机器学习 (L01)：导论

授课教师: 王贝伦 / 助教: 张嘉琦，周嘉莹，黄旭

1 课程介绍和考核要求

1.1 课程介绍

教师: 王贝伦
英语教材
联系方式:

- E-mail:beilun@seu.edu.cn
- QQ 群

1.2 考核要求

分数组成:

平时分 10%（网课采用随机数点名回答上次课相关问题）
作业 10%
期末考试 60%
实验 20%（编程作业 6-8 次）
附加分 <=5 分

额外原则:

缺课两次不及格
期末考试满分 = 总分满分

2 机器学习基础概念介绍

2.1 背景

数据无处不在

- 生物医学
  - 患者病例，脑成像，核磁共振扫描
  - 基因组序列，生物结构，药物效应信息
- 科学
  - 历史文献，书籍扫描，天文学数据库，环境数据，气候记录
- 社交媒体
  - 社交互动记录，推特脸书记录，在线评论
- 娱乐
  - 互联网图像，好莱坞电影，音频文件

大数据时代的机遇和挑战

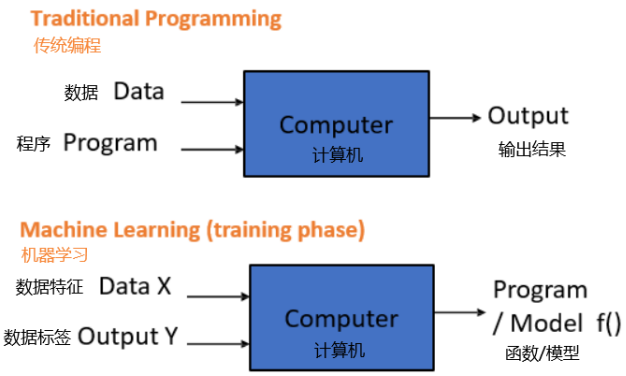
- 机遇
  - 办公效率提升
  - 科研突破
  - 生活质量提升：医疗，能源节约/生产，环境，养老，交通运输
- 挑战
  - 数据捕获（传感器，智能设备，医疗仪器等）
  - 数据传输
  - 数据存储（云计算）
  - 高性能数据处理

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

- 数据可视化（如，人机交互）
- 数据安全和隐私（如，多人情况）
- ★ 数据分析（本课内容）
  - 如何对大数据进行分析？——**机器学习**
  - 机器学习的应用——数据挖掘，语音识别，图像识别，文本分析，控制学习...

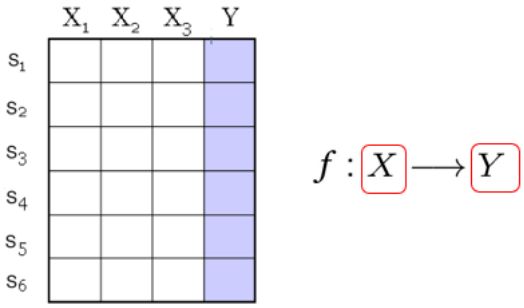
2.2 基本概念

- 什么是机器学习
  - 目标：构建一个能够从过去经验中学习和适应环境的计算机系统
  - 途径："经验"来源于**数据样本**（数据集）
  - 数据样本：在特征空间 X(feature space) 上的数据点 (data points)
  - 例：监督学习
    - 目标：从给定的训练数据集中学习出一个函数（模型参数）$f(\mathrm{x})$，使其预测值 $f(\mathrm{x})$ 和真实值 y 之间的差异尽可能小
    - 例 1:
      输入 (Input) x: 顾客对饭店的评价：虽然饭店的位置有点偏，但是菜品和好吃，老板也很热情.
      输出 (Output) y: 文章语义为褒义或贬义？$y \in \{1/Yes, -1/No\}$
    - 例 2:
      输入 (Input) x: 二维数据：图片上的一个点（点的颜色与位置有关）.$\mathrm{x} = (x_1, x_2)$
      输出 (Output) y: 这个点是红色或蓝色.$y \in \{1, 0\}$
      函数 (Function) f: 线性函数：划分红色点集和蓝色点集的线段或平面 $f(\mathrm{x}, \mathrm{w}, b) = sign(\mathrm{w}^\top \mathrm{x} + b)$
- 传统编程和机器学习的流程对比



传统计算机工作：给定数据和指令，计算机执行指令输出结果。
机器学习：给定数据和结果，计算机从中学习一个模型参数。

- 基本概念



上图为数据集的一个抽象结构图。其中每一行 s_i 表示一个样本，前三列 $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3$ 表示样本的特征，最后一列 y 表示样本的标签，f 表示从 x 到 y 的函数

- 样本 (data)：数据集中的一条记录，或者说一个对象。
  - 例：数据集为班级，样本为班里的一位同学
  - 上图每一行代表一个样本 s_i
- 特征 (features)：对象的属性。
  - 例：同学的身高、体重。

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

  - 上图 $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3$ 表示样本的三个属性
- 标签 (Target)：训练集样本的结果信息。
  - 标签可以是离散型的，例如性别，也可以是连续的，如体重。
- 损失函数 (Loss function): 反映预测结果和实际结果之间的差别
  - 例：二分类使用的 hinge loss

$$\sum_{i=1}^{L} \mathcal{L}(f(\mathrm{x}_i), y_i) = \sum_{i=1}^{L} \max(0, 1 - y_i f(\mathrm{x}_i))$$

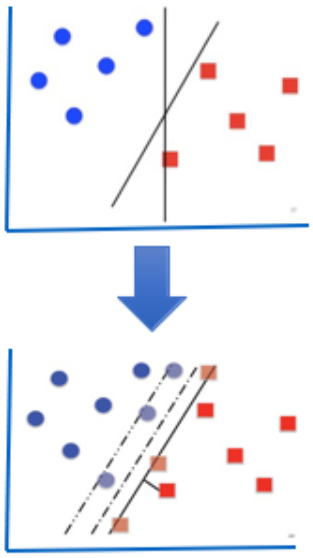
  - 例：排序使用的 pairwise ranking loss
- 训练 (Training)：学习模型参数 w, b
  - 训练集：
    样本 (sample):$\mathrm{x}_1, \mathrm{x}_2 ... \mathrm{x}_L$
    对应标签 (labels):$y_1, y_2 ... y_L$
  - 通过最小化损失 (minimize loss) 来学习参数 (w, b)

$$(\mathrm{w}, b) = \arg\min_{w,b} \sum_{i=1}^{L} \mathcal{L}(f(\mathrm{x}_i), y_i)$$

    其中 $\mathcal{L}$ 为特定的损失函数，用来衡量预测值 $f(\mathrm{x})$ 与真实值 y 之间的误差。w 为权重，类似于直线的斜率，用来衡量每个样本对目标函数（模型）的贡献。b 为偏置值，类似于直线的截距。只要这两个特征确定，那么 $f(\mathrm{x})$ 也就确定了。所以权重 w 与偏置 b 是 $f(\mathrm{x})$ 分布的关键特征。
    数据集总体的损失等于各个样本的损失之和。
- 测试 (Testing)：评估当前模型下特征点（对模型参数影响最大的点或靠近边界的点）的表现
  - 测试集中预测值 $(f(\mathrm{x}))$ 和真实值 (y) 的偏差
  - 测试集中的样本不在训练集中出现
- 泛化 (Generalisation)：学习模型来预测未知数据
- 正则化 (Regularization)：减小测试误差，避免过拟合（在训练数据上能够获得比其他假设更好的拟合，但是在训练数据外的数据集上却不能很好地拟合数据）
  - 例：在损失函数后面添加一个系数的“惩罚项”是正则化的常用方式

$$C \sum_{i=1}^{L} \mathcal{L}(f(\mathrm{x}_i), y_i) + \frac{1}{2}\|\mathrm{w}\|^2$$

