人工智能的数学基础

华东师范大学 数学科学学院 黎芳(教授) 2019年9月9日

第一章 统计学习方法概论

Table of Contents

1.1 统计学习

统计学习特点:

Herbert A. Simon----如果一个系统能够通过执行某个过程改进它的性能，这就是"学习"。

统计学习就是计算机系统通过运用数据及统计方法提高系统性能的机器学习。

统计学习的对象：数据

统计学习的目的：数据预测与分析

统计学习的方法：监督学习、无监督学习和强化学习

步骤：

1. 得到一个有限的训练数据集合
2. 确定包含所有可能的模型的假设空间，即学习模型的集合
3. 确定模型选择的准则，即学习的策略
4. 实现求解最优模型的算法，即学习的算法
5. 通过学习方法选择最优模型
6. 利用学习的最优模型对新数据进行预测或分析

统计学习的研究：

1. 统计学习方法
2. 统计学习理论（统计学习方法的有效性、效率和基本理论）
3. 统计学习应用

统计学习的重要性：

统计学习是处理海量数据的有效方法

统计学习是计算机智能化的有效手段

统计学习是计算机科学发展的一个重要组成部分

1.2 统计学习的分类

1.2.1 基本分类：

1.监督学习（Supervised learning）

输入空间，特征空间和输出空间

样本： $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

$x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$

联合概率分布： 假设输入输出的随机变量X,Y服从联合分布$P(X,Y)$，对学习系统来说，该分布未知. 训练数据和测试数据被看作是依联合概率分布$P(X,Y)$独立同分布地产生的.

假设空间：监督学习的目的是学习一个从输入和输出的映射，用模型表示，假设空间就是这些模型的集合. 模型可以是概率模型或非概率模型，由条件概率$P(Y|X)$或决策函数$Y=f(X)$表示.

问题的形式化：

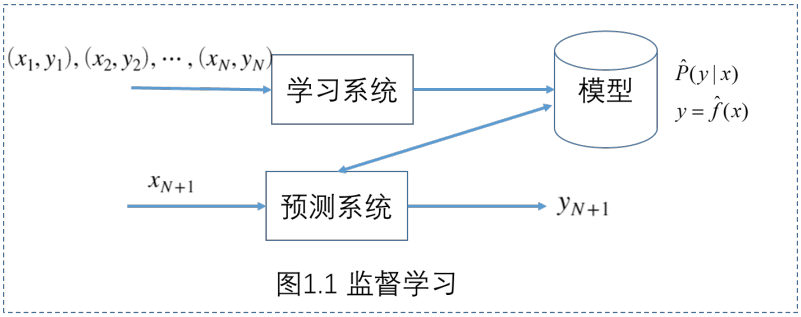


图1.1 监督学习

2.无监督学习（Unsupervised learning）

学习数据中的统计规律或潜在结构，可以实现对数据的聚类、降维或概率估计.

假设X是输入空间，Z是隐式结构空间. 要学习的模型可以表示为：

$z = g(x)$，$p(z|x)$ OR $p(x|z)$

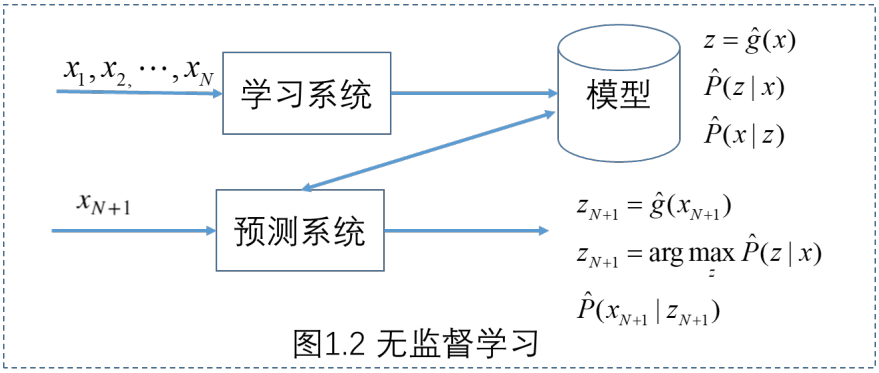


图1.2 无监督学习

3.强化学习（Reinforcement learning）

强化学习是指智能系统在与环境的连续互动中学习最优行为策略的机器学习问题. 假设智能系统与环境的互动是基于马尔可夫决策过程，智能系统能观测到的是与环境互动得到的数据序列. 强化学习的本质是学习序贯决策.

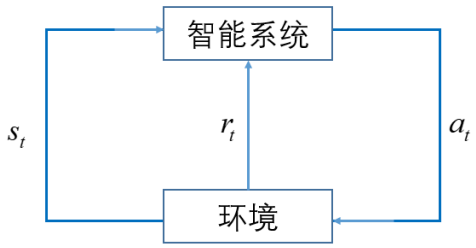


图1.3 智能系统与环境的互动

s -- state, r -- reward, a -- action

强化学习的马尔可夫决策过程是状态、奖励、动作序列上的随机过程，由五元组(S, A, P, r, γ)组成.

