人工智能

语文出版社

课时分配表

|  |  |  |  |
| --- | --- | --- | --- |
| **章序** | **课程内容** | **课时** | **备注** |
| **1** | 初识人工智能 | 6 |  |
| **2** | 人工智能应用 | 7 |  |
| **3** | 新一代人工智能生态 | 6 |  |
| **4** | 大模型实践 | 6 |  |
| **5** | 人工智能伦理 | 8 |  |
| **6** | 机器学习 | 3 |  |
| **总计** |  | 36 |  |

# 第6课 机器学习

|  |  |  |
| --- | --- | --- |
| **课 题** | 机器学习 | |
| **课 时** | 6课时（270 min）。 | |
| **教学目标** | **知识技能目标：**  1．深入掌握机器学习模型基本概念、工作原理以及不同类型的模型的特点和适用场景。  2．了解数据在机器学习中的作用，掌握数据收集、预处理等关键步骤。  **思政育人目标：**  让学生通过学习机器学习，培养个人持续学习的习惯，主动追踪机器学习领域的最新进展，通过自学不断提升个人的技术水平和知识储备，以适应技术发展的快速变化。 | |
| **教学重难点** | **教学重点：**数据预处理技术、特征工程方法  **教学难点：**模型训练的基本流程 | |
| **教学方法** | 讲授法、问答法、讨论法 | |
| **教学用具** | 电脑、投影仪、多媒体课件、教材 | |
| **教学设计** | 第1节课：考勤（2min）--知识讲解（40min）--作业布置（3min）  第2节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第3节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min） | |
| **教学过程** | **主要教学内容及步骤** | **设计意图** |
| **考勤**  **（2min）** | ■【教师】清点上课人数，记录好考勤  ■【学生】班干部报请假人员及原因 | 培养学生的组织纪律性,掌握学生的出勤情况 |
| **知识讲解**  **（40min）** | **【教师】**讲解千变万化—— 机器学习模型  **一、机器学习概念**  机器学习（Ma chine Learning）是探索计算机如何模拟人类学习行为，进而获取新知、重塑知识结构并持续优化其性能的技术。作为人工智能的核心，它赋予了计算机智能的基石，广泛应用于人工智能的各个领域。与主要依赖演绎法不同，机器学习侧重于归纳与综合。近十年来，这一技术推动了自动驾驶、语音识别及网络语义搜索等领域的显著进步。  **二、机器学习的过程**  机器学习的实现涵盖学习与测试两大环节。  在学习环节，首先，需依据实际问题选定合适的模型，模型即为一组函数的集合。其次，要确定衡量函数优劣的标准，即损失函数，其选择需根据具体问题来定，例如回归问题常用欧几里得距离，而分类问题则倾向于交叉熵代价函数。最后，核心在于快速且准确地从众多函数中找出“最优”函数，这通常借助梯度下降算法、最小二乘法等方法来实现。测试环节同样关键，将学习得到的“最优”函数应用于新样本进行测试，唯有在新样本上亦能表现出色，该函数方可被视为真正“好”的。  **三、机器学习与人类学习对比**  机器学习属于人工智能的一个分支。所以，学习是一种智能，图 6-1 给出了机器学习在人工智能学科中的地位。图 6-2 给出了机器学习与人类学习的对比。    **四、机器学习模型**  面对众多实际问题，由于我们无法知晓如何从给定输入精确计算出期望输出（缺乏算法），或即便知晓，计算成本也可能极其高昂（涉及指数级复杂度），导致传统编程方法难以奏效，因为系统设计者无法明确指定输入数据到输出的转换路径。此时，一种有效的解决策略便是机器学习，它使计算机能够从示例中自动学习并建立起输入数据到输出的函数映射关系。  如果把机器学习问题看作拟合问题，那么机器学习模型就是从无限的函数中找到满足一定条件的拟合函数。  拟合函数有简有繁（a、c 简单，b、d 复杂），拟合度有高有低（a、b、d 拟合度高，c 拟合度低）。如果拟合函数用 h（x）表示，可能的拟合函数集合为假设空间 H，如果实际的输出为 y（x），则机器学习的任务就是在 H 中寻找 h（x），使 | y（x）-h（x）| 最小。注意这里的 | y（x）-h（x）|是误差度量函数，不一定指“差的绝对值”。