# 统计学习方法概论

## 统计学习

### 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

## 监督学习

**输入空间**：所有输入可能取值的集合

**输出空间**：所有输出可能取值的集合

**特征向量**：每个具体的输入是一个实例

**特征空间**：所有特征向量存在的空间

**特征**：特征空间的每一维度

**回归问题**：输入输出变量均为连续变量的预测

**分类问题**：输出变量为有限个离散变量的预测

**标注问题**：输入输出变量均为变量序列的预测

注：有时候假设输入空间与特征空间为相同的空间；有时候输入空间与特征空间为不同的空间，将实例从输入空间映射到特征空间。模型实际上都是定义在特征空间上的。

记：特征向量：（均以列向量表示）

第i个特征向量：

训练数据集：

## 统计学习三要素



### 模型

例如决策函数表示非概率模型，条件概率表示概率模型。

### 策略

损失函数与风险函数，用表示

1. 二分类(准确率，召回率)
2. 平方损失函数(最小二乘法)
3. 绝对损失函数
4. 对数损失函数(最大似然法)

期望损失：，又称平均损失、经验损失、经验风险。

**经验风险最小化**

一般认为：在所有模型中，值最小的模型为**最优模型**。

**结构风险最小化**(等价于**正则化**)(防止过拟合)

一般认为：在所有模型中，值最小的模型为**最优模型**。

其中为模型的复杂度，是定义在假设空间上的泛函，模型越复杂，复杂度就越大；反之模型越简单，复杂度就越小。是系数，用以权衡经验风险和模型复杂度，结构风险小需要经验风险和模型复杂度同时小，结构风险小的模型往往对训练数据以及未知的测试数据都有较好的预测。

### 算法

学习模型的具体计算方法，用什么样的计算方法求解最优模型。

## 模型评估和模型选择

## 正则化和交叉验证

## 泛化能力

## 生成模型和判别模型

## 分类问题

## 标注问题

## 回归问题

# 感知机

# K邻近法

# 朴素贝叶斯

# 决策树

# 逻辑斯谛回归和最大熵模型

# 支持向量机

# 提升方法

# EM算法及其推广

# 隐马尔可夫模型

# 条件随机场

# 统计学习方法总结