- S是有限状态集合
- A是有限动作集合
- P是状态转移概率函数：$P(s'|s,a) = P(s_{t+1} = s'|s_t = s, a_t = a)$

- r是奖励函数：$r(s,a) = E(r_{t+1}|s_t = s, a_t = a)$
- γ是衰减系数：$\gamma \in [0,1]$

策略π定义为给定状态下动作的函数$a = f(s)$或条件概率$P(a|s)$.

状态价值函数：$v_\pi(s) = E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots |s_t = s]$

动作价值函数：$q_\pi(s,a) = E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots |s_t = s, a_t = a]$

强化学习的目标是找出具有最优价值的π^*.

4.半监督学习与主动学习（Semi-**supervised learning and active** learning）

半监督学习是指利用标注数据和未标注数据学习预测模型的机器学习问题. 半监督学习旨在利用未标注数据中的信息，辅助标注数据，进行监督学习，以较低的成本达到较好的学习效果.

主动学习是指机器不断主动给出实例让专家进行标注，然后利用标注数据学习预测模型的机器学习问题.

1.2.2 按模型分类

1.概率模型与非概率模型（或确定性模型）

$P(y|x),\ y = f(x)$

概率模型的代表是概率图模型，与联合概率分布有关. 概率推理基本规则

加法规则：$P(x) = \sum_y P(x,y)$

乘法规则：$P(x,y) = P(x)P(y|x)$

图1.4 基本概率公式

概率模型：决策树、朴素贝叶斯、隐马尔可夫模型、条件随机场、概率潜在语义分析、潜在狄利克雷分配、高斯混合模型

非概率模型：感知机、支持向量机、k近邻、AdaBoost、k均值、潜在语义分析、神经网络

逻辑斯蒂回归既可以看成概率模型也可以看成非概率模型

2.线性模型和非线性模型

线性模型--$y = f(x), z = g(x)$是线性函数

线性模型：感知机、线性支持向量机、k近邻、k均值、潜在语义分析

非线性模型：核函数支持向量机、AdaBoost、神经网络、深度学习

3.参数化模型和非参数化模型

参数化模型--假设模型参数的维数固定，模型可以由有限维参数完全刻画--适合比较简单的情况

非参数化模型--假设模型参数维数不固定，随着训练数据增加不断增大--适合复杂的现实问题

参数化模型：感知机、朴素贝叶斯、逻辑斯蒂回归、k均值、高斯混合模型

非参数化模型：决策树、支持向量机、AdaBoost、k近邻、潜在语义分析、概率潜在语义分析、潜在狄利克雷分配

1.2.3 按算法分类

在线学习（online learning）-- 每次接受一个样本，进行预测，之后学习模型

批量学习（batch learning）-- 一次接受所有数据，学习模型，之后进行预测

必须在线学习的场景：数据依次到达无法存储，系统需要及时做出处理；数据规模很大，

在线学习可以是监督学习，也可以是无监督学习，强化学习本身就拥有在线学习的特点.

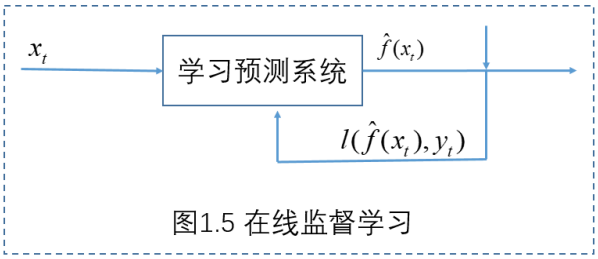


图1.5 在线监督学习

1.2.4 按技巧分类

1.贝叶斯学习

利用贝叶斯定理，计算在给定数据条件下模型的条件概率，即后验概率，并应用这个原理进行模型估计，以及对数据的预测. 用D表示数据.

后验概率：$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$，其中$P(\theta)$是先验概率，$P(D|\theta)$是似然函数.

