（联合开发：合作企业 Logo）

机器学习工程师纳米学位课程大纲

系统掌握监督学习、非监督学习、深度学习、工程部署等技能

更新日期 / 2019-09-24

开始之前

学习目标

机器学习是驱动人工智能领域突破性发展的核心技术。AlphaGo 战胜人类围棋冠军、人脸识别、大数据挖掘，都和机器学习密切相关。

这个课程将培养你成为一名机器学习工程师，学会如何建立模型，并将这些模型应用在各个领域的数据集中，如金融、社交产品、共享经济、医疗和教育等，帮助你脱颖而出成为行业渴望的稀缺人才。

先修知识

一定的 Python 编程能力，了解基础的概率论与统计、线性代数与微积分。如果不具备这些先修，建议学习“Python 人工智能入门”，如需技术能力测试可咨询学习规划师。

项目时长

26 周，建议每周学习时间 10 小时左右

阅读下文 “详细大纲”，了解更多课程信息。

课程大纲详细篇

第 1 部分：监督学习

课程内容

课程标题	学习目标
简介	在深入研究多种机器学习算法之前，我们先了解整个领域的全局知识
线性回归	了解分类与回归的区别，学习如何使用线性回归来做预测
感知器算法	学习神经网络中的感知器，以及如何使用它进行分类
决策树	学习决策树，并使用决策树探索泰坦尼克号乘客存活模型
朴素贝叶斯	学习朴素贝叶斯原理，并构建垃圾邮件分类器
支持向量机	学习如何训练支持向量机以线性分离数据； 使用核方法在非线性可分的数据上来训练 SVMs
集成方法	通过 boosting 提升传统方法；Adaboost
模型评估维度	学习评估模型的常用维度：准确率、精读、召回率等等。
错误与优化	了解训练过程中常见的错误类型，学习如何处理错误来优化模型性能。
Lab：为银行提供精准营销方案	练习机器学习基础技能，在实战中掌握预测的 pipeline。

实战项目：航班延误预测

在该实战项目中，你将把机器学习技术应用到航空领域中。航班延误一直是旅客们头疼的问题，台风，雾霾或飞机故障等因素都有可能导致大面积航班延误的情况。大面积的延误往往带来很多不便，你将分析携程与飞常准提供的历史航班动态起降数据、历史城市天气、机场城市与特情相关数据集，在计划起飞前 2 小时预测航班延误情况，从

而帮助出行的旅客更好地规划行程。在完成这次的项目过程中，你将会学到：处理实际问题中噪声数据的技能；tabular数据特征工程技术；lightgbm、xgboost 等决策树模型的构建技术；模型调参技术。

第 2 部分：深度学习

课程内容

课程标题	学习目标
神经网络简介	学习深度学习基础，包括 softmax、one-hot encoding 和 cross entropy 学习感知器与梯度下降
实现梯度下降	了解如何实现梯度下降，并实现一个反向传播
训练神经网络	学习如何训练神经网络，包括早期停止、正则化、dropout等知识
PyTorch	学习如何使用 PyTorch 构建神经网络，并在 Fashion-MNIST 数据集上学习图像分类与迁移学习。

实战项目：人脸识别

在这个项目中，你将尝试完成计算机视觉领域中的热门专题——人脸识别。具体来说，训练一个能够将不同的人准确识别出具体是谁的模型。在这个任务中，你将会学习深度学习框架——Keras，并将此用于人脸识别任务实践中；同时，你将会学到：卷积神经网络（CNN），CNN高级结构比如VGG、ResNet等；另外，还将学习数据增强技术在cv中的应用。

第 3 部分：非监督学习

课程内容

课程标题	学习目标
聚类	学习如何聚类算法，并尝试使用 k-means 对数据进行聚类
层次聚类法与密度聚类	学习单连接聚类法和层次聚类法，DBSCAN

高斯混合模型与聚类验证	学习高斯混合模型及相关示例，以及聚类分析过程和如何验证聚类结果。
降维和PCA	了解降维的作用，并学习 PCA 的原理和使用场景
随机投影与 ICA	学习随机投影与独立成分分析，并通过 Lab 学习如何应用这些方法