所以，学习的过程就是寻找逼近 y 的拟合函数 h（x）的过程。  人工神经网络（Artificial Neural Network，即 ANN），是 20 世纪 80 年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互连接构成。每个节点代表一种特定的输出函数，称为激励函数（activation function）。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式、权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。最近十多年来，人工神经网络的研究工作不断深入，已经取得了很大的进展，其在模式识别、智能机器人、自动控制、预测估计、生物、医学、经济等领域已成功地解决了许多现代计算机难以解决的实际问题，表现出了良好的智能特性。  **五、机器学习知识框架**  学习机器学习的知识框架，如图 6-3 所示。 | **讲解千变万化—— 机器学习模型，让学生更加仔细的了解知识，从而激发学生的学习欲望。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了千变万化—— 机器学习模型，让学生了解机器学习的基本概念、原理及其在不同领域中的应用。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  机器学习与人类学习在学习方式、知识获取以及应用场景上有哪些异同点？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解去伪存真—— 数据准备  **一、数据集划分**  **（一）训练集**  训练数据（Train data）集用于构建模型，其中每个样本均附带标签，即正确答案。模型在训练集上表现优异，并不直接等同于模型整体性能优良；我们更期望模型能对未知数据展现出良好的泛化能力。训练是建模阶段的一环，通常在线下完成。若将机器学习过程比作高考历程，那么训练就如同考前的日常练习。  **（二）验证集**  为确保模型对未知数据表现优异，我们采用验证数据集来全面评估模型性能。若评估结果未达预期，将调整构建学习模型的参数，直至获得满意的训练模型。然而，即便模型在验证集上表现出色，也并不能完全证明其优越性，我们依然期望模型能对未见过的数据展现良好的泛化能力。验证是建模阶段的重要环节，通常在线下进行。若将机器学习比作高考之旅，验证则如同月考或周考，是对学习成果的一种阶段性检验。  **（三）测试集**  测试数据（Test data）集是一个在建模阶段没有使用过的数据集。我们希望模型在测试集上有好的表现，即强泛化能力。测试属于模型评估阶段，线上进行。如果把机器学习过程比作高考过程，验证相当于高考取得的成绩。  **（四）数据集划分标准**  一般来说采用 70/15/15，但这不是必须的，要根据具体任务确定划分比例。  **二、数据标注**  数据标注是一种由数据加工人员执行的行为，他们可能会借助像 BasicFinder 这样的标记工具来对样本数据进行加工处理。通常来说，数据标注的类型有很多种，包括图像标注、语音标注、文本标注以及视频标注等。以图像标注为例，其基本形式涵盖了标注画框、3D 画框、类别标注、图像打点以及描绘目标物体的轮廓线等多种方式，如图 6-5 所示。  以人脸识别为例，虽然大多数人知道这一功能是由智能算法实现的，但很少有人深入了解其背后的原理。智能算法之所以能够识别人脸，是因为它经历了大量的学习过程。在这个过程中，人工会对人脸样本进行细致的标注，如标注脸型、眼型、嘴型等特征，将算法原本无法理解的图像内容转化为容易识别的数字形式。随后，算法模型会依据这些标注后的图像内容进行学习，逐步建立起人脸特征与身份之间的逻辑关联。因此，海量的、覆盖各种脸型特征的数据是训练出优秀模型算法的关键，它们为算法提供了全面的学习材料，使其能够准确识别并区分不同的人脸。    数据标注的质量影响学习的效果，数据标注的成本非常高，如何实现自动化数据标注是机器学习领域研究的热点。 | **通过教师讲解，熟悉去伪存真—— 数据准备。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了去伪存真—— 数据准备，让学生了解数据标注的基本概念、流程和常用工具，能够准确高效地对数据进行标注。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  当数据集规模较小时，如何有效利用有限的数据划分来避免过拟合或欠拟合的问题？