预测：$P(x|D) = \int P(x|\theta, D)P(\theta|D)d\theta$, x是新样本

贝叶斯估计**vs.**极大似然估计

频率学派--极大似然估计--点估计--假设θ为一个固定值--预测用一个θ

贝叶斯学派--贝叶斯估计--分布估计--θ是随机变量，服从某先验分布--预测用所有θ

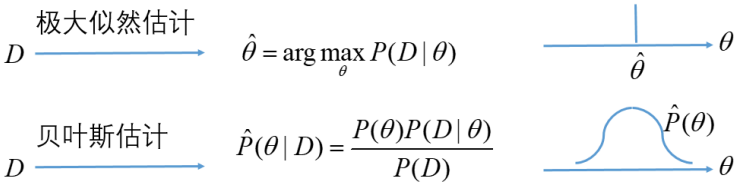


图1.6 贝叶斯估计与极大似然估计

例：假设有一个造币厂生产某种硬币，现在我们拿到了一枚这种硬币，想试试这硬币是不是均匀的。

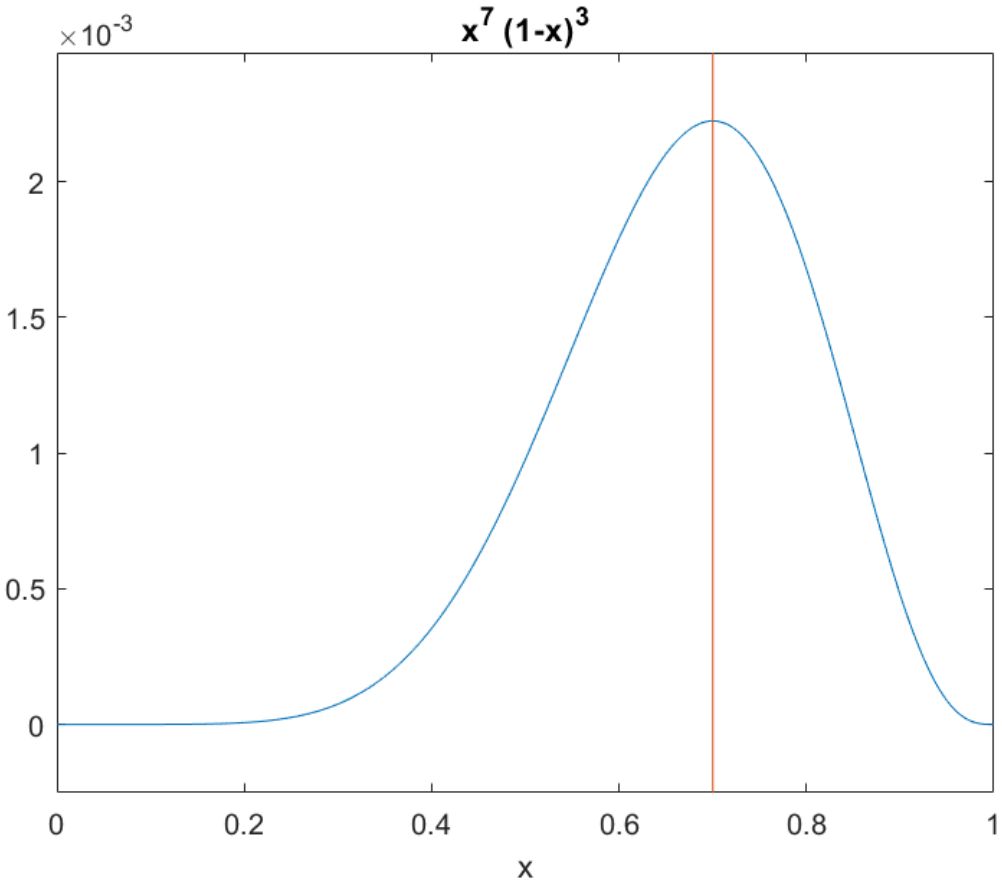
于是我们拿这枚硬币抛了10次，得到的数据x_0是：反正正正正反正正正反。我们想求的正面概率θ是模型参数，而抛硬币模型我们可以假设是二项分布。

MLE：出现实验结果x_0的似然函数

$$p(x_0|\theta) = (1-\theta)\times\theta\times\theta\times\theta\times\theta\times(1-\theta)\times\theta\times\theta\times\theta\times(1-\theta) = \theta^7(1-\theta)^3 = f(\theta)$$

$$\max_\theta p(x_0|\theta) \Rightarrow \theta = 0.7$$

```
close all
figure,ezplot('x^7*(1-x)^3',[0,1])
hold on; plot([0.7,0.7], [-1 0.02])
```