实战项目：创建客户细分

在此项目中，你将针对葡萄牙里斯本批发商的客户收集的产品支出数据应用非监督学习技术，以确定隐藏在数据中的客户群。你将首先通过选择一个小样本子集来探索数据，并确定是否有任何产品类别彼此高度相关。之后，你将通过缩放每个产品类别来预处理数据，然后识别（并移除）不需要的异常值。借助良好，干净的客户支出数据，你将对数据应用PCA转换，并使用聚类算法以对转换后的客户数据进行细分。最后，你将比较发现的细分和附加标签，并考虑这些信息是否可以帮助批发商进行未来服务变更。

第 4 部分：软件工程

课程内容

课程标题	学习目标
软件工程练习	编写清晰、注释充分的模块化代码 重构代码并提高代码效率 创建检验程序的单元测试 通过日志记录进程操作和结果 审阅代码
编程	了解何时使用面向对象编程 构建和使用类 了解如何创建大型模块化 Python 软件包并使用面向对象编程
将软件包上传到 PyPI	作品集练习：构建你自己的 Python 软件包

案例演练：构建 Python 软件包

在之后的几个章节中，你将学习如何构建机器学习算法并使这些算法能够进入可拓展的生产系统。构建这些系统的第一步是了解如何编写生产级代码，你可以选择编写一个你自己的 Python 软件包。在此 Lab 中，你将练习到以下技能：面向对象编程、整洁的模块化代码、代码文档注释。

第 5 部分：模型部署

课程内容

课程标题	学习目标
部署简介	了解云端和部署术语 了解生产环境中的机器学习工作流程 了解机器学习的工作场所用例
部署模型	在 SageMaker 中部署模型 使用 SageMaker 上的 XGBoost 预测波士顿房价 使用 SageMaker 上的 XGBoost 判断影评情感
网络托管	学习如何从网站提供端点访问权限； 使用 API Gateway 和 Lambda 将机器学习模型集成到网络应用中；
模型监控	了解如何监控模型随时间推移的行为； 使用 SageMaker 的自动化超参数调节工具调节 XGBoost 模型的超参数； 在 SageMaker 上运行 A/B 测试，比较调节过的模型与未调节的模型；
更新模型	在监控模型的过程中发现数据有变化后，相应地更新模型； 了解如何处理在情感分析过程中添加到模型中的新短语；

实战项目：部署情感分析模型

在此项目中，你将利用递归神经网络判断影评的情感，使用的数据集包含标为正面或负面情感的文本影评。你将使用 Amazon SageMaker 创建和部署此模型。部署模型之后，你将创建一个简单网络应用，该应用会与部署的模型 交互并对新输入的影评进行分类。

第 6 部分：机器学习案例研究

课程内容

课程标题	学习目标
利用 SageMaker 进行总体分割	使用 AWS SageMaker 了解可用的算法广度； 了解如何通过 SageMaker 使用非监督式算法分析数据； 使用 SageMaker 部署非监督式模型； 通过提取模型属性了解数据；
检测信用卡欺诈行为	构建并改善能识别付款欺诈行为的线性模型； 处理训练数据类别不平衡的问题； 在 SageMaker 中根据特定指标调节模型并改善模型性能；
部署自定义模型	使用 SageMaker 部署自定义 PyTorch 模型； 编写自定义训练脚本，并训练你设计的模型；
时间序列预测	处理时间序列数据并调整数据格式，使数据适合训练机器学习模型； 使用 SageMaker 的DeepAR 算法进行时间序列预测； 部署模型并使用模型预测未来的数据点；

实战项目：抄袭检测器

运用机器学习技能比较两个文本来源，并判断是否存在剽窃行为。在此项目中，你将提取相关的文本特征并训练你 自己设计的模型来检测剽窃行为。然后，你将使用 Amazon SageMaker 部署你训练的模型。