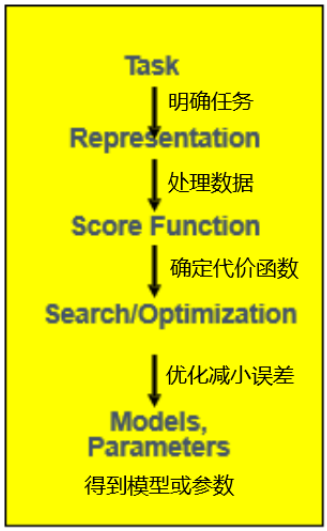
($\frac{1}{2}\|\mathrm{w}\|^2$ 为“惩罚项”)



(正则化后拟合直线的斜率被限制)

- 总结——机器学习流程导图

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。



3 机器学习发展史

3.1 发展历程

- 1950s:
  1956 年，在达特茅斯学院举行的一次会议上，不同领域（数学，心理学，工程学，经济学和政治学）的科学家正式确立了人工智能为研究学科。
  - Arthur Samuel 在五十年代中期和六十年代初开发的国际象棋程序，棋力已经可以挑战具有相当水平的业余爱好者
- 1960s:
  达特茅斯会议之后是大发现的时代。对很多人来讲，这一阶段开发出来的程序堪称神奇：计算机可以解决代数应用题、证明几何定理、学习和使用英语。在众多研究当中，搜索式推理、自然语言、微世界在当时最具影响力。
  - 感知器（深度学习雏形）被提出，由 Rosenblatt 提出会，是神经网络和支持向量机的基础。感知机接收多个输入信号，输出一个信号。
- 1970s:
  70 年代初，AI 遭遇到瓶颈。研究学者逐渐发现，虽然机器拥有了简单的逻辑推理能力，但遭遇到当时无法克服的基础性障碍，AI 停留在“玩具”阶段止步不前，远远达不到曾经预言的完全智能。
  - 当时主要问题：
    1. 计算机运算能力遭遇瓶颈，无法解决指数型爆炸的复杂计算问题
    2. 常识和推理需要大量对世界的认识信息，计算机达不到“看懂”和“听懂”的地步
    3. 无法解决莫拉维克悖论
    4. 无法解决部分涉及自动规划的逻辑问题
    5. 神经网络研究学者遭遇冷落
  - 一类名为“专家系统”的 AI 程序开始为全世界的公司所采纳，人工智能研究迎来了新一轮高潮。
    专家系统是一种程序，能够依据一组从专门知识中推演出的逻辑规则在某一特定领域回答或解决问题。由于专家系统仅限于一个很小的领域，从而避免了常识问题。“知识处理”随之也成为了主流 AI 研究的焦点。
- 1980s:
  - 最早的决策树算法是由 Hunt 等人于 1966 年提出，Hunt 算法是许多决策树算法的基础，包括 ID3、C4.5 和 CART 等
  - 1986 年，由认知心理学家 McClelland 和 Rumellhart 在神经网络训练中引入了 (Delta)Δ 学习规则，该规则亦可称为连续感知器学习规则（与离散感知器学习规则相并行）
  - PCA 理论出现，是一种多变量统计方法。通过析取主成分显出最大的个别差异，也用来削减回归分析和聚类分析中变量的数目
- 1990s:
  - 1989 年 8 月召开的第 11 届国际人工智能联合会议的专题讨论会上首次出现了知识发现 (Knowledge Discovery in Database, 以下简写为 KDD)，它泛指所有从源数据中发掘模式或联系的方法。KDD 描述了整个数据发掘的过程，包括最开始的制定业务目标到最终的结果分析，而数据挖掘 (data mining) 描述使用挖掘算法进行数据挖掘的子过程。最近人们逐渐习惯于使用数据挖掘来涵盖整个过程。
  - 神经网络与强化学习的结合，即深度强化学习，为强化学习带来了真正的机会，强化学习取得进展
- 2000s:
  - 1995 年，Corinna Cortes 和 Vladimir Vapnik 发表了他们在支持向量机上的工作。

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

  - 在 1995 年诞生了两种经典的算法-SVM 和 AdaBoost。SVM 代表了核技术的胜利，这是一种思想，通过隐式的将输入向量映射到高维空间中，使得原本非线性的问题能得到很好的处理.
  - 概率图模型是机器学习算法中独特的一个分支，它是图与概率论的完美结合。在这种模型中，每个节点表示随机变量，边则表示概率。

- 2010s:
  在摩尔定律下，计算机性能不断突破。云计算、大数据、机器学习、自然语言和机器视觉等领域发展迅速，人工智能迎来第三次高潮。
  - 语音识别和翻译兴起，2011 年苹果发布语音个人助手 Siri，2012 年 Google 发布个人助理 Google Now
  - 2014 年百度发布 Deep Speech 语音识别系统
  - 2017 年 AlphaGO 在围棋网络对战平台以 60 连胜击败世界各地高手
  - 2017 年 Google 发布了 ARCore SDK

3.2 相关领域

人工智能 (Artificial Intelligence):研究机器模拟人类的学习活动、获取知识和技能的理论和方法，以改善系统性能

数据挖掘 (Data Mining): 数据降维、数据筛选

概率和统计 (Probability and Statistics): 概率图模型、贝叶斯网

信息理论 (Information theory): 交叉熵损失、信息增益

数值优化 (Numerical optimization): 梯度下降、拉格朗日数乘

计算复杂性理论 (Computational complexity theory)：量子机器学习

控制理论 (自适应)(Control theory (adaptive)): 自动驾驶、机器人

心理学 (发育、认知)(Psychology (developmental, cognitive))：类脑研究、智能机心理学

神经生物学 (Neurobiology)：计算神经科学

语言学 (Linguistics)：自然语言处理

哲学 (Philosophy)

3.3 人工智能研究的目标

人工智能即包含理论研究的内容又包含工程方面的内容。人工智能的研究更注重系统的效果而不是单纯的对人的智能行为的模拟。

- 根本目标：要求计算机不仅能模拟而且可以延伸，扩展人的智能的理论，达到甚至超过人类智能的水平。
- 近期目标，使现有的计算机不仅能做一般的数值计算及非数值信息的数据处理，而且能运用知识处理问题，能模拟人类的部分职能行为。
- 作为工程技术学科，人工智能的目标是提出建造人工智能系统的新技术、新方法和新理论，并在此基础上研制出具有智能行为的计算机系统。
- 作为理论研究学科，人工智能的目标是提出能够描述和解释智能行为的概念与理论，为建立人工智能系统提供理论依据。

3.4 人工智能研究取得进展的原因

- 海量数据

  随着人工智能进入应用时代，数据的应用量得到了大幅提升。数据可以被视为支撑人工智能运行的原材料。总的来说，一个由合格的普通研究人员设计的、以大量训练数据为基础的算法，将胜过一个由最优秀的人工智能科学家设计的但用较少数据进行训练的算法。

