1.1 统计学习方法

1.1.1 监督学习 (supervised learning)

利用训练数据集(输入\输出数据对）学习一个模型，再用模型对测试样本集进行预测。主要用于**分类**、**标注**、**回归**分析。

1. 联合概率分布

监督学习假设输入与输出的随机变量 X 和 Y 遵循联合概率分布 $P(X,Y)$. $P(X,Y)$ 表示分布函数，或分布密度函数. 注意，在学习过程中，假定这一联合概率分布存在，但对学习系统来说，联合概率分布的具体定义是未知的. 训练数据与测试数据被看作是依联合概率分布 $P(X,Y)$ 独立同分布产生的. 统计学习假设数据存在一定的统计规律，X 和 Y 具有联合概率分布的假设就是监督学习关于数据的基本假设.

2. 监督问题的形式化

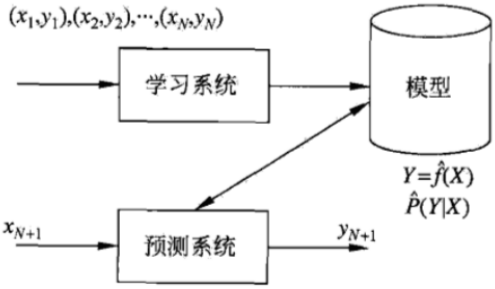

图 1.1　监督学习问题

3. **分类、标注、回归** 对比

监督学习	输入变量	输出变量	应用
分类	离散或连续	离散	文本分类、客户类型分类、垃圾邮件过滤等
标注	观测序列	状态(标记)序列	词性标注、信息抽取等
回归	连续	连续	股价预测问题等

4. 分类问题评价标准

TP —— 将正类预测为正类

FN —— 将正类预测为负类

FP —— 将负类预测为正类

TP —— 将负类预测为负类

精确率：$P=\frac{TP}{TP+FP}$ → 预测正确的正类/预测为正类的总数

召回率：$R=\frac{TP}{TP+FN}$ → 预测正确的正类/总的正类

F1值：$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{R}$ → 精确率和召回率的调和均值，两个率都高时，F1也会高

1.1.2 非监督学习 (unsupervised learning)

用于学习的数据集只有输入（未标记的样本），学习的任务是对于数据进行分析，找到输出。主要用于**聚类**。

1.1.3 半监督学习

监督学习和非监督学习的结合，它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题，主要用于**半监督分类**、**半监督回归**、**半监督聚类**、**半监督降维**。

1.1.4 强化学习

学习者在学习过程中不断与环境交互，会从环境中得到一定的奖赏，根据奖赏再不断的学习，直到达到一个更优的策略。

1.1.5 实现统计学习方法的步骤

(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) 确定模型选则的准则，即策略；

(4) 实现求解最优模型的算法，即算法；

(5) 选择最优的算法；
(6) 利用最优模型对新数据进行预测或分析。

1.2 统计学习三要素

1.2.1 模型

即所要学习的条件概率分布或决策函数

1.2.1.1 假设空间

包含所有可能的条件概率分布或决策函数，可以定义为决策函数的集合或条件概率分布族，即

$$F = \{f|Y = f(X)\}或F = \{P|P(Y|X)\}$$

1.2.1.2 参数空间

包含决策函数或条件概率分布模型中涉及的所有参数向量

$$F = \{f|Y = f_\theta(X)\ ,\ \theta \in R^n\}或F = \{P|P_\theta(Y|X)\ ,\ \theta \in R^n\}$$