实战项目：毕业项目开题报告

在正式开始毕业项目之前，你需要撰写一份开题报告来阐述自己的建模过程，并得到专业审阅者的建议。

实战项目：毕业项目（三选一）

在毕业项目中，你将运用你在此纳米学位中学到的机器学习算法和方法选择一个你感兴趣的问题来解决。首先你需要定义 (define) 一个你想要解决的问题，调研可能的解决方案，并解释其衡量指标。其次，你需要通过数据可视化和数据挖掘分析 (analyze) 这个问题，以便对适于解决该问题的算法和特征有一个更好的了解。接着，你需要实现 (implement)你的算法和衡量指标。你需要记录下预处理、改善、后处理的整个过程。接下来，你需要收集这些模型的表现结果 (results)，把重要的部分可视化，来验证或证明这些结果。最后，你要在你的结果基础上得出结论 (conclusions)，讨论一下你的实现是否真的解决了这个问题。

自然语言处理方向

项目1 - 句子相似度匹配

Quora Querstion Pairs数据集是Quora于2017年公开的句子匹配数据集，其通过给定两个句子的一致性标签标注，从而来判断句子是否一致。

数据挖掘方向

项目2 - 预测 Rossmann 未来的销售额

Rossmann 是欧洲的一家连锁药店。 在这个源自Kaggle比赛 Rossmann Store Sales，我们需要根据 Rossmann 药妆店的信息（比如促销，竞争对手，节假日）以及在过去的销售情况，来预测它未来的销售额。

计算机视觉方向

项目3 - 猫狗大战

使用深度学习方法识别一张图片是猫还是狗。

注：由于技术的快速迭代，中国区教研团队将根据当前热点对毕业项目选题进行实时更新，请以教室中的选题为准。

先修知识：Python 与数据分析

课程内容

课程标题	学习目标
数据类型和运算符	你将学习 Python 中用到的所有数据类型和运算符，如字符串、列表、元组、集合、字典等；
控制流	学习如何通过控制流为你的程序创建逻辑；
函数	你将学习如何定义函数。还将学习如何将程序拆分为多个部分，使得代码的结构更加合理。
脚本编写	了解如何编写脚本，学习用来开发程序的不同环境，有助于之后与他人合作编程；
NumPy	在这节课，你将学习 NumPy 基本知识，以及如何使用 NumPy 创建和操作数组。
Pandas	在这节课，你将学习 Pandas Series 和 DataFrame 基本知识，以及如何使用它们加载和处理数据。

先修知识：SQL 与数据分析

课程内容

课程标题	学习目标
基本 SQL	学习在单个表中使用 SQL 的基础知识；学习以许多不同方式过滤表的关键命令；
SQL JOIN	学习如何将多个表格中的数据组合到一起；
SQL 聚合	学习如何使用 SUM、AVG 和 COUNT 等 SQL 函数整合数据。此外，CASE、HAVING 和 DATE 函数是非常强大的问题解决工具；

SQL 子查询和临时表格	学习使用子查询来回答更加负责的商业问题；
SQL 数据清理	数据清理是数据分析的重要步骤，你在本课中将会学到如何使用 SQL 进行数据清理；
窗口函数	学习强大的数据分析工具——窗口函数；
SQL 全连接与性能优化	学习 SQL 的高级用法，了解如何在大数据上使用查询语句；

先修知识：命令行

课程内容

课程标题	学习目标
Shell 讲习班	Unix shell 对所有领域的开发工程师来说都是一款强大的工具。在这节课，我们将快速讲解下在计算机上使用该工具的基本知识。

先修知识：使用 Git 和 GitHub 进行版本控制

课程内容

课程标题	学习目标
使用 Git 和 GitHub 进行版本控制	了解版本控制的优势并安装 Git； 学习如何创建仓库和其他常用操作； 学习如何利用 git 的分支实现隔离开发过程； 学习如何在 GitHub 上创建远程仓库，以及如何获取和推送对远程仓库的更改；

