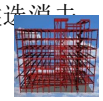
# 机器学习基础


主讲：王星
单位：中国人民大学统计学院
助教：玉冰
自助
电话：86-10-82500167
上课时间：周四18:00-20:30
上课地点：0405
Email: wangxingwisdom@126.com
办公地点：明德主楼 1019


## 第一章 数据科学与机器学习

1.机器学习概念及其应用
2.概念学习基本概念
  2.1 数据集、属性、样本空间
  2.2 训练、假设、两种学习任务
  2.3 泛化、假设空间和版本空间
3.两类概念学习算法-归纳Find-S*和候选消去
4.机器学习基本理论



## 1.1.什么是机器学习？

- 计算机中，“经验”通常是以“数据”形式储存下来，因此机器学习所研究的主要内容是关于如何从观测数据中不断学习和总结经验，从数据中产生“模型”（model）帮助计算机做出准确判断的自动化技术，这种技术称为算法，也称为“learning algorithm”.
- 两个注释：模型是算法的结果，也包括模式。

## 机器学习

- 定义：如果一个计算机程序针对某个任务T，用P作为性能的衡量标准，根据经验E自我完善，就称这个计算机程序是从经验E中学习任务T，衡量性能为P。
- (Tom M. Mitchell1997)To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E
- 例如：
  T：识别和分类图象中的手写文字。
  P：分类的正确率；
  E：已知分类的手写文字数据库。

## 什么是机器学习？

- 计算机中，“经验”通常是以“数据”形式储存下来，因此机器学习所研究的主要内容是关于如何从观测数据中**不断学习**和总结**经验**，从数据中产生“**模型**”（model）帮助计算机做出准确判断的自动化技术，这种技术称为算法，也称为“learning algorithm”.
- 两个注释：
  - 模型是算法的结果，也包括模式。
  - “经验”可以从哲学、社会、人文等多个角度进行解读。归纳起来，“经验”是人与客观事物接触过程中，通过感官获得的，关于客观事物的现象和外部联系的认practical contact with and observation of facts or events.
- 简单来说，通过观察客观现象与结果之间的联系所发现的规律就是经验。
- 机器学习的特点：起源于数据，产出模型，并能动态自我更新的算法。

## 小结

- 什么是机器学习？
- 机器学习的算法特点是什么？
- 从思维方式上，机器学习算法到底是如何学习的：
  - 归纳性思维；
  - 逻辑演绎思维；
  - 启发性思维。
- 具体算法是怎样产生的？

## 2.基本概念

**2.1 指导学习是什么？**
- 指导学习的三种观点：搜索、生成、模型
- 概念学习的几个问题

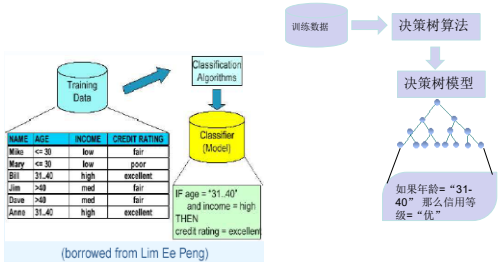
**2.2 数据集、属性、样本空间**

**2.3 训练、假设、两种学习任务**

**2.4 泛化、假设空间和版本空间**

### 2.1.1有指导的学习(Supervised Learning):

用一组输入变量 (predictors, inputs, features, independents)
对输出变量 (responses, outputs , dependent) 产生预测




(borrowed from Lim Ee Peng)

**Classification:分类问题**
**Learn a method for predicting the instance class from pre-labeled (classified) instances**

给定一个带有输入和输出的训练集，我们希望在输入和输出之间建立一个函数关系

- **例** 1（人员定位 (person detection) ）给定一个图象，我们希望回答这个人是否是我们要找的人？结果是两个可能的分类 "是" 或 "否"。
- 

在以上两个问题中，输入的都是一个高维向量 $x$, $x$, 代表某一图象块的灰度或颜色。

**涉及到算法层面会遇到两个难题：**
**1.分类会有错误；（为什么会出错，对谁会出错？如何优化？）**
**2.有些个案不存在最佳匹配（哪些数据，如何退出？）**

## 学习问题的一般表示

- $\mathcal{X}$ 输入空间（ $\mathcal{X} \subseteq \mathbb{R}^d$ ），每个元样例 $x_i = \{x_{i1}, \cdots, x_{id}\}$.
- $\mathcal{Y}$ 输出空间每个元样例 $y_i$.

  分类问题（分类的输出）： $\mathcal{Y} = \{c_1, \cdots, c_k\}$;

  回归问题（连续的输出）： $\mathcal{Y} \subseteq \mathbb{R}$
- $S = \{(x_i, y_i)\}_{i=1}^m$ ：训练样本

有监督的学习模型问题就是要计算出一个最优的函数，该函数可以恰当地描述输入和输出之间的关系。



## 指导学习的定义和类型

- **定义：指导**学习的目标是学习输入到输出的映射关系，其中正确值已部分地由指导者通过训练数据给出。
- **类型：**
  - 概念学习：0-1学习
    - 特点1：将学习问题转化为一个搜索问题；
    - 特点2：强调假设空间的性质、搜索算法和评价准则；
  - 生成式学习：
  - 模型学习（统计学习）：回归
    - 特点1：将学习问题转化为一个估计问题，特别是分布的特征估计问题；
    - 特点2：强调分布选择，估计的性质和模型的解释；



## 概念学习

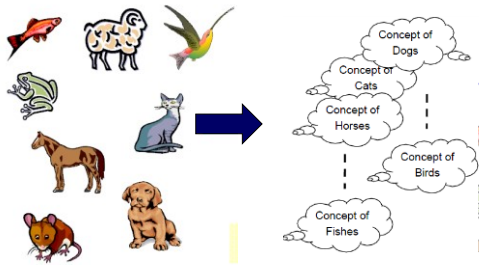
- **1.学习的基本框架：假设空间的搜索，FIND-S和候选消去算法**
- **2.概念学习的几个基本问题**
  - **样本的复杂性：学习模型收敛到成功的假设需要的样本量**
  - **计算的复杂性：学习模型收敛到成功假设需要的计算量**
  - **出错边界的估计：在成功收敛到一个假设之前，学习方法对训练数据的错误分类有多少次**
- **3.有指导学习的的几个基本概念:**
  - **假设的一致性(Consistent)**
  - **打散性(Shatter)**
  - **PAC学习估计理论**