1.2.2 策略

1.2.2.1 损失函数与风险函数

1. **损失函数**（loss function）或**代价函数**（cost function）
   对于给定的输入x，由模型f(X)给出相应的输出，但是预测的输出f(x)与真实值Y可能存在不一致，用一个损失函数或者代价函数来**度量预测错误的程度**。损失函数L(Y,f(X))是预测值f(X)和真实值Y的**非负实值函数。损失函数值越小，模型就越好**。常见的损失函数:

   a. 0-1损失函数

   $$L(Y, f(X)) = \begin{cases} 1 & Y \neq f(x) \\ 0 & Y = f(X) \end{cases}$$

   b. **平方损失函数**

   $$L(Y, f(X)) = (Y - f(X))^2$$

   c. 绝对损失函数

   $$L(Y, f(X)) = |Y - f(X)|$$

   d. 对数损失函数(logarithmic loss function)或对数似然损失函数

   $$L(Y, f(X)) = -logP(Y|X)$$

2. **风险函数**(risk function) 或 **期望损失** (expected loss)
   模型$f(x)$关于**联合分布**$P(X,Y)$的平均意义下的损失:

   $$R_{exp} = E_P[L(Y, f(X))] = \int_{X \cdot Y} L(Y, f(X))P(X,Y)dxdy$$

3. **经验风险** (empirical risk)
   模型$f(X)$关于**训练数据集**的平均损失，根据**大数定律**，当样本容量N趋于无穷时，经验风险趋于期望

风险，所以可以用**经验风险估计期望风险**。

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

其中，$T = \{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$ 为训练集

1.2.2.2 经验风险最小化和结构风险最小化

1. **经验风险最小化**（Empirical Risk Minimization，ERM）
   ERM的策略认为经验风险最小的模型是最优的模型 → **极大似然估计**（某些条件下），即

   $$min\{R_{emp}\}$$

   a. 当样本容量足够大时，经验风险最小化能保证有很好的学习效果

   b. 当样本容量很小时，经验风险最小化学习的效果未必很好，甚至会产生“**过拟合**”问题

2. **结构风险最小化**（structural risk minimization，SRM）
   为了**防止过拟合**而提出的策略，结构风险在经验风险上加上表示模型复杂度的正则化项（regularization）或罚项（penalty term）。SRM策略认为结构风险最小的模型就是最优模型 → 贝叶斯估计中的**最大后验概率估计**。

   $$min\{R_{emp} + \lambda J(f)\}$$

   其中，$J(f)$ 为模型的复杂度，$f \in F$

1.2.2 算法

→ 经验风险或结构风险**最优化问题**

1.3 模型评估与模型选择

1.3.1 训练误差与测试误差

1. **训练误差**
   模型关于**训练数据集**的平均损失（经验风险）

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

2. **测试误差**
   模型关于**测试数据集**的平均损失（经验风险）

   $$e_{test} = \frac{1}{N'}\sum_{i=1}^{N'} L(y_i, \hat{f}(x_i))$$

3. 误差率 (error rate)

   $$e_{test} = \frac{1}{N'}\sum_{i=1}^{N'} I(y_i \neq \hat{f}(x_i))$$

其中，I 为指示函数(indicator function)，$I=\begin{cases}1 & y \neq \hat{f}(x) \\ 0 & y=\hat{f}(x)\end{cases}$

4. 准确率 (accracy rate)

$$r_{test}=\frac{1}{N'}\sum_{i=1}^{N'}I(y_i=\hat{f}(x_i))$$

1.3.2 过拟合

1. **过拟合** (over-fitting)
   学习时选择的模型所包含的参数过多（复杂度过高），以致于出现这一模型**对已知数据预测得很好，但对未知数据预测得很差的现象**。

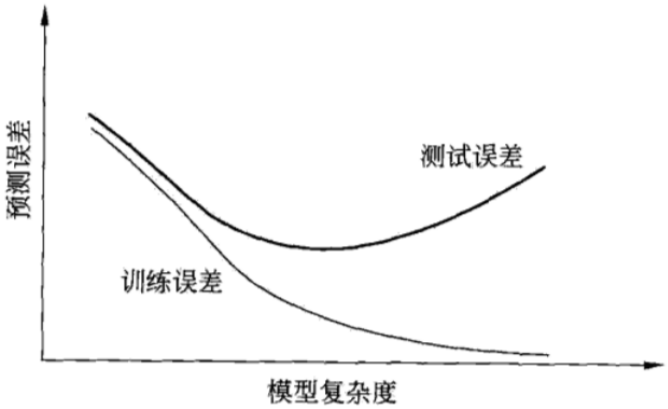

图 1.3　训练误差和测试误差与模型复杂度的关系

1.3.3 模型选择

1.3.3.1 正则化 (regularization)

