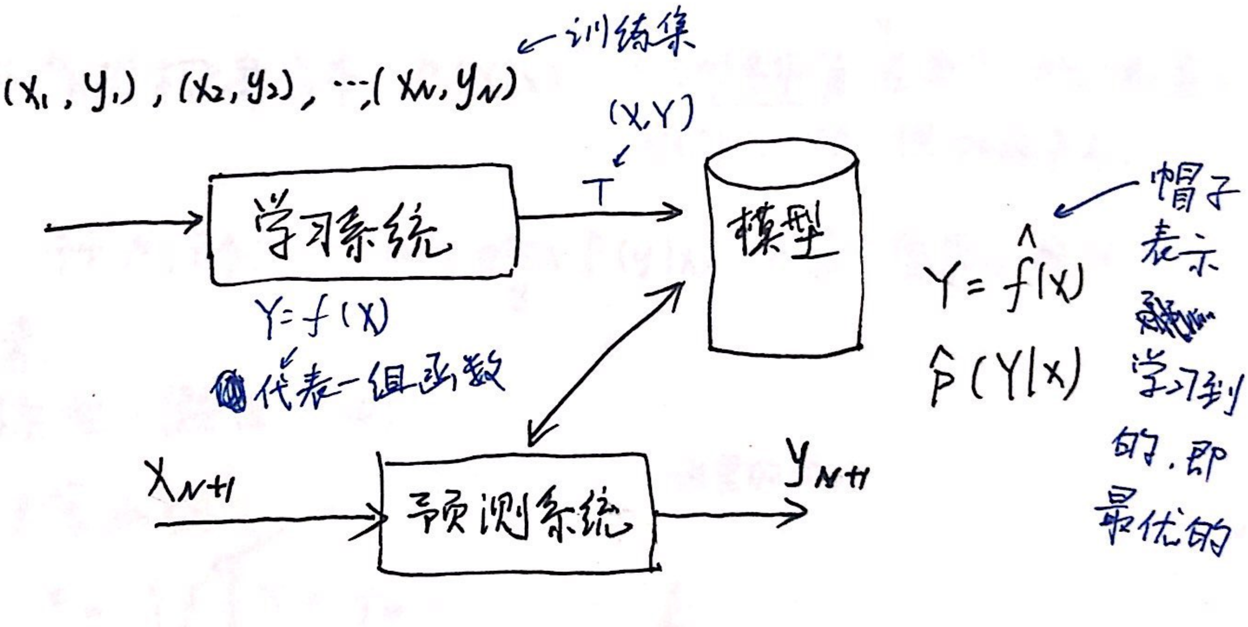
1.1 统计学习

监督学习的实现步骤

1. 有限训练集合
2. 解设模型假设空间，即备选模型确定
3. 确定模型选择准则，即学习策略
4. 实现求解最优模型的算法

（2–4：学习系统）

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or分析



1.2 监督学习

拿到的数据有随机性，
∴用联合概率分布表示

训练集

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

实例x的特征向量

$x=(x^{(1)},x^{(2)},\cdots,x^{(n)})^T$

一般输入空间=特征空间
但，若x为输入空间，但我们要将 $x \longrightarrow x^2+x^3$，则 x^2 与 x^3 为特征空间

模型：

1) 决策函数 $Y=f(X)$ 表示一组假设空间的函数

预测形式 $y=\hat{f}(x)$ 最优的，就一个值

2) 条件概率分布 $P(Y|X)$ 表示 Y有很多取值，在每个不同的取值上概率不一样，但加起来为1

预测形式 $\arg\max\limits_{y} P(y|x)$ 取最大值对应的y.

三要素

① 模型（假设空间）

决策函数

代表每个函数（一组里的一个）

$F=\{f \mid Y=f_\theta(X),\theta\in R^n\}$ 给定假设空间，求出θ，即求出了最优。

例：$Y=a_0+a_1X$，$\theta=(a_0,a_1)^T$

条件概率分布

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

例 $Y\sim N(a_0+a_1X,\sigma^2)$，$\theta=(a_0,a_1)^T$

② 策略：

损失函数（针对每个实例来说的）

针对决策函数

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

真实值　预测值

$$L(Y, f(X)) = (Y - f(X))^2$$ （回归问题） 平方损失（对差值大的实例敏感）

$$L(Y, f(X)) = |Y - f(X)|$$ （回归问题） 绝对值损失

针对条件概率分布

$$L(Y, P(Y|X)) = \underbrace{-\log P(Y|X)}_{\text{对数似然函数}}$$

经验风险最小化：

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

结构风险最小化

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \underbrace{\lambda \boxed{J(f)}}_{\text{正则项}}$$