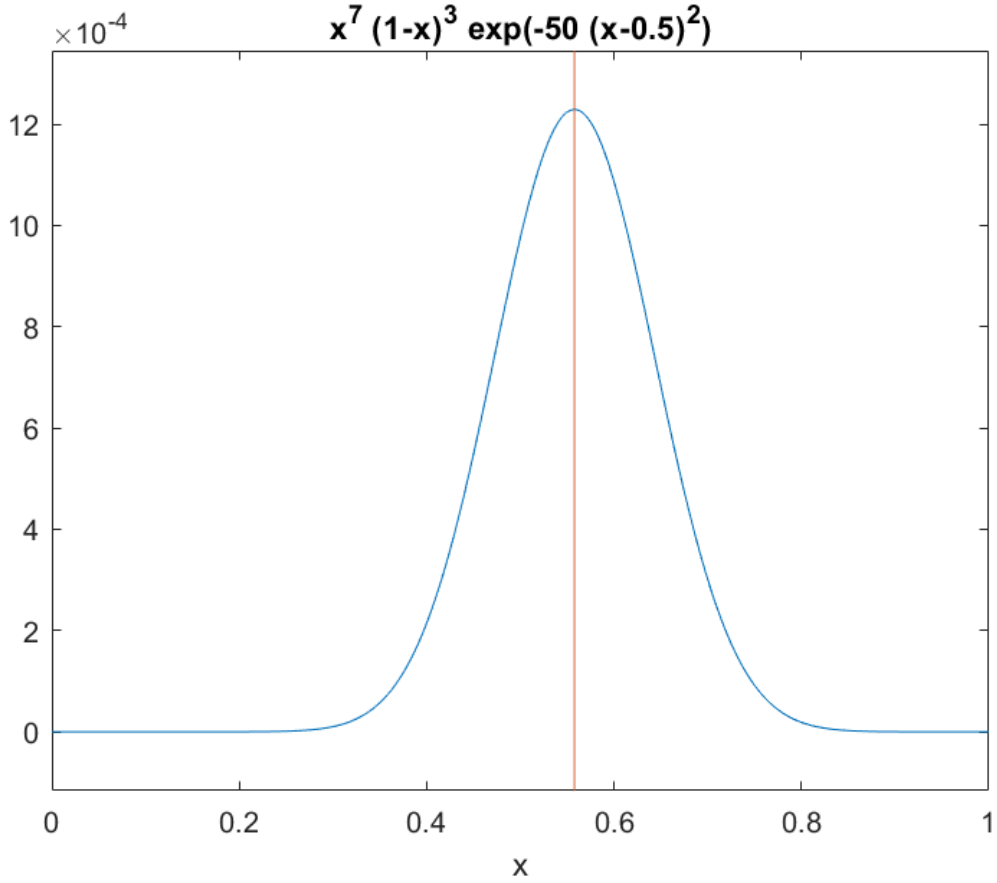


MAP：$\max_\theta p(\theta|x_0) \Leftrightarrow \max_\theta \frac{p(x_0|\theta)p(\theta)}{p(x_0)} \Leftrightarrow \max_\theta p(x_0|\theta)p(\theta)$

若假设P(θ)为均值0.5，方差0.1的高斯函数

$$\max_\theta \theta^7(1-\theta)^3 \exp\left(-\frac{(\theta-0.5)^2}{0.02}\right)$$

```
figure,ezplot('x^7*(1-x)^3*exp(-50*(x-0.5)^2)',[0,1])
hold on; plot([0.5577,0.5577], [-1 0.02])
```




```
fminsearch(@(x)-x^7*(1-x)^3*exp(-50*(x-0.5)^2),0.5)
```

```
ans = 0.5577
```

2.核方法

图1.6 输入空间到特征空间的映射

$K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$

核方法：核函数支持向量机，核PCA, 核k均值方法

1.3 统计学习三要素

统计学习方法=模型+策略+算法

1.3.1 模型

假设空间用$\mathcal{F}$表示，可定义为决策函数的集合：$\mathcal{F}=\{f|Y=f(X)\}$

X,Y是输入空间和输出空间$\mathcal{X},\mathcal{Y}$上的变量，这时$\mathcal{F}$通常是一个由参数向量决定的参数族：

$\mathcal{F}=\{f|Y=f_\theta(X),\theta\in\mathbf{R}^n\}$

假设空间也可定义为：$\mathcal{F}=\{P|P(Y|X)\}$，$\mathcal{F}=\{P|P_\theta(Y|X),\theta\in\mathbf{R}^n\}$

1.3.2 策略

1.损失函数和风险函数

- 0-1损失函数(0-1 loss function)：$L(Y,f(X))=\begin{cases}1, & Y\neq f(X)\\ 0, & Y=f(X)\end{cases}$
- 平方损失函数(quadratic loss function)：$L(Y,f(X))=(Y-f(X))^2$
- 绝对损失函数(absolute loss function)：$L(Y,f(X))=|Y-f(X)|$
- 对数损失函数(logarithmic loss function)：$L(Y,P(Y|X))=-\log P(Y|X)$
- 损失函数的期望：$R_{\exp}(f)=E_P[L(Y,f(X))]=\int_{\mathcal{X}\times\mathcal{Y}}L(y,f(x))P(x,y)\mathrm{d}x\mathrm{d}y$，称为风险函数 risk function 期望损失 expected loss
- 经验风险 empirical risk，经验损失 empirical loss：$R_{\text{emp}}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$，根据大数定律，$N\to\infty\Rightarrow R_{\text{emp}}(f)\to R_{\exp}(f)$

2.经验风险最小化和结构经验风险最小化

- 经验风险最小化最优模型：$\min_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$
- 当样本容量很小时，经验风险最小化学习的效果未必很好，会产生“过拟合over-fitting”
- 结构风险最小化：$\min_{f\in\mathcal{F}}S_{\text{rmf}}=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$
- 为防止过拟合提出的策略，等价于正则化（regularization），加入正则化项（regularizer），或罚项（penalty term）.