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解灵活多变—— 机器学习方式  **一、有监督学习**  有监督学习是计算机（作为求知的学生）从环境（扮演老师的角色）中汲取知识与信息的过程，其中环境提供带有正确或错误标签的训练集，并给出最终答案的指导。计算机通过学习这一过程，不断积累经验和技能，形成模型，从而能够对未学习过的问题，即测试集，做出准确的预测和解答。  简单地说，就是通过训练集学习得到一个模型，然后用这个模型进行预测。根据预测数据是否连续，有监督学习分为两类。  （1）回归：预测数据为连续型数值。  （2）分类：预测数据为类别型数据，并且类别已知。  **（一）线性回归**  如果希望知道自变量 X 是怎样影响因变量 Y 的，以一元线性回归为例，从数学角度，就是  建立模型：    其中，β=（β0，β1）称为回归系数。  参数 β0 和 β1 决定了回归直线相对于训练集的准确程度，即模型预测值与训练集中实际值之间的差距，如图 6-7 中的 ei，称为建模误差。    我们希望建模误差越小越好，用公式表示：    一般使用梯度下降算法求使建模误差最小化的参数 β0 和 β1 的值，如下：    **（二）决策树**  1. 基本思想  决策树通过模拟人类的决策流程，依据属性的重要程度顺序，对数据集的每个属性进行逐一判断，最终得出输入数据所对应的预测结果。  2. 基本概念  决策树由根节点、多个内部节点以及叶子节点构成，其中叶子节点代表决策的最终结果，内部节点则代表对样本特定属性的判断，而从根节点通往任意叶子节点的路径则构成了测试序列。  3. 决策树构建过程  决策树的构建始于依据某种分类准则确定最优分割特征，并计算相应的最优特征子函数，据此创建特征分割节点，依据该节点将数据集分割为多个子数据集。随后，在每个子数据集上再次应用判别准则，生成新的节点并作为树的分支。这一过程递归进行，直至达到递归停止条件为止。  4. 划分特征选择  合理选择其内部节点所对应的样本属性，使节点所对应样本子集中的样本尽可能多地属于同一类别，即具有尽可能高的纯度。特征选择的准则主要有 3 种：信息增益、信息增益比、基尼指数。  **（三）支持向量机**  支持向量机（Support Vector Machine，SVM）是一种基于有监督学习的二元分类广义线性分类器，其决策边界源于对学习样本求解得到的最优分类超平面。该方法由 Cortes 和 Vapnik 在1995 年首次提出，在处理小样本、非线性以及高维模式识别任务时展现出独特优势。与传统统计模式识别方法侧重于经验风险最小化不同，SVM 考虑了单纯经验风险最小化可能导致的“过拟合”问题及其较差泛化能力。根据统计学习理论，机器学习的实际风险由经验风险值和置信范围值共同构成。  **二、无监督学习**  无监督学习是在没有老师的情况下，学生自学的过程，如图 6-9 所示。无监督学习不局限于解决像有监督学习那样有明确答案的问题。因此，它的学习目标并不十分明确。常见的无监督学习任务是聚类、关联分析和降维。  **（一）聚类**  聚类模型是将物理或抽象对象的集合，分组为由类似的对象组成的多个类的分析过程。通过计算两个观测数据之间的距离来衡量其相似性，并以此为依据进行分组。与有监督学习不同，聚类是一种无监督学习方法，无须依赖预先标注的数据。  1. K-means 聚类  K-means 是最简单的聚类算法之一，其运用十分广泛。该方法有两个前提：通常要求已知类别数；只适用于连续型变量。  K-means 的计算方法如下。  （1）随机选取 k 个中心点。  （2）遍历所有数据，将每个数据划分到最近的中心点中。  （3）计算每个聚类的平均值，并作为新的中心点。  （4）重复 Step2、Step3，直到这 k 个中心点不再变化（收敛），或执行了足够多的迭代。  2. 层次聚类  层次聚类（Hierarchical Clustering，HC）是通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法。  **（二）关联分析**  1. 模型原理  为确定哪些商品顾客可能会在一次购物中同时购买，可以对商店的顾客事务、零售数量进行购物篮分析，如图 6-10 所示。该分析通过探索顾客“购物篮”中商品间的关联，来揭示顾客的购物习惯，从而帮助零售商识别哪些商品常被顾客一同购买，并据此制订更有效的营销策略。    2. 基本术语    3. Apriori 算法  1994 年，Agrawal 等人创立了项目集格空间理论，并在此基础上提出了著名的 Apriori 算法。