衡量数据集上经验与模型复杂度重要程度

表示f这个模型的复杂度

力求 $\boxed{\text{经验风险小 + 模型复杂度低}}$

1.4 模型评估与模型选择

训练误差 (在训练集上)

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

测试误差 (在测试集上)

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

过拟合

例1: 多项式拟合问题 (书上M次多项式拟合)

M次项增加会导致"噪声"增加，∴力求模型复杂度低

1.5 正则化与交叉验证 (模型选择的2个方法)

考虑

正则化:

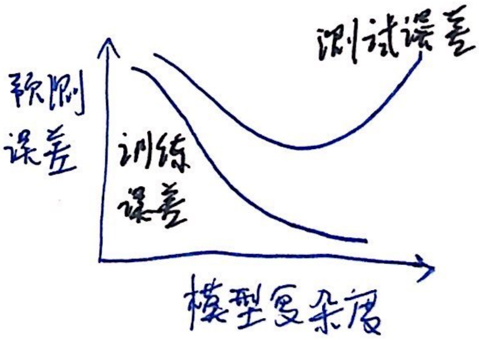
最小化结构风险

$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

交叉验证

训

训 验



M=1

M=3

M=9

1.6 泛化能力

定理 1.1 泛化误差上界

对于二分类问题，当假设空间是有限个函数的集合 $F=\{f_1, f_2, \cdots, f_d\}$ 时，对任意一个函数 $f \in F$，至少以概率 $1-\delta$，以下不等式成立：

念 yī pè xiu

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

期望风险（在所有数据上）

经验风险（在训练数据集上）

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

备选模型的个数 $d \uparrow$, $R(f) \uparrow$

样本量 $\uparrow$, $R(f) \downarrow$

1.7 生成模型与判别模型

生成方法 (X与Y都是随机的变量) 给定X，不仅要学习Y的概率分布，还要学习X与Y的联合概率分布.

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

判别方法 (不考虑X的随机)

$f(X)$ 或 $P(Y|X)$

(给定X，学习 $f(x)$ 或 Y的概率分布)

1.8 分类问题 (分类问题中的模型叫分类器)

(2分类代表2个类型)

0 1

TP — 将正类预测为正类 (1)

FN — 正 → 负 (0)

FP — 负 → 正

TN — 负 → 负

精确率

$$P = \frac{TP}{TP+FP}$$

召回率

$$R = \frac{TP}{TP+FN}$$

1.9 标注问题

输入

$X=(x^{(1)},x^{(2)},\cdots,x^{(n)})^T$

输出

$y=(y^{(1)},y^{(2)},\cdots,y^{(n)})^T$

例：文本分类

输入：At Microsoft Research

$$X=(\underset{\text{At}}{x^{(1)}},\underset{\text{Microsoft}}{x^{(2)}},\underset{\text{Research}}{x^{(3)}})$$

输出：At/O Microsoft/B Research/E

Microsoft Research

$$y=(\underset{O}{y^{(1)}},\underset{B}{y^{(2)}},\underset{E}{y^{(3)}})$$

意义

可以将文本分为一个一个词组

1.10 回归问题

输出为连续的值

★ 极大似然估计

例：掷硬币，设出现正面向上的概率为 θ

$$X_i=\begin{cases}1, & 正\\ 0, & 反\end{cases}, \quad X_i \sim b(1,\theta), \quad P(X=x)=\theta^x(1-\theta)^{1-x}$$

$$\max L(\theta)=P(X_1=x_1|\theta)\cdots P(X_n=x_n|\theta)$$

$$=\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}$$

∵ 每次掷硬币独立，
∴ 可以把联合概率分布写成连乘形式

$$\max \ln L(\theta)=\sum\left[\ln\theta^{x_i}+\ln(1-\theta)^{1-x_i}\right]$$

$$=\sum x_i\ln\theta+(n-\sum x_i)\ln(1-\theta)$$

极大似然目的：
想找到1个 θ，使得样本出现的概率最大，即要最大化这个似然函数.

等价为：
最大化似然函数的对数.

求导

$$\frac{\partial \ln L(\theta)}{\partial\theta}=\frac{\sum x_i}{\theta}-\frac{n-\sum x_i}{1-\theta}=0$$