- 适合深度神经网络 (Deep Neural Networks，以下简写为 DNN) 的计算架构

  未来 AI 芯片或传感器系统（如计算机视觉、雷达或光达）供货商不仅提供硬件，而且还会提供自己的高速、高效的 DNN——为应用而设计的深度神经网络架构。任何供货商都会为不同的运算平台匹配各自所需的网络架构。

- 强大的 DNN 平台/库

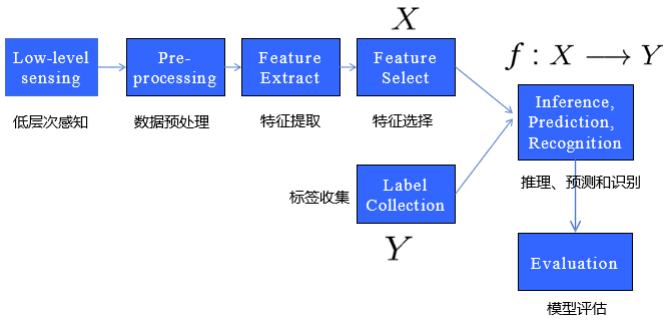
  如 pytorch、tensorflow、keras 等深度学习框架可以快速搭建神经网络。

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

4 机器学习主要方向

4.1 机器学习研究方向介绍

4.1.1 机器学习的一般流程



- 低层次感知 (Low-level sensing): 如通过传感器等（例如 CMOS）来获得数据。
- 预处理 (Preprocessing): 将原始数据标准化和规范化，以进行进一步研究，包括处理缺失值、处理偏离值、数据规范化、数据的转换等。
- 特征提取 (Feature extract): 主要是通过属性间的关系，如组合不同的属性得到新的属性，这样就改变了原来的特征空间，从而减少特征数据集中的属性 (或者称为特征) 的数目。
- 特征选择 (Feature select): 从原始特征数据集中选择出有决定性影响的子集，从而对原始数据降维预处理、特征提取和特征选择，概括起来就是特征表达。良好的特征表达，对最终算法的准确性起了非常关键的作用。
- 推理、预测和识别 (Inference,prediction,recognition): 机器学习的部分，绝大部分的工作是在这方面做的，包括模型的选择与训练。
- 评估 (Evaluation): 客观地评价模型的预测能力，模型评价阶段不做参数调整。可以根据分类、回归、排序等不同问题关心的问题选择不同的评价指标来评估模型的优缺点。

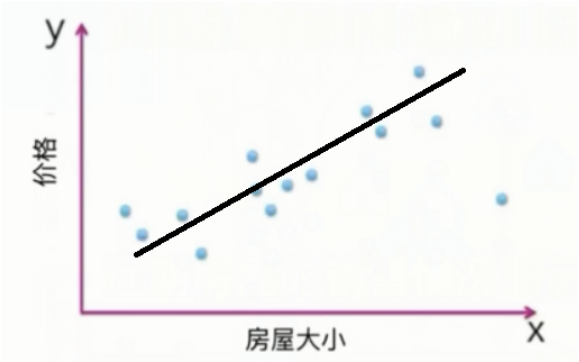
4.1.2 有监督学习 (Supervised Learning)

有监督学习，即人工给定一组数据，每个数据的属性值也给出，对于数据集中的每个样本，我们想要算法预测并给出正确答案。例如，回归问题，分类问题。

- 回归 (regression)：y 是实数 vector。回归问题，就是拟合 (x,y) 的一条曲线，使得损失函数 $\mathcal{L}$ 最小。

  例：根据楼房的大小、地段等预测房价。

  每个楼房代表一个样本 (sample)，输入值 (Input) 为样本特征 (feature)，例如大小、所处地段，输出值 (output) 为房屋价格，为实数。



  通过计算平均预测误差 (即最小二乘法) 作为损失函数来衡量回归模型的优劣。损失函数为：

$$J(\theta) = \frac{1}{2m}\sum_{i=0}^{m}(h_\theta(\mathrm{x}_i) - y)^2$$

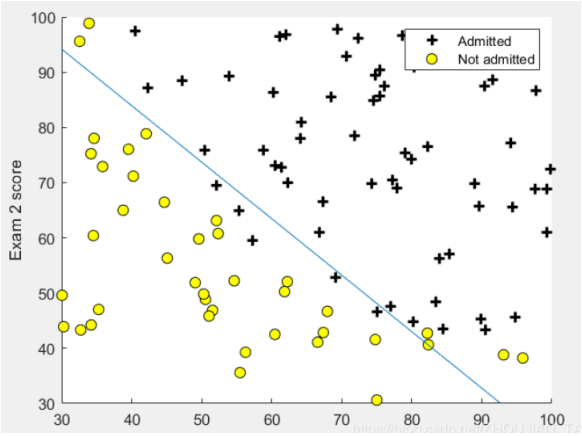
  其中 m 为样本数，h 为模型函数，θ 为要求解的参数，x_i 为样本，$h_\theta(\mathrm{x}_i)$ 为预测房价，y 为真实房价

  损失函数越小，说明回归模型越好。

- 分类 (classification)：与回归原理类似，不同点为 y 值属于一个有限集合 r，可以看做类标号。

  例：使用合适的决策边界（直线）将坐标平面划分为两个区域，使得坐标平面上的点尽可能按照颜色分类到同一个区域

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

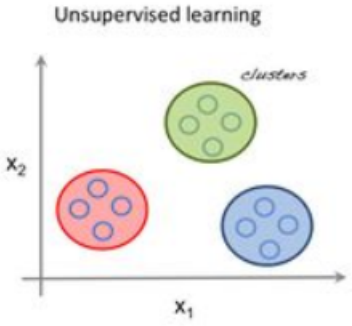


每个点代表一个样本，输入值为点的坐标 (x_i, x_k)，输出值为点的颜色 $y \in \{0, 1\}$。$y = 1$ 为黑色，$y = 0$ 为黄色。

4.1.3 **无监督学习** (Unsupervised Learning)

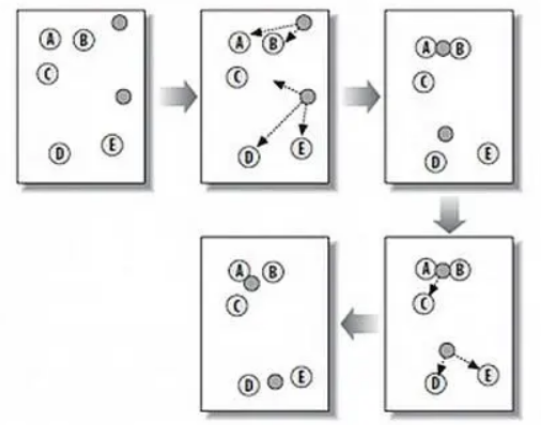
无监督学习也称作聚类学习，在无监督学习中，数据并不被特别标识，学习模型是为了推断出数据的一些内在结构。常见的应用场景包括关联规则的学习以及聚类等。常见算法包括 Apriori 算法以及 K-Means 算法。

- 例：将点集按照位置分成数类



可以采用 K-Means 方法。

- 1. 随机在图中取 K（这里 $K = 2$）个种子点。
  2. 对图中的所有点求到这 K 个种子点的距离，假如点 P_i 离种子点 S_i 最近，那么 P_i 属于 S_i 点群。（上图中，我们可以看到 A B 属于上面的种子点，C D E 属于下面中部的种子点）
  3. 移动种子点到属于他的“点群”的中心
  4. 重复第 2 和第 3 步，直到，种子点不再移动