选修内容：Python 与数据可视化

课程内容

课程标题	学习目标
数据分析中的数据可视化	了解数据可视化成为数据分析重要组成部分的原因以及它的适用范围。
可视化的设计	了解可视化设计的元素，尤其要避免使用那些可能导致不良可视化的元素；
单变量数据探索	学习如何利用 matplotlib 和 seaborn 可视化单变量数据；
双变量数据探索	学习使用 matplotlib 和 seaborn 探索双变量数据，根据对变量的理解，构建变量之间的关系；
多变量数据探索	学习如何使用 Matplotlb 和 Seaborn 可视化多个变量（三个或三个以上）之间的关系；
解释性数据可视化	学习如何修饰图表让你的信息得到精准地传达；
可视化案例分析	运用之前学到的关于解释性和探索性可视化的技巧，来探索钻石价格的影响因素；
补充主题	在本附加课程中，你可以阅读有关本课程的其他可视化方法和图类型的信息。

选修内容：应用统计学

课程内容

课程标题	学习目标
描述统计学	学习数据类型，中心度量和统计表达式的基础知识； 了解与定量数据相关的离散程度测量，形状和异常值，并学习了解推论统计。

录取案例分析	学习辛普森悖论，学会提正确的问题。
概率	利用硬币和骰子获得概率基础知识。
二项分布	学习概率中最流行的分布之一：二项分布。
条件概率	并不是所有事件都是独立的。学习相关事件的概率规则。
贝叶斯规则	学习概率中最流行的一种规则：贝叶斯规则。
Python 概率练习	利用上节课所学知识，应用到 Python 实践中。
正态分布理论	学习从掷硬币到正态分布背后的数学知识。
抽样分布与中心极限定理	学习置信区间和假设检验的基础：抽样分布。

选修内容：线性代数

课程内容

课程标题	学习目标
简介	简要了解精彩的线性代数以及为何它是一个很重要的数学工具。
向量	了解线性代数的基本概念——向量。
线性组合	了解如何伸缩向量和将向量相加，以及如何可视化求解过程。
线性变换和矩阵	什么是线性变换，它与矩阵有何直接联系？你将学习如何运用数学知识并可视化这些概念。

选修内容：NLP 基础

课程内容

课程标题	学习目标

NLP 简介	了解自然语言模型中文本的表示方式； 使用“Bag-of-Words”、“ TF-IDF”、“ Word2Vec”和“ GloVE”之类的方法来转换文本。
Pytorch 实现 RNN 和 LSTM	学习如何用代码表示记忆功能。然后在 PyTorch 中定义和训练 RNN 并将它们用于处理序列数据。
在线Lab：Pytorch实现情感分析 RNN	实现一个判定影评是正面还是负面影评的情感分析 RNN。
线性变换和矩阵	什么是线性变换，它与矩阵有何直接联系？你将学习如何运用数学知识并可视化这些概念。

选修内容：卷积神经网络

课程内容

课程标题	学习目标
卷积神经网络	卷积神经网络可以识别空间图案。Alexis 和 Cezanne 将介绍卷积神经网络如何帮助我们显著提高图像分类的效果。
迁移学习	学习如何通过迁移学习将预训练的网络应用到新任务上。
权重初始化	在这节课，你将学习如何为神经网络设置合适的初始权重。合适的初始权重使神经网络更接近最佳模型
自动编码器	自动编码器是用于数据压缩，图像降噪和降维的神经网络。在这里，你将使用 PyTorch 构建自动编码器。

选修内容：Flask 网页部署

课程内容

课程标题	学习目标
网页部署	使用Flask，Bootstrap，Plotly和Pandas开发数据仪表板。

项目组合练习：部署数据仪表板	自定义上一课的数据仪表板，并将仪表板上传到网络。