# 学习方法的特点

- 学习的方法的是要从学习样例中学习规律推导出算法，就像孩子通过玩游戏掌握游戏取胜的技巧和方法。
- 学习方法是要从有限的样例中分离出结构，这些方法应该具有以下特点
  1. 平稳性：找到的规律适用于大部分一般情况，而不只针对少部分样例；
  2. 高效性：推导解的时间不应随着数据量的增大而以指数增长；
  3. 稳健性：找到的规律不应对少部分样例或个别样例的变化过于敏感

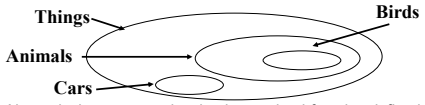
13

# 2.1.2概念是什么?




# 概念是什么

- A Concept is a subset of objects or events defined over a larger set [Example: The concept of a ***bird*** is the subset of all objects (i.e., the set of all ***things*** or all ***animals***) that belong to the category of bird.]概念是一组研究对象或事件的集合。它是较大集合中选取的子集，或在较大集合中定义的布尔函数



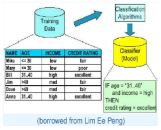

- Alternatively, a concept is a boolean-valued function defined over this larger set [Example: a function defined over all ***animals*** whose value is ***true*** for birds and ***false*** for every other animal].

Lecture Notes for Xing Wang 2008 Introduction to Machine Learning Stat Ruc ?
(参考Tom Mitchell machine learning(chap1,2) 1997)

15

# 概念学习的本质
# Concept-Learning?

Given a set of examples labeled as members or non-members of a concept, ***concept-learning*** consists of automatically inferring the general definition of this concept.
从输入输出样例中推断布尔函数




In other words, ***concept-learning*** consists of approximating a boolean-valued function from training examples of its input and output.

# 学习方法的特点

- 学习的方法的是要从学习样例中学习规律推导出算法，就像孩子通过玩游戏掌握游戏取胜的技巧和方法。
- 学习方法是要从有限的样例中分离出结构，这些方法应该具有以下特点
  1. 平稳性：找到的规律适用于大部分一般情况，而不只针对少部分样例；
  2. 高效性：推导解的时间不应随着数据量的增大而以指数增长；
  3. 稳健性：找到的规律不应对少部分样例或个别样例的变化过于敏感

17

# 2.2 数据集、属性到样本空间

| 编号 | 色泽 | 根蒂 | 敲声 | 好瓜 |
|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 是 |
| 2 | 乌黑 | 蜷缩 | 清脆 | 是 |
| 3 | 青绿 | 硬挺 | 沉闷 | 否 |
| 4 | 乌黑 | 稍蜷 | 沉闷 | 否 |

- 以样本的属性为坐标轴张成的多维空间，也叫属性空间、输入空间。
- 上例中，每行样本包含三个属性：色泽、根蒂、敲声，则可以以这三个属性为坐标轴，生成一个三维空间，每个西瓜（只要用这三种属性描述）都能在该空间中找到其对应的坐标位置。

## 概念的要素构成学习例1

**Database:**

| Day | Sky | AirTemp | Humidity | Wind | Water | Forecast | WaterSport |
|---|---|---|---|---|---|---|---|
| 1 | Sunny | Warm | Normal | Strong | Warm | Same | Yes |
| 2 | Sunny | Warm | High | Strong | Warm | Same | Yes |
| 3 | Rainy | Cold | High | Strong | Warm | Change | No |
| 4 | Sunny | Warm | High | Strong | Cool | Change | Yes |

**class**

**假设的形式:**

**合取Conjunction of constraints on each attribute** where:

- "?" means "any value is acceptable"
- "0" means "no value is acceptable"

**Example of a hypothesis: <?,Cold,High,?,?,?>**

(If the air temperature is cold and the humidity high then it is a good day for water sports)

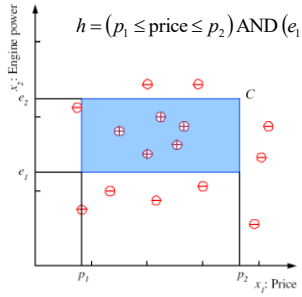
19

表 1.1 西瓜数据集

| 编号 | 色泽 | 根蒂 | 敲声 | 好瓜 |
|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 是 |
| 2 | 乌黑 | 蜷缩 | 浊响 | 是 |
| 3 | 青绿 | 硬挺 | 清脆 | 否 |
| 4 | 乌黑 | 稍蜷 | 沉闷 | 否 |

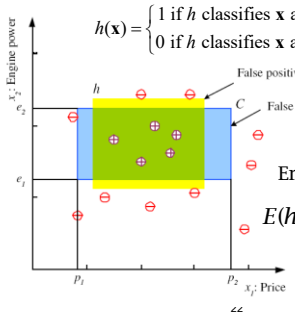
## 概念学习连续的问题：真实的概念Class $C$

$$h = (p_1 \le \text{price} \le p_2) \text{ AND } (e_1 \le \text{engine power} \le e_2)$$




## 假设空间 Hypothesis class $\mathcal{H}$

$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ classifies } \mathbf{x} \text{ as positive} \\ 0 \text{ if } h \text{ classifies } \mathbf{x} \text{ as negative} \end{cases}$$




Error of h on $\mathcal{H}$

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

## 一般意义上的概念学习

Concept: "days on which my friend Tom enjoys his favourite water sports"