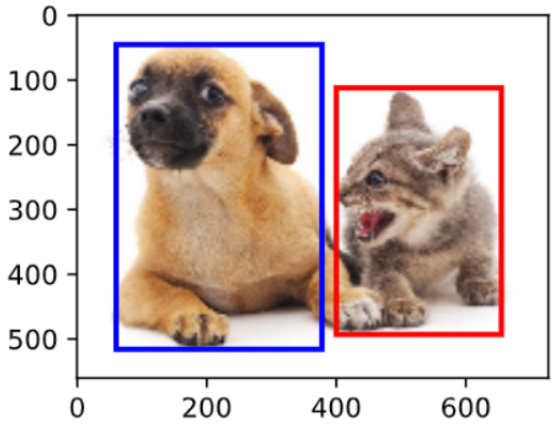
4.1.4 **结构化输入/输出学习** (Structural Output/Input Learning)

结构学习就是输入或输出是有结构的数据，比如说语句、列表、树和边界框（bounding box），而通常的网络学习之中，输入和输出都是向量。

在结构学习中，我们需要学习的是一个函数 f，它的输入是一种形式，输出是另外一种形式。例如，输入语音，输出对应的文本；输入中文，输出英文；输入图像，输出边界框（用来捕捉物体位置），等等。

- 例：目标检测。输入一张图片，输出包含检测到的物体的边界框

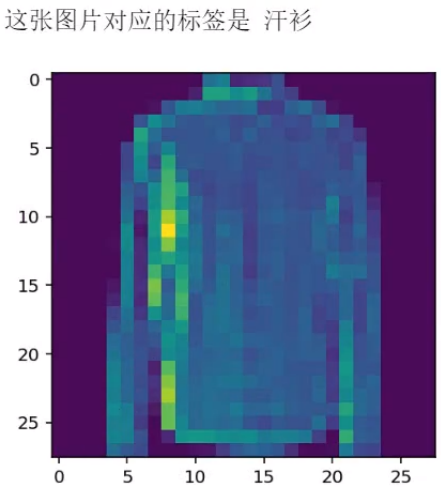
© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。



4.1.5 **深度学习** (Deep Learning)

深度学习使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象。深度学习通过层次化的结构对输入信息进行逐层提取和筛选，从而从多个角度学习事物的特征。深度学习是机器学习中一种基于对数据进行表征学习的方法

- 例：神经网络通过学习 Fashion-MNIST 时尚物品数据集完成对物品图像的识别



深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征。

4.1.6 **强化学习** (Reinforcement Learning)

强化学习，用于解决智能体（Agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题

强化学习主要由智能体（Agent）、环境（Environment）、状态（State）、动作（Action）、奖励（Reward）组成。智能体执行了某个动作后，环境将会转换到一个新的状态，对于该新的状态环境会给出奖励信号（正奖励或者负奖励）。随后，智能体根据新的状态和环境反馈的奖励，按照一定的策略执行新的动作。上述过程为智能体和环境通过状态、动作、奖励进行交互的方式。

强化学习模式下，输入数据作为对模型的反馈，不像监督模型那样，输入数据仅仅是作为一个检查模型对错的方式。在强化学习下，输入数据直接反馈到模型，模型必须对此立刻作出调整。例如，动态系统以及机器人控制等。常见算法包括 Q-Learning 以及时间差学习 (Temporal difference learning)

- 例：微软 AI 强化学习算法进行吃豆人游戏

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

深度强化学习将深度学习的感知能力和强化学习的决策能力相结合，可以直接根据输入的图像进行控制，是一种更接近人类思维方式的人工智能方法。

4.1.7 大规模机器学习 (Large-Scale Machine Learning)

获取高性能的机器学习系统途径是采用低偏差的学习算法，并用大数据进行训练。当使用规模庞大且高频变化的特征和样本作为训练集，传统的算法往往因为代价过大而无法使用。所以大规模机器学习中，希望找到能够用于大数据上的替代算法或者更有效的方法。

- 例：淘宝每天就有 100 亿规模的用户行为数据
- 例：2017 年双 11 购物狂欢节当天，小时级 XNN 模型在猜你喜欢和天猫推荐

4.2 课程安排

- 回归 (supervised)
- 分类 (supervised)
- 无监督学习
- 学习理论
- 概率图模型
- 强化学习

引用

[1] Prof. Andrew Moore's tutorials.

[2] Prof. Raymond J. Mooney's slides

[3] Prof. Alexander Gray's slides

[4] Prof. Eric Xing's slides

[5] http://scikit-learn.org/

[6] Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

[7] Prof. M.A. Papalaskar's slides

[8] https://qiyanjun.github.io/2019f-UVA-CS6316-MachineLearning/

5 Anaconda 安装教程

Anaconda 可以便捷获取包且对包能够进行管理，同时对环境可以统一管理。Anaconda 包含了 conda、Python 在内的超过 180 个科学包及其依赖项。

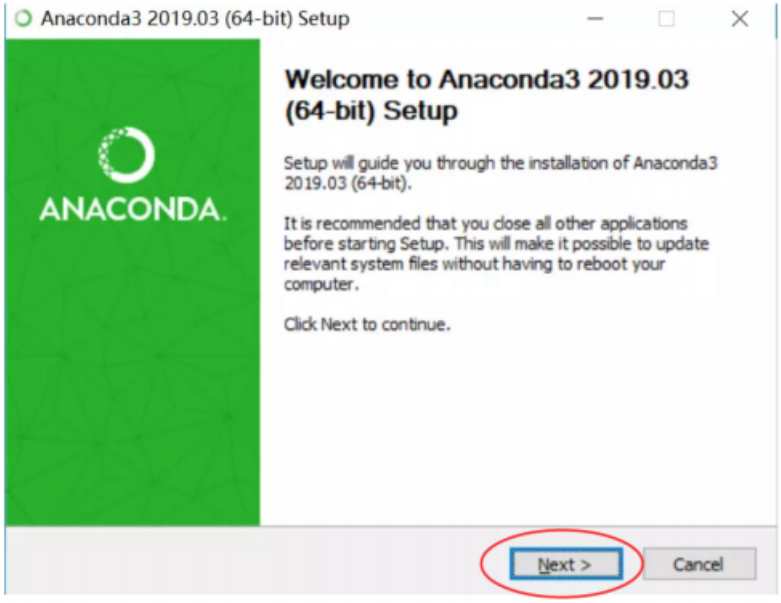
5.1 下载 Anaconda

官网下载较慢，使用清华镜像源
https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/
选择相应的.exe 版本

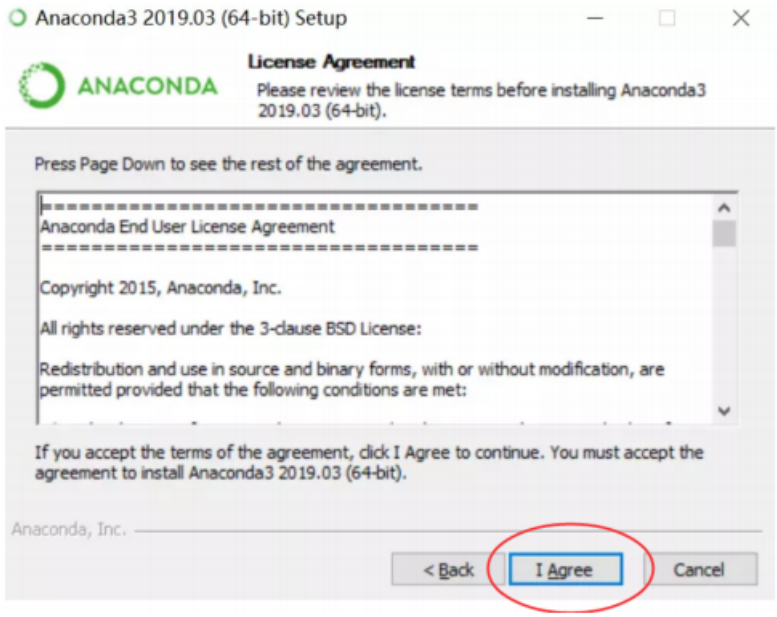
© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