1.3.3 算法

- 如果最优化问题有显式的解析式，算法比较简单
- 但通常解析式不存在，就需要数值计算的方法

1.4 模型评估与模型选择

1.4.1 训练误差与测试误差

- 训练误差，训练数据集的平均损失：$R_{\text{cmp}}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right)$
- 测试误差，测试数据集的平均损失：$e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} L\left(y_i, \hat{f}(x_i)\right)$
- 损失函数是 0-1 损失时：$e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} I\left(y_i \neq \hat{f}(x_i)\right)$
- 测试数据集的准确率：$r_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} I\left(y_i = \hat{f}(x_i)\right)$
- $r_{\text{test}} + e_{\text{test}} = 1$

1.4.2 过拟合与模型选择

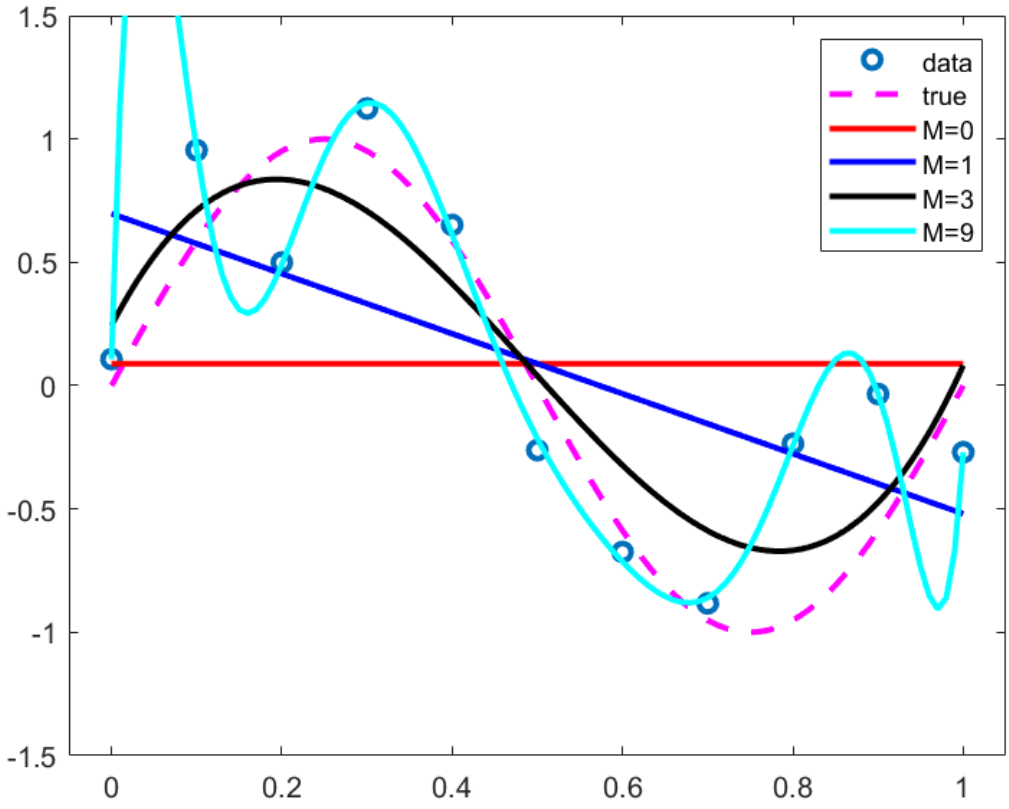
过拟合是指学习时选择的模型包含的参数过多，以至于出现这一模型对已知数据预测得很好，但对未知数据预测得很差的现象.

```
close all
clear all
x = 0:0.1:1;
rng('default')
epsilon = 0.2*randn(size(x));
t = sin(2*pi*x)+epsilon; % noisy
x0 = 0:0.01:1;
t0 = sin(2*pi*x0);
plot(x,t,'o','LineWidth',2)
hold on, plot(x0,t0,'--m','LineWidth',2)
axis([-.05,1.05,-1.5,1.5])
M = [0 1 3 9];
for i = 1:4
p = polyfit(x,t,M(i))
y(i,:) = polyval(p,x0);
end
```

```
p = 0.0890
p = 1×2
   -1.2167    0.6973
p = 1×4
   14.6603  -21.5010    6.6795    0.2430
p = 1×10
10^5 ×
    0.3898   -1.8121    3.5628   -3.8519    2.4880   -0.9730    0.2216   -0.0264 ···
```

```
hold on; plot(x0,y(1,:),'r','LineWidth',2)
hold on; plot(x0,y(2,:),'b','LineWidth',2)
hold on; plot(x0,y(3,:),'k','LineWidth',2)
```

```
hold on; plot(x0,y(4,:),'c','LineWidth',2)
legend('data','true','M=0','M=1','M=3','M=9')
```