Task: predict the value of "Enjoy Sport" for an arbitrary day based on the values of the other attributes

attributes

| Sky | Temp | Humid | Wind | Water | Fore-cast | Enjoy Sport |
|---|---|---|---|---|---|---|
| Sunny | Warm | Norr | | Warm | Same | Yes |
| Sunny | Warm | High | example | Warm | Same | Yes |
| Rainy | Cold | High | Strong | Warm | Change | No |
| Sunny | Warm | High | Strong | Cool | Change | Yes |

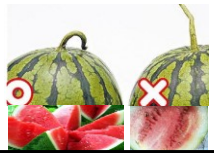
23

## 2.2.样本空间、训练集、泛化、假设空间和版本空间

- 训练集属性色泽、根蒂、敲声分别有3、2、3种可能取值
- 假设空间：，采用属性合取式描述假设空间假设空间由形如"(色泽=？)∧(根蒂=？)∧(敲声=？)"的所有假设组成。
- 泛化：通过对训练集中"好瓜"的经验归纳出对没有见过的瓜进行判断的能力。
- 参考周志华《机器学习》，2006

表 1.1 西瓜数据集

| 编号 | 色泽 | 根蒂 | 敲声 | 好瓜 |
|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 是 |
| 2 | 乌黑 | 蜷缩 | 浊响 | 是 |
| 3 | 青绿 | 硬挺 | 清脆 | 否 |
| 4 | 乌黑 | 稍蜷 | 沉闷 | 否 |



# 作为搜索的概念学习要素：实例数、概念数、假设数

- Sky: Sunny, Cloudy, Rainy
- AirTemp: Warm, Cold
- Humidity: Normal, High
- Wind: Strong, Weak
- Water: Warm, Cold
- Forecast: Same, Change

#distinct instances : 3*2*2*2*2*2 = 96

#distinct concepts : $2^{96}$

#syntactically distinct hypotheses : 5*4*4*4*4*4=5120

#semantically distinct hypotheses : 1+4*3*3*3*3*3=973



# 假设的一般到特殊序：偏序
# 很多假设空间的假设存在序结构

- Consider two hypotheses:
  - $h_1$=< Sunny,?,?,Strong,?,?>
  - $h_2$=< Sunny,?,?,?,?,?>
- Set of instances covered by $h_1$ and $h_2$:

  $h_2$ imposes fewer constraints than $h_1$ and therefore classifies more instances x as positive h(x)=1. $h_2$ is a more general concept.

Definition: Let $h_j$ and $h_k$ be boolean-valued functions defined over X. Then $h_j$ is **more general than or equal to** $h_k$ (written $h_j \geq h_k$) if and only if

$$\forall x \in X : [(h_k(x) = 1) \rightarrow (h_j(x) = 1)]$$

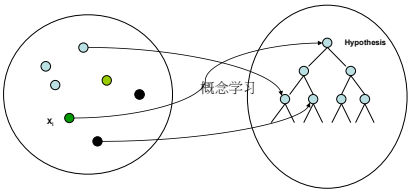
- The relation ≥ imposes a partial order over the hypothesis space H that is utilized in many concept learning methods.



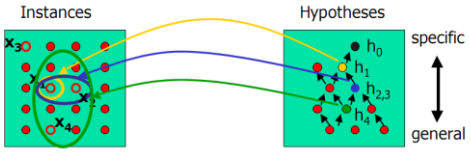
# Learning

学习问题经常归结为搜索问题，即对一个假设空间进行搜索，以确定最佳拟和观察到的数据和学习器中已有的假设。

样本空间（实例空间）X　　假设空间 H




# 举例：




$x_1$=<Sunny,Warm,Normal,Strong,Warm,Same>+　$h_0$=< Ø, Ø, Ø, Ø, Ø, Ø,>

$h_1$=< Sunny,Warm,Normal, Strong,Warm,Same>

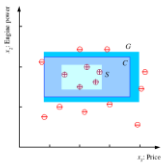
$x_2$=<Sunny,Warm,High,Strong,Warm,Same>+

$x_3$=<Rainy,Cold,High,Strong,Warm,Change> -　$h_{2,3}$=< Sunny,Warm,?, Strong,Warm,Same>

$x_4$=<Sunny,Warm,High,Strong,Cool,Change> +　$h_4$=< Sunny,Warm,?, Strong,?,?>



# 最特殊的假设，最一般的假设




涵盖所有正例不包括任何负例的最小的假设
<Ø, Ø, Ø, Ø,..., Ø >

most general hypothesis, G
<?,?,?,?,...,?>涵盖所有正例不包括任何负例的最大的假设

h ∈ $\mathcal{H}$, between S and G is consistent

and make up the version space
(Mitchell, 1997)

**一致性**定义:一个假设h和训练样本集称为一致,当且仅当对D中每个样例(x,C(x)),有C(x)=h(x)

$$\text{consistent}(h, D) = \{\forall(x, c(x)) \in D, h(x) = C(x)\}$$

# 版本空间基本概念

- Any $h \in H$ between $S$ and $G$
- Consisting of valid hypotheses with no error (consistent with the training set)

1.**版本空间**定义: 假设空间H和训练数据集D的变型空间是H中每个与训练样本D一致的假设构成的子集

$$VS_{H,D} = \{h \in H, Consistent(h, D)\}$$

2.关于假设空间H和训练数据集D的**一般边界**(General Boundary)

$$G = \{g \in H, Consistent(g, D) \wedge [\neg\exists((g' > g) \wedge Consistent(g', D))]\}$$

3.关于假设空间H和训练数据集D的**特殊边界**(Specific Boundary)

$$S = \{s \in H, Consistent(s, D) \wedge [\neg\exists((s > s') \wedge Consistent(s', D))]\}$$

# 归纳学习（Inductive Leanring）

- 机器学习的目标是从假设空间H中找到h(x):X →{0,1}，使得h(x)=c(x).
- **归纳学习的基本假设：**任何一个假设如果在足够大的训练样例中很好的逼近目标函数，那么也可能在未见的实例中更好地逼近目标函数。

Given:

- Instance Space X : Possible days decribed by the attributes Sky, Temp, Humidity, Wind, Water, Forecast
- Target function c: EnjoySport X → {0,1}
- Hypothese Space H: conjunction of literals e.g.
  < Sunny ? ? Strong ? Same >
- Training examples D : positive and negative examples of the target function: $<x_1,c(x_1)>,\ldots, <x_n,c(x_n)>$

Determine:

- A hypothesis h in H such that h(x)=c(x) for all x in D.