5.2 安装

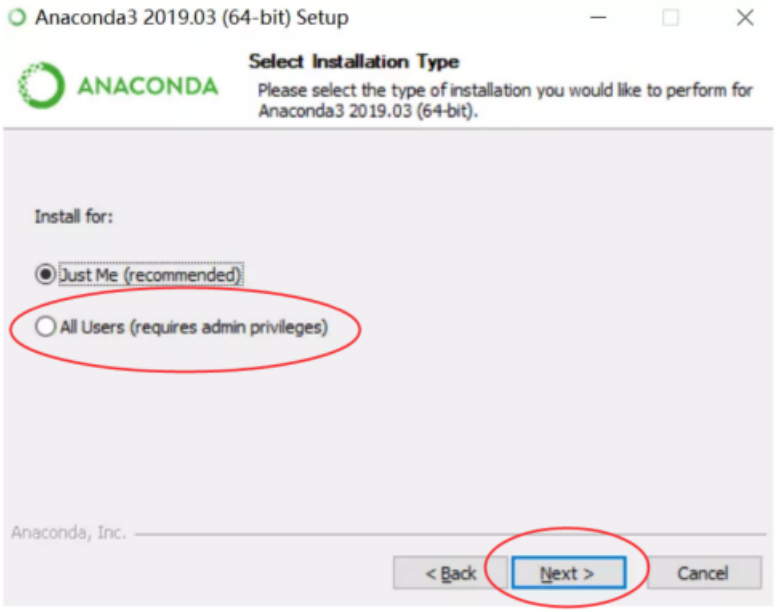
点击 next



点击 "I Agree"