M次多项式：$f_M(x,w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$

$$\min L(w) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i,w) - y_i)^2 = \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=0}^{M} w_j x_i^j - y_i\right)^2$$

$$\Rightarrow w_k = \sum_{i=1}^{N}(\sum_{j=1}^{M} w_j x_i^j - y_i)x_i^k = 0$$

$$\sum_{j=1}^{M} w_j \sum_{i=1}^{N} x_i^j x_i^k = \sum_{i=1}^{N} y_i x_i^k$$

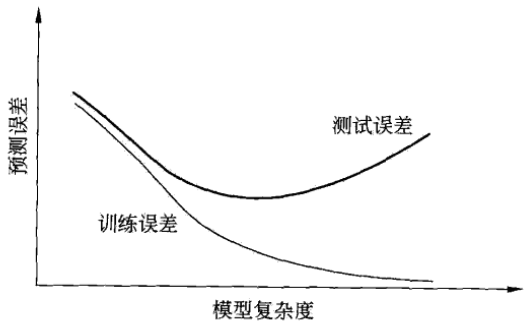
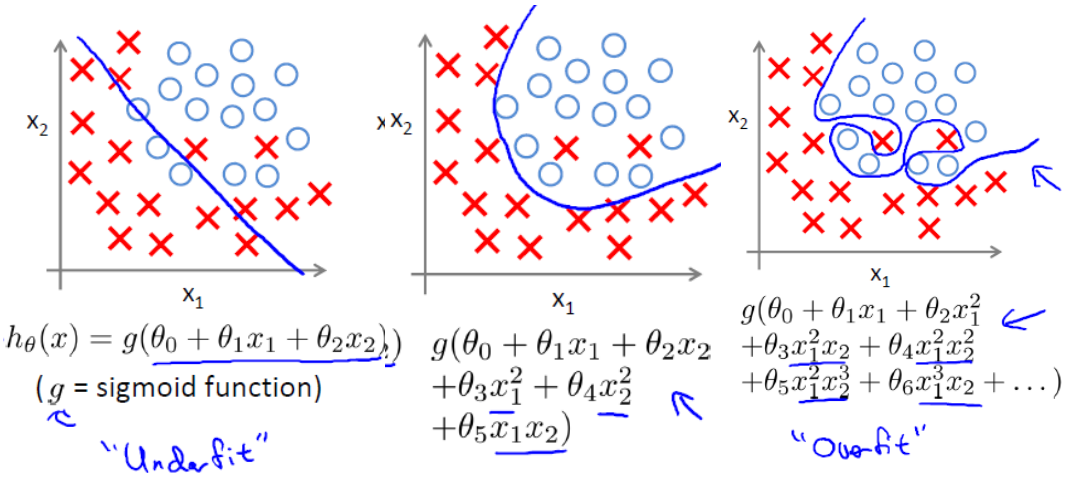


图 1.3　训练误差和测试误差与模型复杂度的关系

逻辑回归同样也存在欠拟合和过拟合问题



1.5 正则化与交叉验证

1.5.1 正则化

一般形式：$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$

回归问题中：

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i; w) - y_i)^2 + \frac{\lambda}{2} \|w\|^2$$

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i; w) - y_i)^2 + \lambda \|w\|_1$$

正则化符合奥卡姆剃刀（Occam's razor）原理：Entities should not be multiplied without necessity.

1.5.2 交叉验证（Cross Validation）

#训练集 training set：　用于训练模型

#验证集 validation set：　用于模型选择

#测试集 test set：　用于最终对学习方法的评估

- 简单交叉验证（**Holdout**）：首先随机地将已给数据分为两部分， 一部分作为训练集，另一部分作为测试集（例如， 70%的数据为训练集， 30%的数据为测试集）： 然后用训练集在各种条件下（例如，不同的参数个数）训练模型， 从而得到不同的模型；在测试集上评价各个模型的测试误差， 选出测试误差最小的模型.
- S折交叉验证（**S-**fold）：首先随机将已给数据切分为S个互不相交的大小相同的子集；然后利用S-1个子集的数据训练模型， 利用余下的子集测试模型； 将这一过程对可能的S种选择重复进行；最后选出S次评测中平均测试误差最小的模型.
- 留一交叉验证（**LOOCV**）：S折交叉验证的特殊情形是S=N, 称为留一交叉验证(leave--0ne--0ut cross validation), 往往在数据缺乏的情况下使用. 这里，N是给定数据集的容量.