# 作为搜索的概念学习要素：实例数、概念数、假设数

- Sky: Sunny, Cloudy, Rainy
- AirTemp: Warm, Cold
- Humidity: Normal, High
- Wind: Strong, Weak
- Water: Warm, Cold
- Forecast: Same, Change

#distinct instances : 3*2*2*2*2*2 = 96

#distinct concepts : $2^{96}$

#syntactically distinct hypotheses : 5*4*4*4*4*4=5120

#semantically distinct hypotheses : 1+4*3*3*3*3*3=973



# 假设的一般到特殊序：偏序 很多假设空间的假设存在序结构

- Consider two hypotheses:
  - $h_1$=< Sunny,?,?,Strong,?,?>
  - $h_2$=< Sunny,?,?,?,?,?>
- Set of instances covered by $h_1$ and $h_2$:

$h_2$ imposes fewer constraints than $h_1$ and therefore classifies more instances x as positive h(x)=1. $h_2$ is a more general concept.

Definition: Let $h_j$ and $h_k$ be boolean-valued functions defined over X. Then $h_j$ is **more general than or equal to** $h_k$ (written $h_j \geq h_k$) if and only if

$$\forall x \in X : [(h_k(x) = 1) \rightarrow (h_j(x) = 1)]$$

- The relation ≥ imposes a partial order over the hypothesis space H that is utilized in many concept learning methods.



# Concept Learning

Experience

| Example | Sky | AirTemp | Humidity | Wind | Water | Forecast | EnjoySport |
|---|---|---|---|---|---|---|---|
| 1 | Sunny | Warm | Normal | Strong | Warm | Same | Yes |
| 2 | Sunny | Warm | High | Strong | Warm | Same | Yes |
| 3 | Rainy | Cold | High | Strong | Warm | Change | No |
| 4 | Sunny | Warm | High | Strong | Cool | Change | Yes |

Low Weak

Prediction

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | Sunny | Cold | Normal | Strong | Warm | Same | ? |
| 6 | Rainy | Warm | High | Strong | Warm | Change | ? |



# Concept Learning

- Learning problem:
  - Task T: classifying days on which my friend enjoys water sport
  - Performance measure P: percent of days correctly classified
  - Training experience E: days with given attributes and classifications



# Concept Learning Review

- Learning problem:
  - Concept: a subset of the set of instances X
    c: X → {0, 1}
  - Target function:
    Sky × AirTemp × Humidity × Wind × Water × Forecast → {Yes, No}
  - Hypothesis:
    Characteristics of all instances of the concept to be learned = Constraints on instance attributes
    h: X → {0, 1}

# Candidate-Elimination Algorithm

$x_1$=<Sunny,Warm,Normal,Strong,Warm,Same>+

$x_2$=<Sunny,Warm,High,Strong,Warm,Same>+

$x_3$=<Rainy,Cold,High,Strong,Warm,Change> -

$x_4$=<Sunny,Warm,High,Strong,Cool,Change> +

$S_0$ = {<∅, ∅, ∅, ∅, ∅, ∅>}
$G_0$ = {<?, ?, ?, ?, ?, ?>}

$S_1$ = {<Sunny, Warm, Normal, Strong, Warm, Same>}
$G_1$ = {<?, ?, ?, ?, ?, ?>}

$S_2$ = {<Sunny, Warm, ?, Strong, Warm, Same>}
$G_2$ = {<?, ?, ?, ?, ?, ?>}

$S_3$ = {<Sunny, Warm, ?, Strong, Warm, Same>}
$G_3$ = {<Sunny, ?, ?, ?, ?, ?>, <?, Warm, ?, ?, ?, ?>, <?, ?, ?, ?, ?, Same>}



# Candidate-Elimination Algorithm

| 4 | Sunny | Warm | High | Strong | Cool | Change | Yes |
|---|---|---|---|---|---|---|---|

$S_3$ = {<Sunny, Warm, ?, Strong, Warm, Same>}

S4 = {<Sunny, Warm, ?, Strong, ?, ?>}

<Sunny, ?, ?, Strong, ?, ?>  <Sunny, Warm, ?, ?, ?, ?>  <?, Warm, ?, Strong, ?, ?>

G4 = {<Sunny, ?, ?, ?, ?, ?>, <?, Warm, ?, ?, ?, ?>}

| 5 | Sunny | Warm | Normal | Light | Warm | Same | Yes/No |
|---|---|---|---|---|---|---|---|

$G_3$ = {<Sunny, ?, ?, ?, ?, ?>, <?, Warm, ?, ?, ?, ?>, <?, ?, ?, ?, ?, Same>}



# 概念学习的特点总结

1. 从一个离散的问题触发，学习的过程就是用训练集的样例来去逼近一个布尔函数。也就是说，概念学习是从离散训练样例的输入输出中推断出满足条件的布尔函数的过程。
2. S和G不是唯一的，依赖于训练集和假设类，可能存在多个解空间，分别形成S-集和G-集，其中S-集中的任何一个假设都与所有的训练结果一致，不存在比S更特殊的一致假设；G-集中任何一个假设都与所有的训练结果一致，不存在比G集更一般的一致假设。
3. 搜索假设空间的方法依赖于一种对任意概念学习都有效的结构：假设的一般到特殊序关系。利用假设空间的自然结构，可以在无限的假设空间中进行彻底的搜索，而不需要明确列举所有的假设。