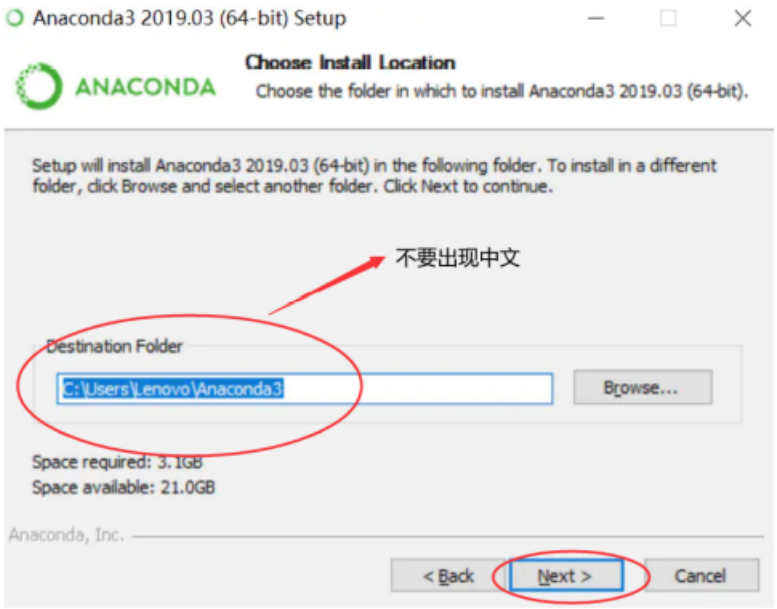


选择 "All Users"，**点击** next

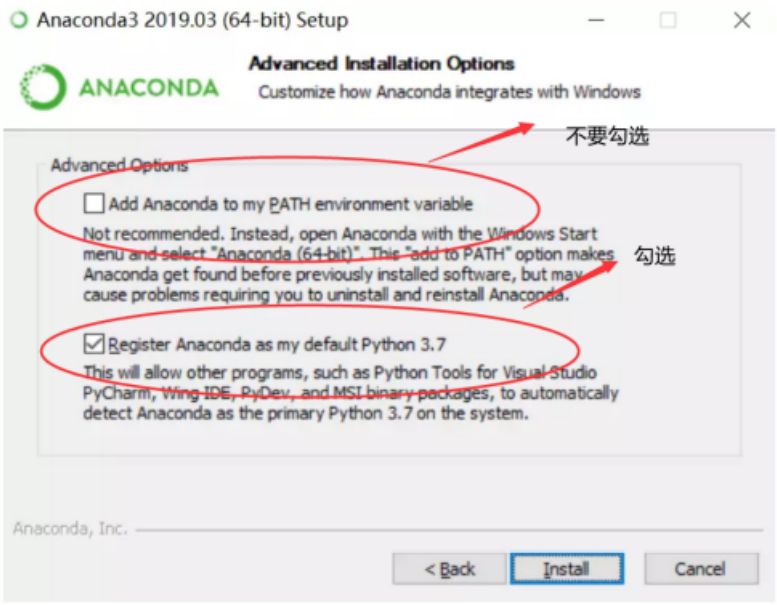


© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

选择安装路径，注意不要出现中文



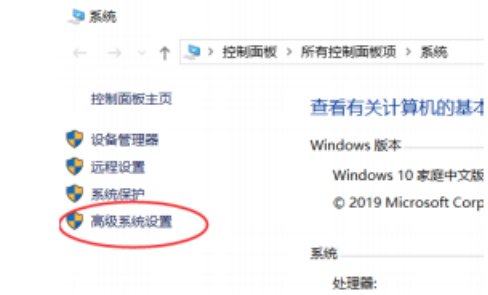
勾选对应项，点击 Install 完成安装

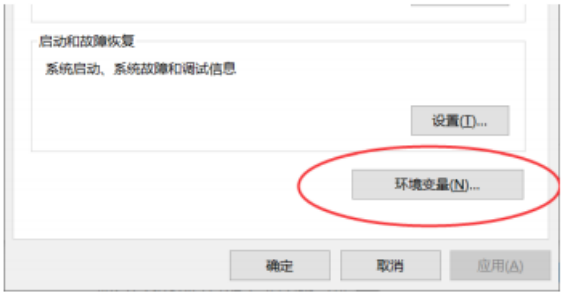


5.3 环境配置

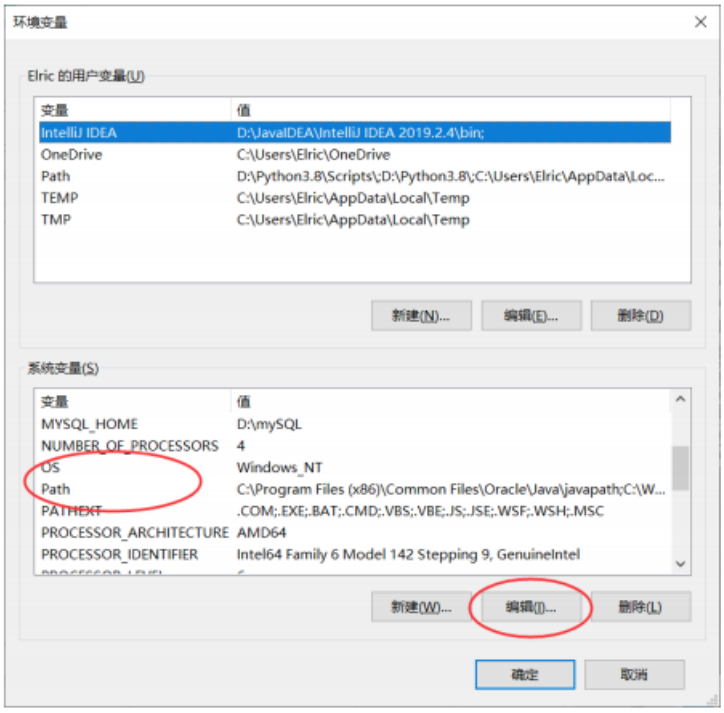
打开我的电脑-> 属性-> 高级系统设置-> 环境变量

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

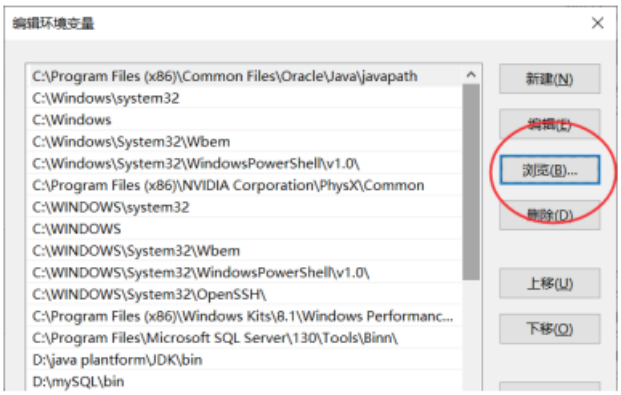




点击“path”，点击“编辑”



点击“浏览”

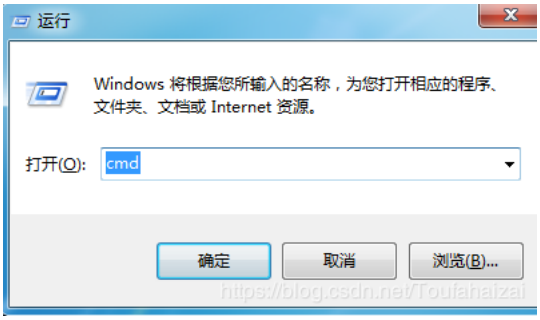


找到 Anaconda 的安装目录（例如 D:\python\anaconda)，点击“确定”，将该路径添加至环境变量。
找到 Anaconda 目录中“Script”目录（例如 D:\python\anaconda\Scrips)，点击“确定”，将该路径添加至环境。

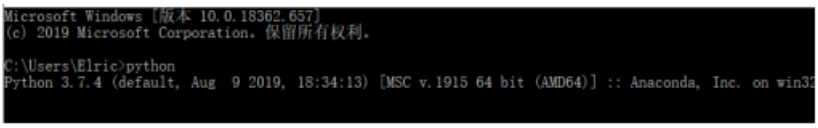
© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

5.4 检验是否安装成功

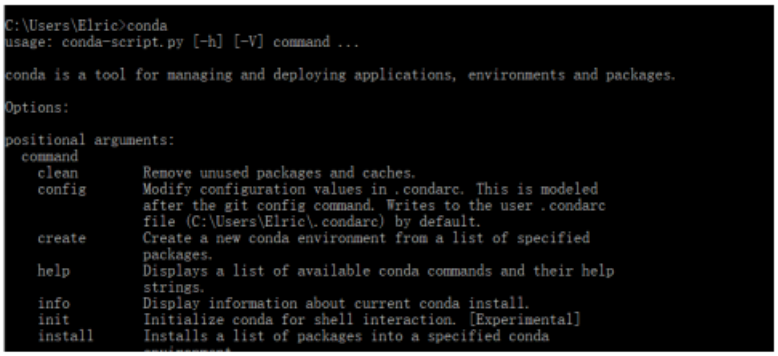
按“win+R”打开“运行”，输入“cmd”打开命令行



输入“python”查看 python 是否正常运行



输入“exit()”退出 python. 接着输入“conda”查看 conda 是否正常运行

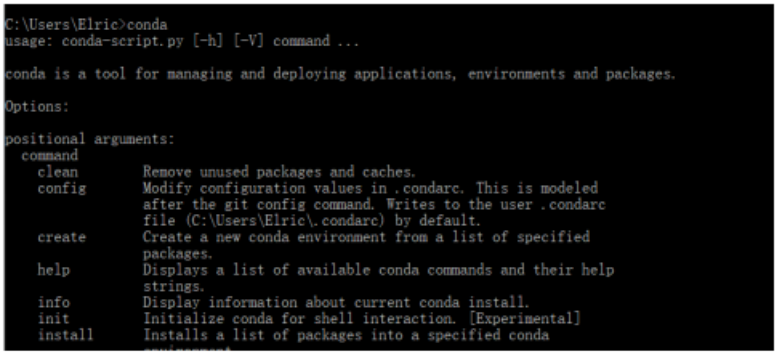


安装完成

6 Jupyter Notebook 使用说明

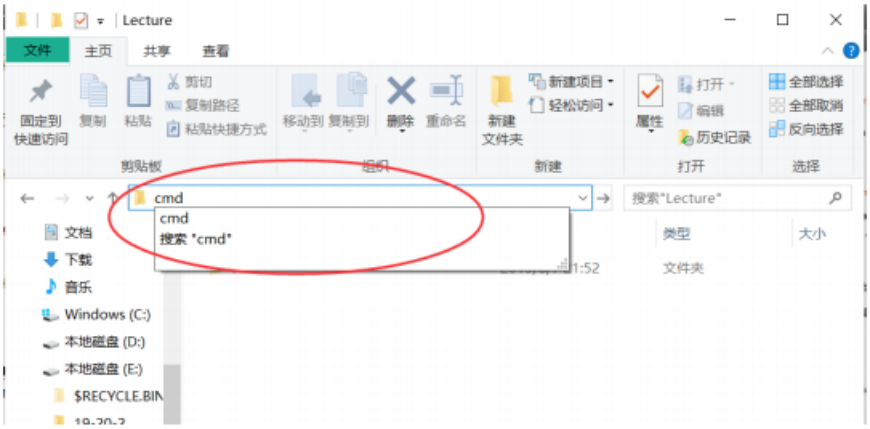
6.1 启动

下载 Anaconda 同时会自动安装 Jupyter Notebook，如果没有，可以通过在控制台输入“pip install jupyter”下载
打开任意一个文件夹，单击文件夹目录，输入“cmd”，然后按下回车，在该文件目录下打开控制台