正则化是**结构风险最小化**策略的实现，是在经验风险上加一个**正则化项**或**罚项**。一般是模型复杂度的单调递增函数，模型越复杂，正则化值越大。**作用**是选择经验风险与模型复杂度同时较小的模型。正则化项的不同形式:

1. L_2范数 (平方损失)

$$L(w)=\frac{1}{N}\sum_{i=1}^{N}(f(x_i;w)-y_i)+\frac{\lambda}{2}||w||_2^2$$

   其中，w 为参数向量，$||w||_2=\sqrt{\sum_{i=1}^{N}w_i^2}$

2. L_1范数 (绝对值损失)

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

1.3.3.2 交叉验证

如果给定的样本数据充足，进行模型选择的一种简单方法是随机地将数据集切成三部分，分别为**训练集**、**验证集**（validation set）和**测试集**。训练集用来训练模型，**验证集用于模型的选择**，测试集用于最终对方法的

评估。但是由于在许多实际应用中数据是不充分的，为了选择好的模型，可以采用交叉验证方法。

1. **基本思想**
   重复的使用数据，把给定的数据进行切分，将切分的数据集组合为训练集和测试集，在此基础上**反复**地进行训练、测试以及模型选择
2. 方法
   a. 简单交叉验证
      将已给数据随机分为两部分，分别用作训练集和测试集；然后用训练集在各种条件下，训练模型，从而得到不同的模型；最后用测试集评价模型。
   b. **S折交叉验证** (S-fold cross validation)
      首先将已给数据随机分为S个互不相交、大小相同的子集；然后利用S-1个子集的数据训练模型，剩余1个子集测试模型；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。
   c. 留一交叉验证
      S折交叉验证的特殊形式是S=N，其中N是给定数据集的容量

1.4 泛化能力

指由该方法学习到的模型对**未知数据**的预测能力。

1. **泛化误差** (generalization error)
   现实中，可以通过**测试误差**来**评价学习方法的泛化能力**（测试数据集的经验风险），但是由于测试数据集有限，所以从理论上进行分析：
   用学习到的模型对未知数据预测的误差即为泛化误差（测试数据集的期望风险）

$$R_{exp}(\hat{f}) = E_P[L(Y, \hat{f}(X))] = \int_{x \cdot y} L(Y, \hat{f}(X)) P(X, Y) dxdy$$

2. 泛化误差上界
   可以理解为泛化误差的可能最大值，等于经验风险+一个函数（参数是样本容量和假设空间容量），即

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

   a. 泛化误差上界是样本容量(N)的单调递减函数，当样本容量增加时，泛化上界趋于0
   b. 泛化误差上界也是假设空间容量(d)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

1.5 生成模型和判别模型

监督学习方法可以分为生成方法和判别方法，所学到的模型分别称为生成模型和判别模型。

1. **生成模型** (generative model)
   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(YIX)作为预测的模型，即生成模型：

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

   之所以称为生成方法，是因为**模型表示了给定输入X产生输出Y的生成关系**。

a. 典型的生成模型

朴素贝叶斯法和**隐马尔可夫模型**

b. 优点

可以**还原联合概率分布**P(X,Y)；学习收敛速度更快；存在隐变量时，仍可以用生成方法学习

2. **判别模型** (discriminative model)

由数据直接学习决策函数f(X)或者条件概率分布P(YIX)作为预测的模型，即判别模型。判别方法关心的是对给定的输入X，应该预测什么样的输出Y。

a. 典型的判别模型

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、条件随机场等

b. 优点

准确率更高；简化学习问题

1.6 引用

> https://www.cnblogs.com/naonaoling/p/5689830.html