1.6 泛化能力（generalization ability）

1.6.1 泛化误差

学习模型的泛化能力是指该方法学习到的模型对未知数据的预测能力.

泛化误差 （generalization error）定义：$R_{\exp}(\hat{f}) = E_p[L(Y, \hat{f}(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{f}(x))P(x, y)\mathrm{d}x\mathrm{d}y$

1.6.2 泛化误差上界

学习方法的泛化能力分析往往是通过研究泛化误差的概率上界进行的，简称为泛化误差上界. 通过比较两种学习方法的泛化误差上界的大小来比较其优劣. 泛化误差上界通常具有以下性质：它是样本容量的函数， 当样本容量增加时， 泛化上界趋于0；它是假设空间容量的函数， 假设空间容量越大， 模型就越难学， 泛化误差上界就越大.

例子：二分类问题的泛化误差上界

$X \in \mathbf{R}^n, \quad Y \in \{-1, +1\}$

期望风险和经验风险：

$$R(f) = E[L(Y, f(X))],\ \hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

经验风险最小化函数是 $f_N = \arg\min_{f\in\mathcal{F}} \hat{R}(f)$

人们更关心的是f_N的泛化能力 $R(f_N) = E[L(Y, f_N(X))]$

定理1.1（泛化误差上界）对二类分类问题，当假设空间是有限个函数的集合$\mathcal{F} = \{f_1, f_2, \cdots, f_d\}$时，对任意一个函数$f \in \mathcal{F}$，至少以概率$1-\delta$，以下不等式成立：

$$R(f) \le \hat{R}(f) + \varepsilon(d, N, \delta)$$

其中 $\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$

证明：Hoeffding不等式

设 $X_1, X_2, \cdots, X_n$ 是独立随机变量，且 $X_i \in [a_i, b_i]$, $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$, 则对任意 $t > 0$, 以下不等式成立：

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

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

对任意函数 $f \in \mathscr{F}$, $R(f) = E[L(Y, f(X))]$, $\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$. 假设损失函数L取值范围是 $[a_i, b_i] = [0, 1]$，由Hoeffding不等式可知，对 $\epsilon > 0$，有

$$P[R(f) - \hat{R}(f) \geq \epsilon] \leq \exp(-2N\epsilon^2)$$

$$P(\exists f \in \mathscr{F} : R(f) - \hat{R}(f) \geq \epsilon) = P(\bigcup_{f \in \mathscr{F}} \{R(f) - \hat{R}(f) \geq \epsilon\})$$

$$\leq \sum_{f \in \mathscr{F}} P(R(f) - \hat{R}(f) \geq \epsilon) \leq d\exp(-2N\epsilon^2)$$

$$\Longleftrightarrow P(R(f) - \hat{R}(f) < \epsilon) \geq 1 - d\exp(-2N\epsilon^2) = 1 - \delta \quad (\delta = d\exp(-2N\epsilon^2))$$

$$\Longleftrightarrow P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$$

1.7 生成模型与判别模型

generative model and discriminative model

生成方法由数据学习联合概率分布 $P(X, Y)$，然后求出条件概率分布 $P(Y|X)$ 为预测模型，即生成模型：

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

模型表示了输入X输出Y的生成关系. 典型例子：朴素贝叶斯、隐马尔可夫模型.

判别方法由数据直接学习决策函数 $f(X)$ 或者条件概率分布 $P(Y|X)$ 作为预测的模型, 即判别模型. 例子：k近邻、感知机、决策树、逻辑斯蒂回归模型、最大熵模型、支持向量机、提升方法和条件随机场.

生成方法的特点：生成方法可以还原出联合概率分布 $P(X, Y)$，而判别方法则不能；生成方法的学习收敛速度更快， 即当样本容量增加的时候， 学到的模型可以更快地收敛于真实模型；当存在隐变量时， 仍可以用生成方法学习， 此时判别方法就不能用.

判别方法的特点： 判别方法直接学习的是条件概率 $P(Y|X)$ 或决策函数 $f(X)$，直接面对预测，往往学习的准确率更高；由于直接学习 $P(Y|X)$ 或 $f(X)$，可以对数据进行各种程度上的抽象、定义特征并使用特征， 因此可以简化学习问题.