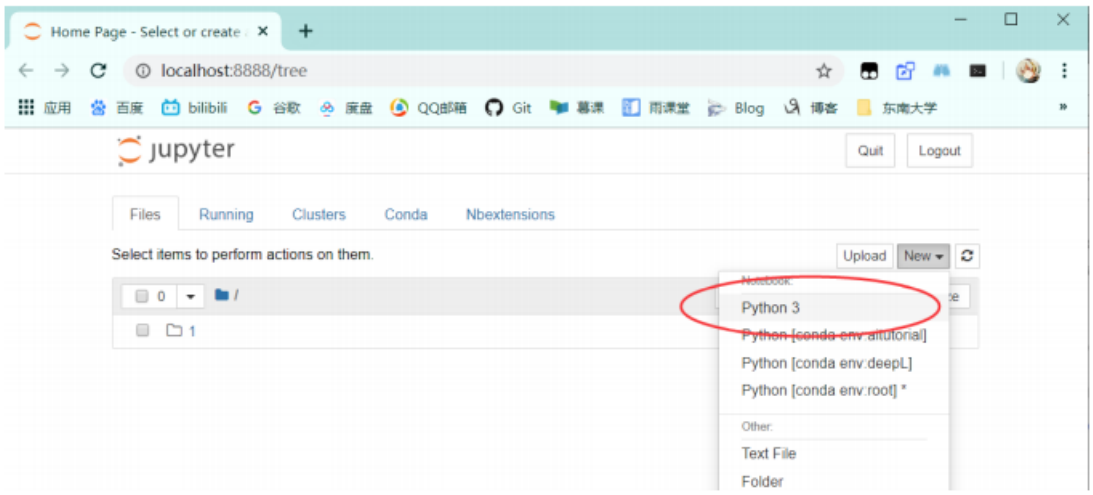


在命令行输入“jupyter notebook”，然后按下回车，启动 jupyter notebook 并自动通过浏览器打开

© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

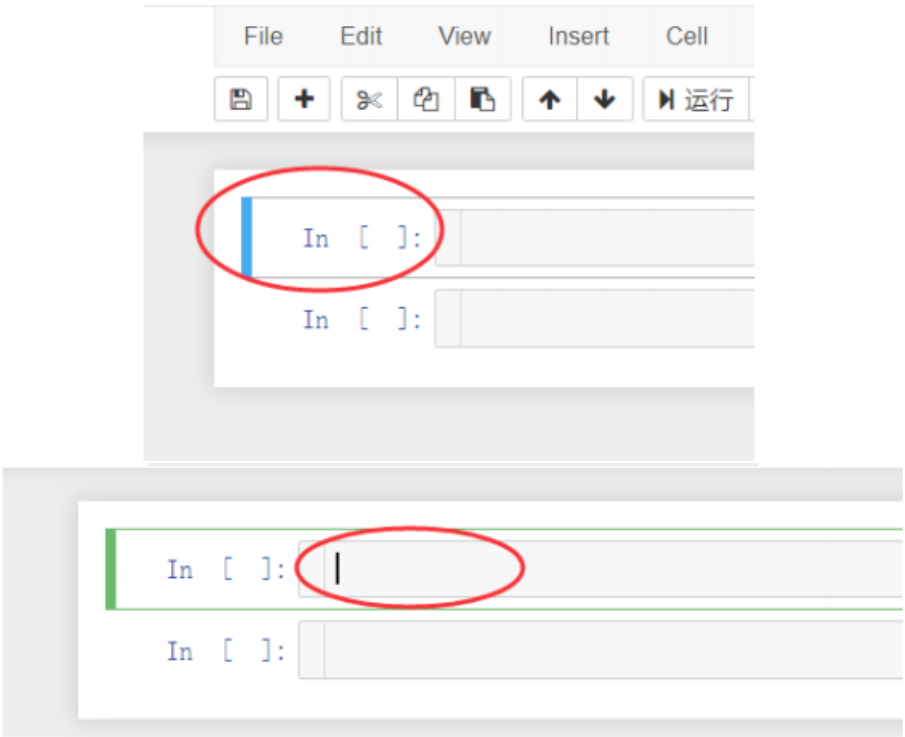


若存在“.ipynb”后缀的文件，则为 jupyter notebook 文件，点击即可打开。若不存在，可新建文件。点击“new”，点击“python3”创建一个 python3 内核的文件



6.2 功能

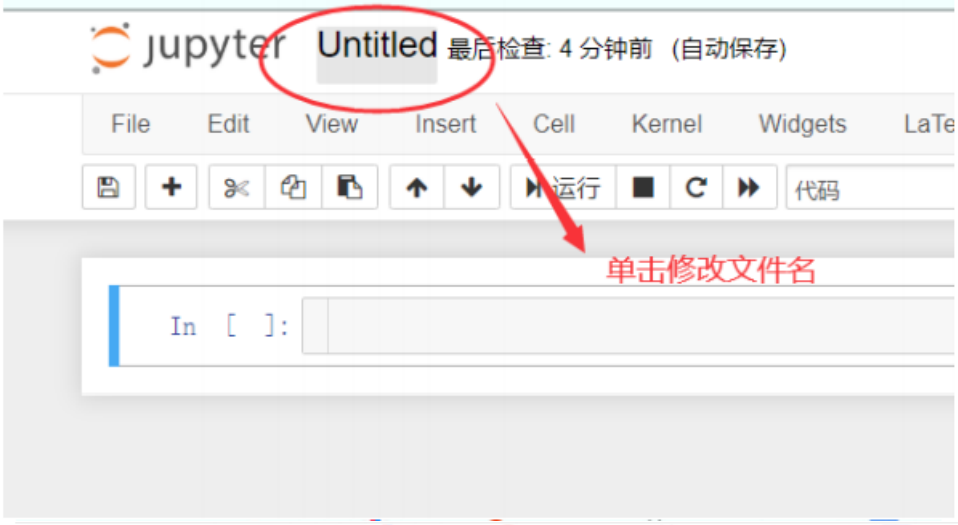
当点击分块（cell）左边区域时，分块会进入“命令行状态”，此时左边会变成蓝色；当点击分块编辑框时，分块会进入“编辑状态”，此时左边会变成绿色

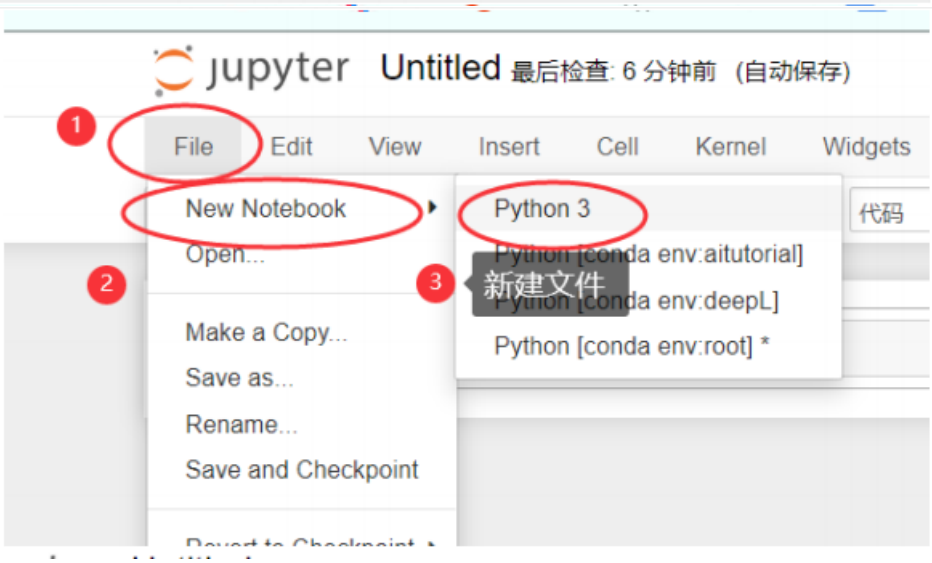


分块有两种类型，一种是代码块，一种是文本块。代码块支持添加 python 代码并可运行代码片段，文本块支持 markdown 语法并可渲染成 markdown 文本。代码运行和文本渲染均可通过选中分块后按“Ctrl+ 回车”或“Shift+ 回车”来执行。

