1. 统计学习

基于数据构建统计模型对数据进行分析

数据：所有可以被记录的都是数据

三要素：模型(model)、策略(strategy)和算法(algorithm)

统计学习分类：有监督学习、无监督学习、强化学习

学习步骤：

1) 明确学习模型

2) 明确评价准则

3) 训练最优模型

2. 统计学习的分类

2.1 基本分类

1. 监督学习：从标注数据中学习预测模型

(1) 输入空间、特征空间和输出空间

输入与输出所有可能取值的集合

每个具体的输入是一个实例(instance)

由特征向量表示

$x=(x^{(1)}, x^{(2)}, \cdots, x^{(p)}) \in R^p$

a) 输入与输出可以看作定义有输入空间上随机变量的取值

b) 监督学习从训练数据(training data)对测试数据(testing data)进行预测 $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ 又称为样本.

c) 输出变量 Y { 连续型 (regression)
离散型 (classification)

(2) 联合概率分布

X Y 联合概率分布为 $P(X, Y)$

认为样本 (x_i, y_i) 依据 $P(X, Y)$ 独立同分布产生

(3)假设空间：输入空间到输出空间映射的集合

2.无监督学习

(1)从无标注数据中学习模型的机器学习问题

(2)学习数据的统计规律或潜在结构

2.2 按模型分类

1.概率模型($P(Y|X)$ eg.朴素贝叶斯)和非概率模型($y=f(x)$, e.g.神经网络)

2.线性模型和非线性模型

3.参数模型和非参数模型

参数化模型可由有限维参数完全刻画

3.统计学习的三要素

1.模型：$\mathcal{F}=\{f|Y=f(X)\}$

2.策略：按何种准则选择最优模型

(1)损失函数和风险函数

损失：度量一次预测的好坏

风险：度量平均意义模型预测的好坏

输入变量为X，模型的预测为$f(X)$，真实值为Y

$L(Y,f(X))$度量预测错误的程度

常见的损失函数：

(a) 0-1损失函数：$L(Y,f(X))=I(Y\neq f(X))$

(b)平方损失函数：$L(Y,f(X))=(Y-f(X))^2$

(c)绝对损失：$L(Y,f(X))=|Y-f(X)|$

(d)对数损失或对数似然：$L(Y,f(X))=-\log P(Y|X)$

风险函数：损失函数的期望

$$E_{exp}(f)=E_p(L(Y,f(X)))=\int L(y,f(x))P(x,y)dxdy$$

给定一个训练集$T=\{(x_1,y_1),\cdots,(x_N,y_N)\}$

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

(2) 经验风险最小化与结构风险最小化

· 经验风险最小化：$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$

eg：极大似然估计（损失函数 → 对数似然函数）

样本量较大时表现较好

· 结构风险最小化：$R_{srm}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$

$J(f)$是模型的复杂度，$J(f)$越大模型越复杂

3. 算法：指模型的具体求解方式，可归结为最优化问题的求解

4. 模型的评估与选择

1. 训练误差和测试误差

训练误差：关于训练集的平均损失 $R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$

测试误差：关于测试集的平均损失 $e_{test} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$

2. 过拟合与模型选择

过拟合指学习时选择的模型参数过多，以至于模型对训练数据拟合较好，但对未知数据预测很差的现象

5. 正则化与交叉检验

1. 正则化

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

λ为调节参数（training parameter），λ越大倾向于选择简单模型

核心：选择经验风险与复杂度同时较小的模型

2. 交叉验证

如果数据充足，将数据分为三部分：训练集、验证集和测试集

验证集用于模型选择，测试集用于模型评估

如果数据不充足，就需要重复利用数据

(1) 简单交叉验证

将数据随机分成两部分，一部分做训练集，一部分做测试集。在不同参数下训练模型，进行模型评估

(2) S折交叉验证

随机将数据切分成S折互不相交的子集，利用S-1折子集进行数据训练，用余下的一折进行模型测试。这种测评进行S次，选择S次误差最小的模型

(3) 留一交叉验证：S=N的S折交叉验证

6. 泛化能力

1. 泛化误差

如果学到的模型是$\hat{f}$，那么用这个模型对未知数据的预测误差就是泛化误差：

$$R_{exp}(\hat{f}) = \int L(y, \hat{f}(x)) P(x,y) dxdy$$

泛化误差就是所学习到模型的期望风险，表示学习方法的泛化能力。