# 例题:在西瓜问题中，如何根据训练集求所对应的版本空间

- ①写出假设空间：先列出所有可能的样本点（即特征向量）（即每个属性都取到所有的属性值）
- ②对应着给出已知数据集，将与正样本不一致的、与负样本一致的假设删除。
- 即可得出与训练集一致的假设集合，也就是版本空间了。

# Candidate-Elimination Algorithm

$S_0$ = {<∅, ∅, ∅>}
$G_0$ = {<?, ?, ?>}

$S_1$ = {<青绿, 蜷缩,浊响 >}
$G_1$ = {<?, ?, ? >}

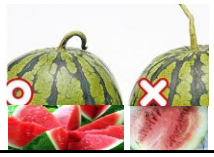
$S_2$ = {<?, 蜷缩,浊响>}
$G_2$ = {<?, ?, ? >}

$S_3$ = {<?, 蜷缩,浊响>}
$G_3$ = { <?,蜷缩, ? >, <?, ?,浊响>}

$S_4$ = {<?, 蜷缩,浊响>}
$G_4$ = { <?,蜷缩, ? >, <?, ?,浊响>}

表 1.1 西瓜数据集

| 编号 | 色泽 | 根蒂 | 敲声 | 好瓜 |
|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 是 |
| 2 | 乌黑 | 蜷缩 | 浊响 | 是 |
| 3 | 青绿 | 硬挺 | 清脆 | 否 |
| 4 | 乌黑 | 稍蜷 | 沉闷 | 否 |



# 使用版本空间的候选消除算法

- 将G中初始化为H中极大一般假设
- 将S中初始化为H中极小特殊假设
- 对每个样例d,进行以下操作:

**如果d是正例**

（1）从G中移出所有与d不一致的假设 1

（2）对S中每个与d不一致的假设s

从S中移除s 2

把s中所有极小泛化h加入到S中，其中h满足与 d一致，而且G的每个成员比h更一般 3

（3）从S中移除所有的假设：它比S中另一假设更一般 4

**如果d是负例**

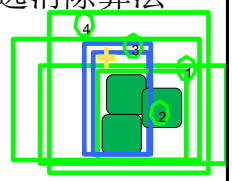
（1）从S中移出所有与d不一致的假设

（2）对G中每个与d不一致的假设g

从G中移除g

把g中所有极小特殊h加入到G中，其中h满足与 d一致，而且S的每个成员比h更特殊

（3） 从G中移除所有的假设：它比G中另一假设更特殊



## Candidate-EliminationAlgorithm(收敛性)

- 候选消去算法的特点：寻找与训练样例一致的假设；
- 原理：通过S泛化和G的特殊化不断缩小变型空间，实现对一致假设的搜索。
- 版本空间虽然并不保证预测性能优良，但在以下条件下可以证明它是收敛的The version space will converge toward the correct target concepts if:
  - contains the correct target concept
    （H中包含了描述目标概念的正确假设(**可知学习**)
  - There are no errors in the training examples
    在训练样本中没有错误（完全学习）
- A training instance to be requested next should discriminate among the alternative hypotheses in the current version space。理想的训练样例是对S和G都有作用，于是可以使边界单调移动，从而有效地推动搜索进程。
- Partially learned concept can be used to classify new instances using the majority rule.不完全学习仍然可以用于预测

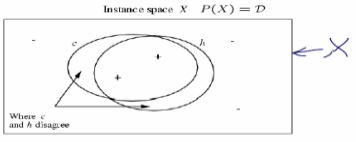


## 机器学习关心的基本问题chap7
(Computational Learning Theory)

- 计算学习理论：
  - **可能的学习**：研究什么时候一个问题是可被学习的?什么样的条件下成功的学习是可能的?假设空间的复杂度（计算复杂度computational complexity）
  - **有效的学习**：一个问题怎样才能被有效解决?
  - 学习算法收敛到正确假设（较高概率）所需要的基本的计算量
  - 训练样本数目与学习能力的关系
  - 需要多少训练样本来保证学习的成功（样本复杂度sample complexity）
  - 训练样本以什么样的方式来指导学习



**定义：**假设h关于目标概念C和分布D的真实错误率（True Error）为：h根据从D随机抽取的实例错误分类的概率




**Definition:** The **true error** (denoted $error_{\mathcal{D}}(h)$) of hypothesis $h$ with respect to target concept $c$ and distribution $\mathcal{D}$ is the probability that $h$ will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

Lecture 4 otes for Xing Wang 2008 Introduction to Machine Learning Stat Ruc

## 评估假设

**Two Notions of Error**

*Training error* of hypothesis $h$ with respect to target concept $c$

- How often $h(x) \neq c(x)$ over training instances D

$$error_{\mathrm{D}}(h) = \Pr_{x \in \mathrm{D}}[c(x) \neq h(x)] = \frac{\sum_{x \in \mathrm{D}} \delta(c(x) \neq h(x))}{|\mathrm{D}|}$$

Set of training examples

*True error* of hypothesis $h$ with respect to $c$

- How often $h(x) \neq c(x)$ over future instances drawn at random from $\mathcal{D}$

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

Probability distribution P(x)

## PAC可学习：可能近似的正确学习
(Probably Approximately Correct, PAC学习)

- 只要求学习算法输出错误率限定在某常数**ε**范围内的假设；
- 要求对所有的随机抽取样例序列的失败的概率限定在某常数δ范围内；
- PAC可学习的定义：
  考虑定义在长度为*n*的实例集合*X*上的一概念类别*C*，学习算法*L*使用假设空间*H*。当对所有*c*∈*C*，*X*上的分布*D*，*ε*和δ满足0<*ε*, δ<1/2，学习算法*L*将以至少1-δ的概率输出一假设*h*∈*H*，使error_*D*(*h*)≤ε，这时称*C*是使用*H*的*L*可PAC学习的，所使用的时间为1/*ε*，1/δ，*n*以及size(*c*)的多项式函数：
  - 1/*ε*和1/δ表示了对输出假设要求的强度;
  - *n*和size(*c*)表示了实例空间*X*和概念类别*C*中固有的复杂度，*n*为*X*中实例的数量，size(*c*)为概念*c*的编码长度
- PAC可学习性的一个隐含的条件：对*C*中每个目标概念*c*，假设空间*H*都包含一个以任意小误差接近*c*的假设。



