的 $H(Z_3)=\ln 6$

(4) VC熵：分类器在所有数量为n 的样本集上的随机熵的期望值，称为该分类器在样本数 n 上的熵，记为 $H(n)=E(\ln N(Z_n))$

例如：三点样本的直线分类器

$H(3)=1/2\times 3\ln 2+1/2\times \ln 6=2\ln 2+0.5\ln 3$

(5) 退火的VC熵：所有数量为n 的样本集不同的正确分类数据$N(Z_n)$ 期望值的对数称为退火熵，记为

$Hann(n) = \ln E(N(Z_n))$

例如：三个样本点直线分类器

$Hann(3) = \ln E(N(Z_3)) = \ln(1/2\times8+1/2\times6)=\ln7$

(6) 生长函数：分类器在所有数量为 n 的样本集上的最大随机熵，称为生长函数，记为 $G(n) = \ln \max_{z_n} N(Z_n)$ 。

生长函数反映了分类器把 n 个样本分成两类的最大可能的分类数，不受具体样本集的影响。

例如：三样本直线分类器的生长函数为$G(3)=3\ln2$

2.4.5 统计学习的三个问题：

① ERM方法一致性收敛的充要条件是什么？

② ERM方法收敛速度快的充要条件是什么？

③ 在什么条件下，不依赖于概率分布，ERM方法一致性收敛并收敛速度快？

学习理论的三个里程碑定理

① 双边一致收敛的充分必要条件(VC熵)：

$$\lim_{n \to \infty} \frac{H(n)}{n} = 0$$

② 收敛速度快的充分条件：$\lim_{n\to\infty} \frac{Hann(n)}{n} = 0$

③ 与样本分布概率无关的收敛充要条件：$\lim_{n\to\infty} \frac{G(n)}{n} = 0$