1.8 监督学习的应用

1.8.1 分类问题

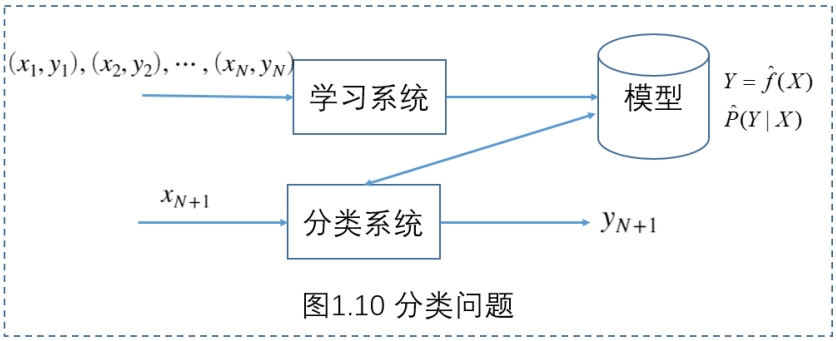


图1.10 分类问题

二分类评价指标

- TP —— true positive 将正类预测为正类
- FN —— false negative 将正类预测为负类
- FP —— false positive 将负类预测为正类
- TN —— true negative 将负类预测为负类

精确率（precision）：$P = \frac{TP}{TP + FP}$（被预测为正的正样本结果数 / 预测为正的样本数）

召回率（recall）：$R = \frac{TP}{TP + FN}$（被预测为正的正样本结果数 / 正样本实际数）

F1值：$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$，$F_1 = \frac{2TP}{2TP + FP + FN}$

准确率（accuracy）：$A = \frac{TP + TN}{TP + FP + FN + TN}$（预测正确的样本数/总样本数）

1.8.2 标注问题 tagging

输入：观测序列 $x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^{\mathrm{T}}, i = 1, 2, \cdots, N$

输出：标记序列或状态序列 $y_i = \left(y_i^{(1)}, y_i^{(2)}, \cdots, y_i^{(n)}\right)^{\mathrm{T}}, i = 1, 2, \cdots, N$

模型：条件概率分布 $P(Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)} | X^{(1)}, X^{(2)}, \cdots, X^{(n)})$

这里$X^{(i)}$取值为所有可能的观测，$Y^{(i)}$取值为所有可能的标记.

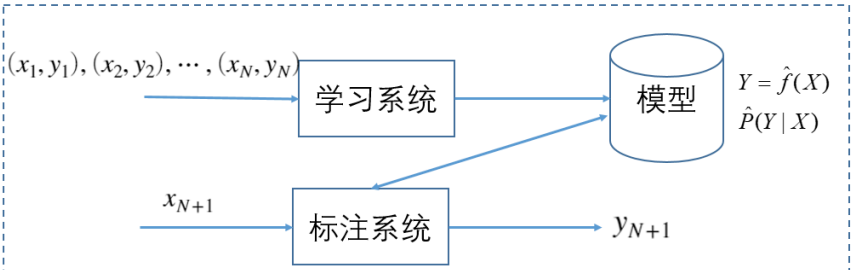


图1.10 标注问题

例：自然语言处理中的词性标注，

例：信息抽取, 如从英文文章中抽取名词短语

标记表示名词短语的“开始”、“结束”或“其他”（分别以B, E, O表示）

输入：At Microsoft Research, we have an insatiable curiosity and the desire to create new technology that will help define the computing experience

输出：At/O Microsoft/B Research/E, we/O have/O an/O insatiable/B curiosity/E and/O the/B desire/E to/O create/O new/B tecbnology/E that/O will/O help/O define/O the/O computing/B experience/E

1.8.3. 回归问题

回归问题的学习等价于函数拟合：选择一条函数曲线使其很好地拟合已知数据且很好地预测未知数据.

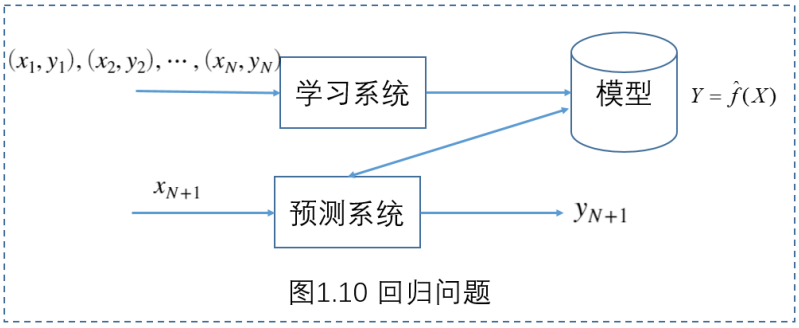


图1.10 回归问题

分类：按输入变量个数分为一元回归和多元回归

按模型的类型分为线性回归和非线性回归

回归学习最常用平方损失函数，此时，可以用最小二乘法求解

例子：股价预测

作业

1.1 说明伯努利模型的极大似然估计以及贝叶斯估计中的统计学习方法三要素．伯努利模型是定义在取值为0与1的随机变量上的概率分布．假设观测到伯努利模型n次独立的数据生成结果，其中k次的结果为1, 试分别用极大似然估计和贝叶斯估计来估计结果为1的概率.

1.2 通过经验风险最小化推导极大似然估计．证明模型是条件概率分布，当损失函数是对数损失函数时，经验风险最小化等价于极大似然估计.