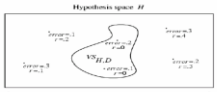
## 训练样例数和样本复杂性

- 在实践中，通常更关心所需的训练样例数，如果*L*对每个训练样例需要某最小处理时间，那么为了使*c*是*L*可*PAC*学习的，*L*必须从多项式数量的训练样例中进行学习。
- 实际上，为了显示某目标概念类别*C*是可*PAC*学习的，一个典型的方法是证明*C*中每个目标概念可以从多项式数量的训练样例中学习到，且处理每个样例的时间也是多项式级。
- 样本复杂度: **(Sample complexity)**
  随着问题规模的增长所带来的所需训练样例的增长称为该学习问题的样本复杂度。成功学习所需要的样本量。

# 1.一致学习算法的样本复杂度

- 定义：**ε-详尽(exhausted)**

  考虑一假设空间*H*，目标概念*c*，实例分布*D*以及*c*的一组训练样例*S*。当$VS_{H,D}$中每个假设*h*关于*c*和*D*错误率小于*ε*时，变型空间被称为*c*和*D*是ε-详尽的。




($r$ = training error, $error$ = true error)

**Definition:** The version space $VS_{H,D}$ is said to be $\epsilon$-**exhausted** with respect to $c$ and $\mathcal{D}$, if every hypothesis $h$ in $VS_{H,D}$ has true error less than $\epsilon$ with respect to $c$ and $\mathcal{D}$.

$$(\forall h \in VS_{H,D})\ error_{\mathcal{D}}(h) < \epsilon$$

# 变型空间的ε-详尽化定理

**Theorem:** [Haussler, 1988].

If the hypothesis space $H$ is finite, and $D$ is a sequence of $m \geq 1$ independent random examples of some target concept $c$, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to $H$ and $D$ is not $\epsilon$-exhausted (with respect to $c$) is less than

$$|H|e^{-\epsilon m}$$

- 若假设空间*H*有限，且*D*为目标概念*c*的一系列*m*>=1个独立随机抽取的样例，那么对于任意0=<*ε*<=1，变型空间$VS_{H,D}$不是*ε*-详尽的概率小于或等于 $|H|e^{-\epsilon m}$



**Learning Conjunctions of Boolean Literals**

How many examples are sufficient to assure with probability at least $(1-\delta)$ that

every $h$ in $VS_{H,D}$ satisfies $error_{\mathcal{D}}(h) \leq \epsilon$

Use our theorem:

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Suppose $H$ contains conjunctions of constraints on up to $n$ boolean attributes (i.e., $n$ boolean literals). Then $|H| = 3^n$, and

$$m \geq \frac{1}{\epsilon}(\ln 3^n + \ln(1/\delta))$$

or

$$m \geq \frac{1}{\epsilon}(n \ln 3 + \ln(1/\delta))$$

- 物理意义：训练样例的数目*m*足以保证任意一致假设是可能（可能性为1-*δ*）近似（错误率为*ε*）正确的;
- *m*随着1/*ε*线性增长，随着1/*δ*和假设空间的规模对数增长



# 例题：布尔合取是PAC可学习的，学习样本理论值：

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

If $H$ is as given in *EnjoySport* then $|H| = 973$, and

$$m \geq \frac{1}{\epsilon}(\ln 973 + \ln(1/\delta))$$

... if want to assure that with probability 95%, $VS$ contains only hypotheses with $error_{\mathcal{D}}(h) \leq .1$, then it is sufficient to have $m$ examples, where

$$m \geq \frac{1}{.1}(\ln 973 + \ln(1/.05))$$
$$m \geq 10(\ln 973 + \ln 20)$$
$$m \geq 10(6.88 + 3.00)$$
$$m \geq 98.8$$

# 2.不可知学习

- **不可知学习算法**

  学习算法不假定目标概念可在*H*中表示，而只简单地寻找具有最小训练错误率的假设，这样的学习算法称为**不可知学习算法**
- **不一致假设**：有非零训练错误率的假设

  令*S*代表学习算法可观察到的特定训练样例集合，*error*S(*h*)表示*h*的训练错误率，即*S*中被*h*误分类的训练样例所占比例

  令*hbest*表示*H*中有最小训练错误率的假设。

**问题是：**

多少训练样例才足以保证其真实错误率***error_D*(*hbest*)**不会多于***ε*+*error*S(*hbest*)**?

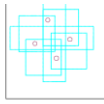


# 3无限假设空间的样本复杂度
# VC Dimension（假设类H的学习能力）

- VC（ Vapnik-Chervonenkis）维是考虑*H*复杂度的一种度量
- 作用和意义：使用VC维代替|H|也可以得到样本复杂度的边界，基于VC维的样本复杂度比|H|更紧凑，还可以刻画无限假设空间的样本复杂度
- *N* points can be labeled in
- $\mathcal{H}$ shatters *N* if there exists $h \in \mathcal{H}$ consistent for any of these: VC($\mathcal{H}$) = *N*

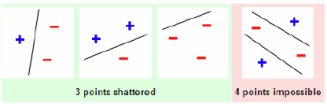
  - 定义：VC维
    - 定义在实例空间X上的假设空间H的Vapnik-Chervonenkis维，是可被H打散的X的最大可能的有限子集的大小
  - 如果X的任意有限大的子集可被H打散，则VC(H)=∞
  - 直观上，被打散的X的子集越大，H的表示能力越强。
  - 对于任意有限的H，VC(H)<=$\log_2$|H|

An axis-aligned rectangle shatters 4 points only !