该算法至今仍是关联规则挖掘领域的经典之作，引发了广泛讨论和深入研究。Apriori 算法专注于挖掘布尔关联规则中的频繁项集，其核心在于利用 Apriori 性质，即频繁项集的所有非空子集也必定频繁。该算法采用逐层搜索的迭代策略，先从数据库中扫描并累积各项计数，筛选出满足最小支持度的项，组成频繁 1 项集 L1。随后，利用 L1 寻找频繁 2 项集 L2，再用 L2 寻找 L3，依次类推，直至无法再找到新的频繁 k 项集。每次寻找 Lk 都需要对数据库进行一次全面扫描。  **（三）降维**  降维是指用较少数量（d 个）的向量来有效表示较多数量（D 个，其中 d ＜ D）向量所含的有用信息。之所以能实现降维，是因为数据中存在冗余，这包括无关信息或重复表达的信息。例如，在一张 512×512 的图像中，若仅中心 100×100 区域有非 0 值，其余部分即为无用信息；又如，对于中心对称的图像，其对称部分的信息是重复的。合理的降维处理能保留原始数据的大部分关键信息，使得降维后的数据能够替代原始输入进行后续操作，从而显著减少计算量，如降至二维或三维以便进行可视化处理。  一般来说，可以从两个角度考虑做数据降维，一种是直接提取特征子集做特征抽取，例如从 512×512 图中只取中心部分，另一种是通过线性 / 非线性的方式将原来的高维空间变换到一个新的空间，这里主要讨论后面一种——主成分分析 PCA。  PCA（Principal Component Analysis）是一种基于从高维空间映射到低维空间的投影的方法，其主要目的就是学习或者算出一个矩阵变换 W，其中 W 的大小是 D×d，d ＜ D，用这个矩阵与高维数据相乘得到低维数据。我们希望降维后的样本点尽可能分散（方差可以表示这种分散程度）。  **三、集成学习**  **（一）基本思想**  整体表现差的学习器在特定样本上可能超越“最好”学习器。在做重要决策时，人们通常会综合考虑多个专家的意见，而非仅依赖一个。类似地，集成学习通过融合多个学习器，能够在  一定程度上弥补单个学习器泛化能力不足的缺陷。例如，图 6-11 展示了如何通过集成三个线性分类器来实现二分类，这说明了即使单个学习器表现不佳，多个学习器的集成也可能在整体上提升性能。‌  **（二）集成学习使用场景**  （1）用于分类的特征可能包含统计特征和结构特征等多种类型，直接组合这些不同类型的特征来构建单个分类器往往具有挑战性。因此，一个有效的策略是分别用分类器对它们进行分类，然后再将结果组合。  （2）特征维数过高时，使用单一分类器进行识别会变得复杂。此时，将高维特征向量拆分为几个低维向量，并分别作为不同分类器的输入，之后再进行组合，不失为一种好方法。这种方法降低了对单个分类器性能的要求，既简化了分类器的构造，又减少了系统开销。  （3）不同分类器之间存在差异性，每种分类方法都有其独特的优势和局限，错误出现的情况也各不相同。利用这种差异性，可以通过多个分类器的互补来提高整体的分类性能。  **（三）半监督学习**  半监督学习是有监督学习和无监督学习相结合的一种学习方式。主要用来解决使用少量带标签的数据和大量没有标签的数据进行训练和分类的问题。  **（四）随机森林**  为了克服决策树容易过度拟合的缺点，随机森林算法（Random Forests，RF）把分类决策树组合成随机森林，即在变量（列）的使用和数据（行）的使用上进行随机化，生成很多分类树，再汇总分类树的结果。随机森林在运算量没有显著提高的前提下提高了预测精度，对多元共线性不敏感，可以很好地预测多达几千个自变量的作用，被称为当前最好的算法之一。随机森林通过自助法（Bootstrap）重复采样技术，从原始训练样本集 N 中有放回地重复随机抽取 k 个样本，生成新的训练集样本集合，然后根据自助样本生成 k 决策树组成的随机森林。它的最终结果是单棵树分类结果的简单多数投票。 | **通过教师讲解，掌握灵活多变—— 机器学习方式。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了灵活多变—— 机器学习方式，让学生了解有监督学习和无监督学习的基本原理、算法和应用场景。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  在有监督学习中，线性回归、决策树和支持向量机分别擅长处理哪些类型的数据和问题？ | 通过课后练习，使学生巩固所学新知识 |
| **教学反思** | 课堂营造更包容的课堂氛围，鼓励学生大胆表达，对积极参与的学生给予更多肯定。同时，建立多样化的反馈渠道。 | |