求得 $\boxed{\hat{\theta}=\frac{\sum x_i}{n}}$

步骤：① 根据样本概率分布，写出样本的联合概率似然函数

② 通过最大化似然函数求得参数估计值

★ 极大似然估计（完全根据样本信息）

★贝叶斯估计（不仅根据样本信息，还有先验信息）

人为认为的他应该是这样的分布

∵一开始有了 $\pi(\theta)$

根据 $x_1,\cdots,x_n$（即样本）

去调整对 θ 概率分布判断

即 ← 条件

$P(\theta \mid x_1,\cdots,x_n)$

$$= \frac{P(\theta, x_1,\cdots,x_n)}{P(x_1,\cdots,x_n)}$$

← 联合密度

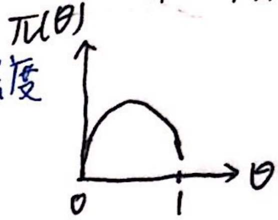
← 边际密度

（根据乘法公式）

$$= \frac{\pi(\theta)\cdot P(x_1|\theta)\cdots P(x_n|\theta)}{\int P(\theta, x_1,\cdots,x_n)\,d\theta}$$

← 联合密度对 θ 求积分

（概率密度分布）



一般[0,1]之间的概率分布可以用 β 分布表示。

∵ $\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ ← 伽马函数

∴分布由 α 与 β 决定

将 $\pi(\theta)$ 代入，为了简化，写成正比例形式 $\propto$（只写与 θ 有关的）

$$= \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\cdot \prod \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta}$$

← 参数为 $\sum x_i+\alpha$，与 $n-\sum x_i+\beta$

步骤：

① 根据参数，给定后验信息的样本分布

② 给 θ 一个具体的值，找一个使得后验分布最大的值，即找 β 分布的众数，为 $\hat{\theta} = \frac{\sum x_i+\alpha-1}{n+\alpha+\beta-2}$

对比极大似然估计与贝叶斯估计

MLE: $\hat{\theta} = \frac{\sum x_i}{n}$

样本个数.

Bayes: $\hat{\theta} = \frac{\sum x_i + \alpha - 1}{n + \alpha + \beta - 2}$ ① 当 $n \to \infty$

$\text{Bayes} \to \frac{\sum x_i}{n}$

② 当 $n=1$

$\text{MLE} \to 1$

$\text{Bayes} \to \frac{\alpha}{\alpha+\beta-1}$

意义: 若样本量大,

先验分布微不足道.

若样本量小,

Bayes 不会出现极端情况

★泛化误差上界

对于二分类问题，当假设空间是有限个函数的集合 $F=\{f_1, f_2, \cdots, f_d\}$ 时，对任意一个函数 $f\in F$，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 $\varepsilon(d,N,\delta)=\sqrt{\frac{1}{2N}\left(\log d+\log\frac{1}{\delta}\right)}$

$\hat{R}(f)$：经验风险

Hoeffding 不等式

① 定义独立随机变量 $X_1, X_2, \cdots X_n$，$X_i\in[a_i, b_i]$

$S_n = X_1+X_2+\cdots X_n$，$ES_n = E(\sum X_i)$

② $P(S_n - ES_n \ge t) \le \exp\left(-\frac{2t^2}{\sum(b_i-a_i)^2}\right)$ ← 损失函数取值

③ 不考虑随机变量的和，考虑随机变量的均值

$\bar{X}_n = \frac{S_n}{n}$ $E(\bar{X}_n) = \frac{ES_n}{n}$

可替换

$P(\bar{X}_n - E(\bar{X}_n) \ge t) = P(S_n - ES_n \ge nt)$

$P(E(\bar{X}) - \bar{X} \ge t)$ $\le \exp\left(-\frac{2n^2t^2}{\sum(b_i-a_i)^2}\right)$ ← n^2阶 ← n阶

$\le e^{-n}$ ← n阶

∴ 当样本量足够大，概率为0

④ $\hat{R}(f) = \frac{1}{N}\sum L(X_i, f(X_i))$ ← 经验　　$R(f)$ ← 期望

代入 Hoeffding 不等式　　有N个候选函数

$P(R(f) - \hat{R}(f) \ge t) \le \exp\left(-\frac{2N^2t^2}{N}\right) = \exp(-2Nt^2)$

(对于1个备选模型来说)　　∵ $[a_i, b_i] = [0,1]$ ∴ 为 N

等价于 不希望有一个备选模型，使得期望风险到经验风险的距离大于 t，即

$$P(\exists f\in F: R(f)-\hat{R}(f)\geq t)$$

并集

$$=P(R(f_1)-\hat{R}(f_1)\geq t \cup \cdots \cup R(f_d)-\hat{R}(f_d)\geq t)$$

$$\leq \sum_i P(R(f_i)-\hat{R}(f_i)\geq t)$$

$$\leq d\exp(-2Nt^2)$$

等价于

$$P(\forall f\in F, R(f)-\hat{R}(f)\leq t)\geq 1-\boxed{d\exp(-2Nt^2)}$$

设为 δ

解得 $t=\sqrt{\frac{1}{2N}(\log d+\log\frac{1}{\delta})}$ $\varepsilon(d,N,\delta)$

书上为 ε