# VC维的例子

- 例子1：实例集合X为二维实平面上的点（x,y），假设空间H为所有线性决策线。此时，H的VC维是多少？
  - 除了三个点在同一直线上的特殊情况，X中3个点构成的子集的任意划分（）均可被线性决策线打散
  - X中4个点构成的子集，无法被H中的任一h打散（分类）
  - 所以，VC(H)=3






# VC维的例子：

- 例子2：假定实例空间X为实数集合，而且H为实数轴上的区间的集合，问VC(H)是多少？

  只要找到能被H打散的X的最大子集，首先包含2个实例的集合能够被H打散，其次包含3个实例的集合不能被H打散，因此VC(H)=2

- 例子3：在r维空间中，线性决策面H的VC维为r+1



# VC维在变型空间中的应用

- VC维表征的ε-详尽化定理：

可以证明，要以1- δ 的概率学习到ε-详尽变型空间（PAC学习任意目标概念），需要的训练样本的边界为：

$$m \geq \frac{1}{\varepsilon}(4\log_2(2/\delta)+8VC(H)\log_2(13/\varepsilon))$$

（Blumer et al. 1989)

- 定理表明：在样本多到什么情况下，可以保证成功学习！
- 可以看到：要成功进行PAC学习，所需要的训练样本数正比于1/ δ的对数，正比于VC(H)，正比于1/ε的对数。



# 样本量的下界

- 样本复杂度的下界定理（Ehrenfeucht et al. 1989）
  - 考虑任意概念类C，且VC(C)>=2，任意学习器L，以及任意0<ε<1/8，0<δ<1/100。存在一个分布D以及C中一个目标概念，当L观察到的样例数目小于下式时：

$$\max\left[\frac{1}{\varepsilon}\log(1/\delta), \frac{VC(C)-1}{32\varepsilon}\right]$$

  L将以至少δ的概率输出一假设h，使$error_D(h)>\varepsilon$
- 定理表明：在样本少到什么情况下，学习器不可能进行成功的PAC学习.
- 该定理提供了成功PAC学习所必要的训练样本的下界。（注意：上一页的定理提供的是充分条件）



从以候选消去算法为例的经典机器学习中有以下总结

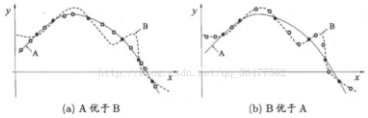
1. 传统的机器学习主要利用假设空间的偏序性质实现对版本空间的缩小从而搜索出与训练样例最佳匹配的概念表示规则；
   不足：只能用于形式较为简单的规则，对更复杂的模型（定量的）存在缺陷。
2. 是否能够实现PAC学习取决于假设空间的复杂性，样本量、版本空间的精度和样本可靠性等因素，但是这些元素需要一个
   不足：未能揭示出假设空间的复杂性与数据结构模式之间的紧密联系。
3. 在同一个训练样本集上进行匹配，可能匹配出多个假设，那么机器学习选择算法的依据是什么呢？

不足：performance的设计过于简单，未能给出可比性的判别准则和实现路径。

# 归纳偏好：什么是一个好模型？

- **归纳偏好：**机器学习算法在学习过程中对某种类型假设的偏好。任何一个有效的机器学习算法必有其归纳偏好。
- **“奥卡姆剃刀”原则：**“若有多个假设与观察一致，则选最简单的那一个。” 注意：奥卡姆剃刀并非唯一可行的原则；

# 什么是最佳拟合




**所有“问题”出现的机会相同、或所有问题同等重要。**

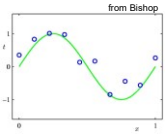
但实际情况并不是这样的，很多时候，我们收集到的样本仅仅体现当时正在关注和正在试图解决的问题。比如，要找到快速从A地到B地的算法，如果我们考虑A地是人民大学东门、B地是北京大学数学楼，那么“骑自行车”是很好的解决方案；但是这个方案对A地是人民大学东门、B地是山东大学的情形显然很糟糕，但研究算法的人所生产出的算法可能被。

所以，NFL定理最重要的寓意，是让我们清楚意识到，**脱离具体问题，空泛地谈论“什么学习算法更好”毫无意义。**

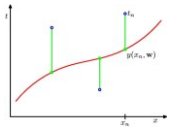
# A simple example: Fitting a polynomial

- The green curve is the true function (which is not a polynomial)噪声
- The data points are uniform in x but have noise in y.

from Bishop

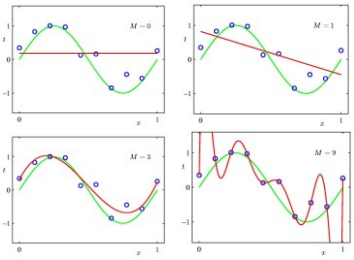


- We will use a loss function that measures the squared error in the prediction of y(x) from x. The loss for the red polynomial is the sum of the squared vertical errors.
- 参考C.bishop(2007)

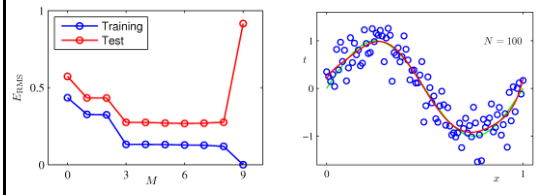


# Some fits to the data: which is best?

from Bishop



- 模型使用不同的阶数拟合的效果呈现出很大的差异；
- 常数（M=0）和一阶（M=1）多项式对数据的拟合相当差，三阶(M=3)似乎效果最好，但我们使用了更高阶的多项式(M=9)，我们发现对每个训练数据有精确的拟合，但是却发生了过拟合(over fitting)。




- 随着多项式阶数增大，系数通常会变大，函数表现出巨大的震荡，有着更大的M表示多项式会被过分调参，多项式与目标值的随机噪声更相符；
- 当数据集规模增加的时候，过拟合问题又变得不太严重，这表明数据量越大，用来拟合数据点的模型应该是更加复杂的，这里的启发是：数据点的数量不应该小于模型可调节参数数量的若干倍，然而后面可以看到的是要求数据量和可调参数之间的比例结构以达到理想的拟合效果是徒劳的，还不如去审视我们的模型空间是不是合适。

