

# 《高等机器学习》课程大纲

研究生课程：3 学分/48 学时（面向研究生，本科生可选）

## 课程目标：

目前国内外高校所讲授的机器学习课程在学术前沿性和应用实战性方面都有所欠缺。清华电子系开有基础性机器学习课程，但尚缺一门全面、深入讲授机器学习前沿知识和主要应用专题的重头课程。基于以上背景，我们申请开设《高等机器学习》课程。

本课程以机器学习的概念、算法、理论为核心，结合机器学习近年来取得的最新进展及其在实际产业中的成功应用，通过课堂讲授、案例分析、面向解决实际问题的课程大作业、研讨会、答辩会等多元的形式，帮助学生对前沿的机器学习技术及其应用方法进行全方位掌握。该课程在技术前沿和产业落地方面，与国内外已有的机器学习课程相比都独树一帜、富有特色。

## 第一部分：机器学习基础（6 学时）

### 第一章 课程导论（1 学时）

- 1.1 机器学习的历史回顾
- 1.2 机器学习的发展现状
- 1.3 如何实现机器学习研究的闭环
- 1.4 课程预备知识简介：微积分、线性代数、最优化方法、概率论等

### 第二章 机器学习框架（4 学时）

## 2.1 机器学习的门类

- 有无监督信号：有监督学习，半监督学习，无监督学习，强化学习与弱监督学习
- 生成与判别问题：生成学习，判别学习

## 2.2 有监督判别学习框架

- 损失函数：回归，分类，排序/结构化预测
- 机器学习模型：线性模型与逻辑回归，支持向量机与核方法，决策树，神经网络等
- 优化方法：随机梯度下降法及其变种，遗传算法，基于离散搜索的优化方法
- 学习理论：优化理论、泛化理论、逼近理论。

## 第三章 机器学习流程（1 学时）

### 3.1 数据清洗与预处理

### 3.2 特征工程与表示学习

### 3.3 训练、交叉验证、测试过程

### 3.4 模型选择与超参调节

## 第二部分：高级机器学习模型(9 学时)

## 第四章 深层神经网络（6 学时）

### 4.1 常用的网络模型

- 深层前馈网络
- 深层卷积网络
- 深层循环网络
- 编解码器框架
- 注意力机制

### 4.2 神经网络的优化算法

- 自适应优化技术：Momentum, AdaGrad, AdaDelta, Adam 等
- 常用技巧：跨层连接，正则化技术，初始化，DropOut 等
- 寻找正确的优化空间：正尺度不变优化算法

#### 4.3 深度学习理论

- 尖锐度、表达能力和泛化性能
- 网络深度对表达能力和泛化性能的影响
- 过参数化对优化过程的影响
- 对跨层连接、批正则化等优化技巧的理论分析

### 第五章 梯度提升树（2 学时）

- 5.1 GBDT 的基本原理及其与神经网络的对比
- 5.2 XGBoost 算法
- 5.3 LightGBM 算法

### 第六章 生成模型（1 学时）

- 6.1 概率图模型与主题模型
- 6.2 生成对抗网络及其变种
- 6.3 贝叶斯深度学习

## 第三部分：高级机器学习范式（15 学时）

### 第七章 强化学习（4 学时）

- 7.1 强化学习基本概念
- 7.2 基于价值函数的优化算法：DQN/C51
- 7.3 策略优化算法：PG/A3C/PP0/TRPO
- 7.4 混合优化算法：DDPG/TD3/SAC
- 7.5 基于模型的强化学习算法
- 7.6 样本效率、探索与效用的平衡

## 第八章 对偶学习（2 学时）

- 8.1 无监督对偶学习
- 8.2 有监督对偶学习
- 8.3 对偶推断
- 8.4 对偶 GAN/循环 GAN
- 8.5 其他小样本学习和无监督学习技术

## 第九章 鲁棒机器学习（1 学时）

- 9.1 对抗攻击
- 9.2 对抗鲁棒性
- 9.3 鲁棒机器学习算法

## 第十章 轻量机器学习（3 学时）

- 10.1 模型压缩
- 10.2 知识蒸馏
- 10.3 LightLDA 算法
- 10.4 LightRNN 算法
- 10.5 LightGBM 算法
- 10.6 机器学习算法的软硬件一体化设计

## 第十一章 分布式机器学习（3 学时）

- 11.1 数据与模型划分
- 11.2 并行框架：MapReduce，参数服务器，数据流
- 11.3 同步与异步通信机制
- 11.4 模型与数据聚合
- 11.4 分布式学习理论

## 第十二章 元学习与教学相长（2 学时）

- 12.1 贝叶斯优化
- 12.2 学习如何学习 (Learning to learn)
- 12.3 神经网络结构优化
- 12.4 教学相长 (Learning to teach)

## 第四部分：机器学习实践 (12 学时)

### 第十三章 机器学习编程 (1 学时)

- 13.1 Python 编程基础
- 13.2 Spark, TensorFlow, PyTorch 工具包
- 13.3 OpenPAI 开放机器学习平台

### 第十四章 计算机视觉 (2 学时)

- 14.1 ResNet 及图像分类
- 14.2 Fast R-CNN 及物体分割

### 第十五章 自然语言处理 (3 学时)

- 15.1 预训练: Word2Vec, BERT 等
- 15.2 阅读理解
- 15.3 机器翻译
  - GNMT, ConvS2S, Transformer
  - 多语言机器翻译
  - 无监督机器翻译

### 第十六章 机器博弈 (3 学时)

- 16.1 围棋与 AlphaGo
- 16.2 德州扑克与 Libratus
- 16.3 星际争霸与 AlphaStar

## 16.4 麻将与 Suphx

## 第十七章 序列分析（3 学时）

### 17.1 供给需求分析

### 17.2 股票走势预测

### 17.3 基因与蛋白质序列分析

## 第五部分：课程总结（6 学时）

## 第十八章 课程大作业汇报会（5 学时）

- 同学自由组队（3 至 5 人为一队），结合第四部分给出应用场景，每队选择一个具体课题，设计相应的机器学习算法，并进行实现。实验过程使用天津研究院人工智能开放平台，和微软亚洲研究院的 OpenPAI 平台。
- 每队将设计思路、算法细节、实验结果及其分析，写成一篇论文格式的总结报告，并参加报告会。每个队约 15 分钟报告时间。论文和答辩占期末成绩的 50%，每部分结束后的综述论文占成绩的另 50%。

## 第十九章 机器学习的技术发展趋势（1 学时）

- 邀请机器学习业内知名学者针对机器学习的未来进行开放式研讨，并和同学们互动。