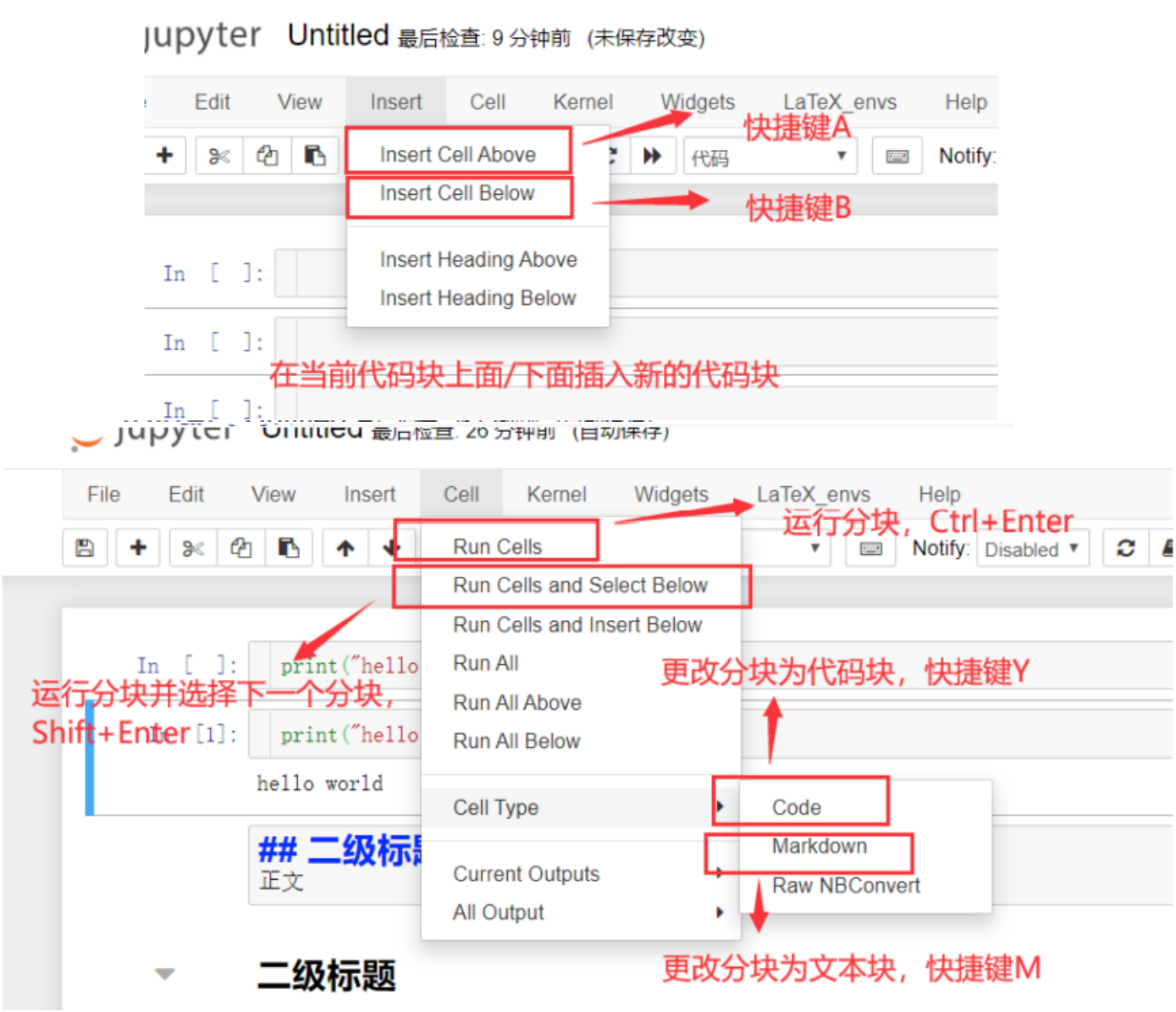
© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。

6.2.1 命令行模式主要功能和快捷键

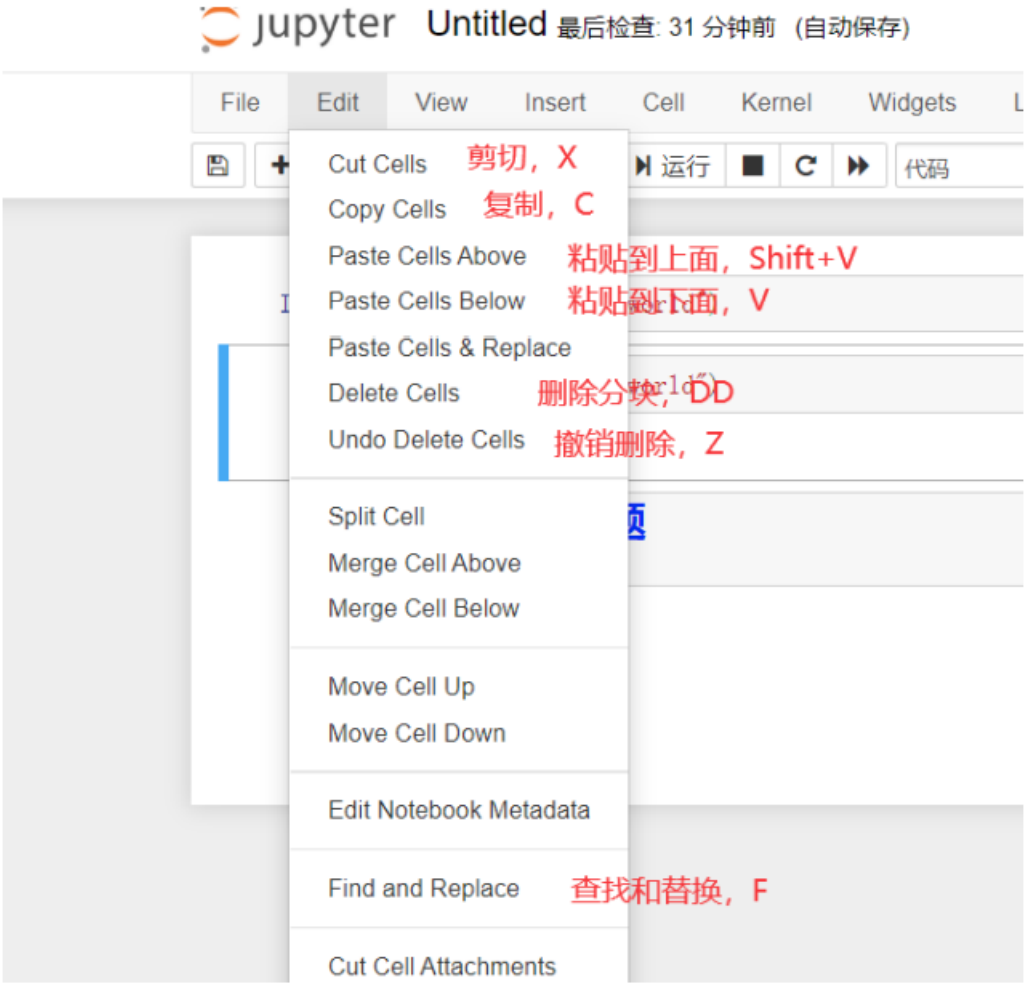




© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。



© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。



© 2020 本课件仅限东南大学机器学习课程使用，版权属于相关作者，严禁传播。