# 不同阶数多项式的系数

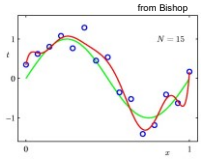
| | M=0 | M=1 | M=3 | M=9 |
|---|---|---|---|---|
| $W_1^*$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $W_2^*$ | | -1.27 | -25.43 | -5321.83 |
| $W_3^*$ | | | 17.37 | 48568.31 |
| $W_4^*$ | | | | -231639.30 |
| $W_5^*$ | | | | 640042.26 |
| $W_6^*$ | | | | -1061800.32 |
| $W_7^*$ | | | | 1042400.18 |
| $W_8^*$ | | | | -557682.99 |
| $W_9^*$ | | | | 125201.43 |

- 对于M=9的情形，训练集的误差是零，符合预期，但是多项式自由度是10，对应于10个系数，而这些系数的数值非常大，对应的函数表现出剧烈的震荡
- 这看起来很矛盾，9阶的多项式包含了所有低阶多项式，也包含了3阶的多项式，如果说3阶多项式是比较理想的，那么4-9阶高阶的多项式系数应该接近于0才是比较正常的，我们原先所设想的使用更复杂多项式的初衷，是因为简单的多项式可以作为复杂多项式的特殊形式而存在，但是当我们用了复杂的多项式以后，我们发现系数的估计过程并没有按照我们预想的方向去进展，而是走向了相反的一面，我们不希望大的系数变得很大，而我们希望很小的系数则变得很大，这表明候选消去的想法是十分可笑的。

# 消除模型复杂度的一种方法
# 压缩系数，防止其变大，变得无法控制
# A simple way to reduce model complexity

- If we penalize polynomials that have big values for their coefficients, we will get less wiggly solutions:

from Bishop




penalized loss function

regularization parameter

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

target value

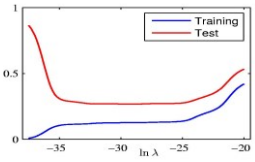
## Polynomial Coefficients如何选择参数

随着多项式阶数增加，系数的估计也会变大，如何调节参数是个问题

| | $\ln\lambda=-\infty$ | $\ln\lambda=-18$ | $\ln\lambda=0$ |
|---|---|---|---|
| $w_0^*$ | 0.35 | 0.35 | 0.13 |
| $w_1^*$ | 232.37 | 4.74 | -0.05 |
| $w_2^*$ | -5321.83 | -0.77 | -0.06 |
| $w_3^*$ | 48568.31 | -31.97 | -0.05 |
| $w_4^*$ | -231639.30 | -3.89 | -0.03 |
| $w_5^*$ | 640042.26 | 55.28 | -0.02 |
| $w_6^*$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^*$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^*$ | -557682.99 | -91.53 | 0.00 |
| $w_9^*$ | 125201.43 | 72.68 | 0.01 |

**Regularization:**

$E_{\mathrm{RMS}}$ **vs.** $\ln\lambda$




为简单起见，假设样本空间 $\mathcal{X}$ 和假设空间 $\mathcal{H}$ 都是离散的. 令 $P(h|X,\mathfrak{L}_a)$ 代表算法 $\mathfrak{L}_a$ 基于训练数据 $X$ 产生假设 $h$ 的概率，再令 $f$ 代表我们希望学习的真实目标函数. $\mathfrak{L}_a$ 的“训练集外误差”，即 $\mathfrak{L}_a$ 在训练集之外的所有样本上的误差为

$$E_{\mathrm{ote}}(\mathfrak{L}_a|X,f)=\sum_h\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a)\,, \tag{1.1}$$

其中 $\mathbb{I}(\cdot)$ 是指示函数，若 $\cdot$ 为真则取值 1，否则取值 0.

考虑二分类问题，且真实目标函数可以是任何函数 $\mathcal{X}\mapsto\{0,1\}$，函数空间为 $\{0,1\}^{|\mathcal{X}|}$. 对所有可能的 $f$ 按均匀分布对误差求和，有

$$\begin{aligned}\sum_f E_{\mathrm{ote}}(\mathfrak{L}_a|X,f)&=\sum_f\sum_h\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a)\\&=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\sum_f\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\\&=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\frac{1}{2}2^{|\mathcal{X}|}\\&=\frac{1}{2}2^{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h\mid X,\mathfrak{L}_a)\\&=2^{|\mathcal{X}|-1}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\cdot 1\,.\end{aligned} \tag{1.2}$$

式(1.2)显示出，总误差竟然与学习算法无关! 对于任意两个学习算法 $\mathfrak{L}_a$ 和 $\mathfrak{L}_b$，我们都有

$$\sum_f E_{\mathrm{ote}}(\mathfrak{L}_a|X,f)=\sum_f E_{\mathrm{ote}}(\mathfrak{L}_b|X,f)\,, \tag{1.3}$$

也就是说，无论学习算法 $\mathfrak{L}_a$ 多聪明、学习算法 $\mathfrak{L}_b$ 多笨拙，它们的期望性能竟然相同! 这就是“没有免费的午餐”定理 (No Free Lunch Theorem，简称 NFL 定理) [Wolpert, 1996; Wolpert and Macready, 1995].

## 2020.10.10作业

- 1.作业阅读Tom Mitchell(1997)教材chap1,chap2,chap7，完成以下阅读报告：
  - 请从网上收集至少5篇有关水上项目应对恶劣天气重大比赛赛事预案和历史事件，列出文献表格进行文献综述，指出这些赛事中主要涉及到哪些项目、哪些天气状况，有哪些相关的预案，时间地点和影响面等信息。

| 编号 | 项目 | 文献类型（预案、事件） | 赛事 | 天气描述 | 时间 | 地点 | 影响描述 | 文献地址 |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |

- 2.编写候选消去程序代码（python,R都可以，以markdown形式输出结果，完成slideP24(周志华1.1表的候选消去版本空间)）