《机器学习基础》
Foundations of Machine Learning

课程复习

重庆大学计算机学院

成绩评定

成绩构成

考核方式	总分	占总成绩的比例
课堂教学	100分	60%
实验项目	10分/项目，共3个项目	30%
作业及实验PPT展示	10分	10%

本课程主要考试范围

主要参照-周志华的专著《机器学习》，主要章节内容如下：

第1章　绪论

第2章　学习模型评估与选择

第3章　线性模型

第4章　决策树

第5章　神经网络

第6章　支持向量机

第7章　贝叶斯分类器

第8章　集成学习

第9章　聚类

第1章：绪论

- 什么是机器学习

“假设用P来评估计算机程序在某任务类T上的性能，若一个程序通过利用经验E在T中任务上获得了性能改善，则我们就说关于T和P，该程序对E进行了学习”

- **监督学习（supervised learning）**
- **无监督学习（unsupervised learning）**
- **分类和回归的原理及区别**
- 泛化能力和归纳偏好（奥卡姆剃刀）
- 没有免费的午餐定理

第2章：模型评估与选择

- 过拟合和欠拟合以及相应的解决方案
- 评估方法：泛化性能
  - 留出法：
  - 交叉验证法：
  - 自助法：
- 性能度量
- 偏差与方差-性能解释

第3章：线性模型

- 线性回归（最小二乘法）
- 二分类任务
  - 对数几率回归
  - 线性判别分析（LDA）
- 多分类学习
  - 一对一
  - 一对其余
  - 多对多
- 类别不平衡问题：欠采样、过采样、再缩放

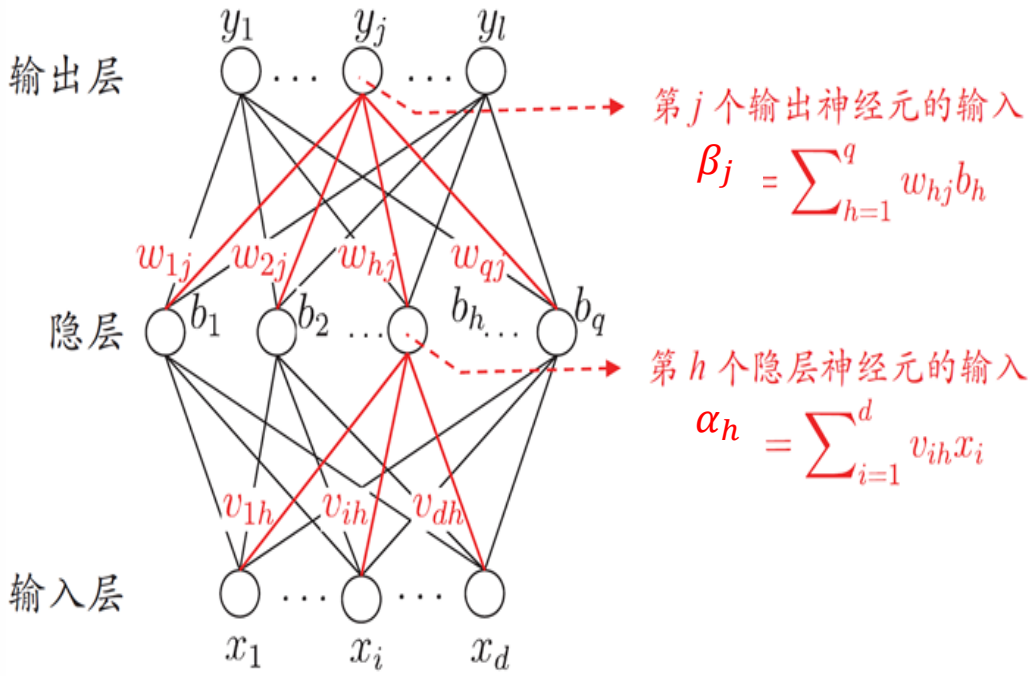
第4章：决策树

- 经典的属性划分方法：
  - 信息增益 ID3
  - 信息增益率 C4.5
  - 基尼指数 CART
- 剪枝处理：“过拟合”的解决手段
  - 预剪枝
  - 后剪枝
- **连续与缺失值**

第5章：神经网络

- 感知机与多层网络
- 多层前馈神经网络
- **误差逆传播算法（BP）：前向计算**
- **梯度下降优化**
- **缓解过拟合的策略**
- 全局最小与局部极小

权重：v_{ih}、w_{hj} **阈值**：θ_j、γ_h $(i=1,\cdots,d, h=1,,\cdots,q, j=1,\cdots,l)$



step1： $b_h = f(\alpha_h - \gamma_h), \alpha_h = \sum_{i=1}^{d} v_{ih} x_i$

step2： $\hat{y}_j^k = f(\beta_j - \theta_j), \beta_j = \sum_{h=1}^{q} w_{hj} b_h$

step3： $E_k = \frac{1}{2} \sum_{j=1}^{l} (\hat{y}_j^k - y_j^k)^2$

$f(x) = \text{sigmoid}(x) = \frac{1}{1+e^{-x}}$

第6章：支持向量机

- 支持向量机：间隔与支持向量
- 核函数：通过向高维空间映射解决线性不可分的问题
- 软间隔：允许支持向量机在一些样本上不满足约束

第7章：贝叶斯分类器

- 判别式模型：决策树，BP神经网络，支持向量机
  - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$，来预测 c
- 生成式模型：贝叶斯分类器
  - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
- **贝叶斯定理：** $P(c \mid \mathbf{x}) = \dfrac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})}$
- 极大似然估计
- **朴素贝叶斯分类器**
- EM算法：对模型中的隐变量做参数估计

假设吸烟的本科生比例为15%，而吸烟的研究生占23%。如果五分之一的大学生是研究生，其余的是本科生，那么随机抽取一名吸烟的学生，该生是本科生还是研究生的可能性大？

$P(\text{吸烟}|\text{本科生}) = 15$

$P(\text{吸烟}|\text{研究生}) = 23$

$P(\text{研究生}) = \frac{1}{5}$

$P(\text{本科生}) = \frac{4}{5}$

求： $P(\text{研究生}|\text{吸烟}) = ?$

根据朴素贝叶斯公式，有：

$$P(\text{研究生}|\text{吸烟}) = \frac{P(\text{吸烟}|\text{研究生}) * P(\text{研究生})}{P(\text{吸烟})}$$

$P(\text{吸烟}) = ?$

$$P(\text{吸烟}) = \frac{N(\text{本科生}and\text{吸烟}) + N(\text{研究生}and\text{吸烟})}{N(\text{本科生}) + N(\text{研究生})}$$

$$= \frac{P(\text{吸烟}|\text{本科生}) * N(\text{本科生}) + P(\text{吸烟}|\text{研究生}) * N(\text{研究生})}{N(\text{本科生}) + N(\text{研究生})}$$

假设全校共5人，根据 $P(\text{研究生}) = \frac{1}{5}$ ，可知
$N(\text{研究生}) = 1\text{人}$， $N(\text{本科生}) = 4\text{人}$。

所以， $P(\text{吸烟}) = \frac{15\% * 4 + 23\% * 1}{1 + 4} = 16.6\%$

最终， $P(\text{研究生}|\text{吸烟}) = \frac{23\% * \frac{1}{5}}{16.6\%} = 27.71\%$

第8章：集成学习

- 集成学习基本原理：好而不同
- **Boosting**
- **Bagging与随机森林（算法基本步骤等）**
- **结合策略**
- 多样性 – 多样性增强

第9章：聚类

- 聚类：无监督学习方法
- k均值算法（应用）