

## 麻省理工学院“机器学习与数据科学”线上科研项目

### 学校简介 Introduction

麻省理工学院 (Massachusetts Institute of Technology)，位于美国马萨诸塞州波士顿都市区剑桥市，主校区依查尔斯河而建，是世界著名私立研究型大学。麻省理工学院创立于 1861 年，早期侧重应用科学及工程学，在第二次世界大战后，麻省理工学院倚靠美国国防科技的研发需要而迅速崛起。在二战和冷战期间，麻省理工学院的研究人员对计算机、雷达以及惯性导航系统等科技发展作出了重要贡献。

2019-20 年度位列 QS 世界大学排名第一、U.S. News 世界大学排名第二、软科世界大学学术排名 (ARWU) 第四、泰晤士高等教育世界大学排名第五；2019-20 年度位列《泰晤士高等教育》世界大学声誉排名世界第二。2021 年位列 QS 世界大学排名第一。

### 项目背景 Program Background

人工智能与机器学习的兴起为未来生活建立了无数新的可能性。在“大数据”时代，海量的数据提供了丰富的建模信息。机器学习模型利用先进的算法提取数据信息进行预测与决策，极大的改变了人们日常生活。目前常用的人脸识别、广告推荐，到未来可能出现的无人驾驶汽车、智能机器人，都充满了机器学习的身影。

“人工智能”与“大数据”的完美结合将改变我们的日常生活，也即将成为各领域研究发展方向的变革工具。

项目将在来自计算机专业排名前列的麻省理工学院的终身教授和博士生的指导下进行，介绍常用机器学习和数据科学理论，以及当下最受欢迎的 Python 编程语言，引导学生学习机器学习理论和实际应用。

### 项目介绍 Program Description

学生将在项目中学习数据科学、机器学习的理论和方法，了解并且掌握 Python 在数据科学和机器学习中的应用。学生将在项目结束时，自选框架和问题，使用 Python 开发机器学习应用，提交项目报告，进行成果展示。

### 导师介绍 Instructor Introduction

麻省理工学院 (MIT) 终身教授

现任麻省理工学院 (MIT) 终身教授，曾获素有“诺贝尔风向标”美誉的美国斯隆研究奖、国际最具声望的博士后奖励 Hubble Fellow。教授的研究兴趣聚焦机器学习、数据科学、人工智能、天体物理，善于利用高性能超级计算机强大的数据处理能力进行数值模拟，训练机器学习和深度学习模型，借助机器学习与数据科学技术分析模拟数据。

### 项目大纲 Syllabus

机器学习理论与应用：

- 机器学习理论综述：理解机器学习的基本框架，监督学习（分类、回归）、无监督学习（聚类、密度估计、降维）、强化学习、序列学习（sequence learning）。理解常用概念：损失函数、训练与预测规则、模型类别与参数估计。介绍最简单的线性分类器。
- 特征表示（feature representation）：数据的特征，不同类别的数据处理方法（离散、连续、文字数据）、数据标准化。简单的数据分析技巧（最值、异常值、绘制分布），理解数据之间的相关性。了解数据降维与数据可视化。课堂编程进行数据分析。
- Logistic regression 和梯度下降算法：了解 logistic regression 的原理，包括目标函数，sigmoid function，分类原理，优化目标，L1 L2 penalty。理解梯度下降算法的原理及其在 logistic regression 中的应用，了解随机梯度下降（SGD）。课堂编程进行 logistic regression 的测试。
- 回归问题（linear regression）：理解线性回归的原理、线性回归的解析解。模型的可解释性。理解几个的正则项对回归问题的影响（ridge regression）。使用 python 进行课堂编程，学会解释线性回归模型。
- 神经网络：理解神经网络模型的原理，包括基本单元、层、激活函数、反向传播算法、模型训练、损失函数、模型参数调整、正则项、batch normalization。
- 前沿机器学习（与深度学习）模型：卷积神经网络的原理与计算、循环神经网络（Recurrent neural network）的原理与计算。
- 其他机器学习模型原理与实践：决策树、K-nearest neighbor、聚类算法、数据降维、主成分分析等。

#### 编程技巧与实战：

- Scikit-Learn：Scikit-Learn 是一个建立在 Scipy 基础上的用于机器学习的 Python 模块，包含众多顶级机器学习算法，主要有六大基本功能，分别是分类、回归、聚类、数据降维、模型选择和数据预处理，学生将了解 Scikit-Learn 的基础架构与应用
- Tensorflow：是一个基于数据流编程的符号数学系统，被广泛应用于各类机器学习算法的编程实现，学生将了解 Tensorflow 的基础架构与应用
- PyTorch：PyTorch 是一个开源的 Python 机器学习库，基于 Torch，用于自然语言处理等应用程序，学生将了解 Pytorch 的基础架构与应用
- Hadoop：Hadoop 是一个分布式系统基础架构，用户可以在不了解分布式底层细节的情况下，开发分布式程序，学生将了解 Hadoop 的基础架构与应用

#### 课程作业与结课项目（Group project）：

- 第一周：以小组（每组 3-4 人）为单位进行课程项目。
- 第二周：组队报名
- 第三周：提交项目计划（一页纸）
- 第四-五周：文献综述、数据处理
- 第六-九周：编程完成课程项目，准备需要的写作材料（如：绘图、表，方法论公式）
- 第十-十一周：论文（或结课报告）写作、海报和 PPT 制作。
- 第十二周：课堂展示、海报展示。
- Office hour：学生在项目进行过程中可与导师及时交流项目进展

**时间：2021年1月22日-3月06日（时间暂定）**

**费用：16800元/学生**

**专业定制：25-30人/班。**

**成果收获：**麻省理工学院教授推荐信、结业证书、成绩单、学术报告、8封网申推荐信（对名校申请起决定作用）、高质量的科研报告可指导发表国际论文或SCI

**咨询方式：**

李老师 手机：18362381882；QQ号：1814958113

**报名须知：**

1. 报名截止时间：机会难得，名额有限，先报先录、额满即止。
2. 面向对象：1.在校本科生、研究生；2.有相关专业基础

**报名步骤：**

第一步：扫码在线填写报名信息



第二步：等待面试及录取邮件通知，缴纳项目费用，签署项目协议

第三步：等待项目组开课通知