第 2 页：

上次分享我们介绍过人工智能的相关内容，是对人的思维的信息过程的模拟。而机器学习就是计算机利用已有的数据(经验)，得出了某种模型，并利用此模型预测未来的一种方法。简单来说，机器学习就是一种实现人工智能的方法，而深度学习则是实现机器学习的技术。

第 3 页：

2020 年 1 月 10 日，中国科学院大数据挖掘与知识管理重点实验室发布了《2019 年人工智能发展白皮书》，白皮书指出计算机视觉技术、自然语言处理技术、跨媒体分析推理技术、智适应学习技术等八大技术是目前人工智能领域的关键技术，安防、金融、零售、交通、教育等产业中蕴含着人工智能的典型应用场景，肯定了人工智能开放创新平台对于全行业的重要推动价值，**并推出全球人工智能企业 TOP20 榜单。**

其中，微软排名第一，市值为 1.21 万亿美元；谷歌和脸书位列第二和第三，市值分别为 9324 亿美元和 5934 亿美元。另外，百度位列全球第四，领跑中国，市值为 438 亿美元。中国共有 7 家企业上榜，其中，百度位列全球第四，领跑中国，市值为 438 亿美元；大疆创新、商汤科技、旷视科技、科大讯飞则“承包”了后四位。另外两家中国企业是松鼠 AI 1 对 1 和字节跳动，位列第十和第十一。中国的 AI 实力正在不断提升，并且在各行各业渗透。

第 4 页：

机器学习可以应用于模式识别、计算机视觉、语音识别、自然语言处理、统计学习、数据挖掘等领域。那么机器学习主要是解决什么类型的问题呢。

第 5 页：

机器学习并不是让机器像人一样会学习，而是通过一种固定的编程模式，对数据进行处理，从大量经验中寻找数学规律，从而在误差很小的情况下预估同类型事件的未来走向。

机器学习擅长解决给定数据的预测问题，如：数据清洗/特征选择； 确定算法模型/参数优化； 结果预测； 不能解决：大数据存储、并行计算或者做一个机器人。

第 6 页：

机器学习可以分为监督学习、无监督学习和强化学习。监督学习是针对有标签的数据而言的，可以通过输入的数据 X 来预测 Y。无监督学习是针对没有标签的数据而言的，可以通过输入的数据 X 来发现其中的规律、现象和特性。强化学习可以解决例如序列决策问题。

第 8 页：

聚类就是按照某个特定标准（如距离准则）把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。即聚类后同一类的数据尽可能聚集到一起，不同数据尽量分离。

主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

降维只要是解决如何将高维空间中的数据点映射到低维度的空间中？

主成分分析，也是处理降维的一种方法：是考察多个变量间相关性一种多元统计方法，研究如何通过少数几个主成分来揭示多个变量间的内部结构，即从原始变量中导出少数几个主成分，使它们尽可能多地保留原始变量的信息，且彼此间互不相关.通常数学上的处理就是将原来 P 个指标作线性组合，作为新的综合指标。

主成分分析的原理是设法将原来变量重新组合成一组新的相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的总和变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上处理降维的一种方法。

第 9 页：

绝大多数问题用典型机器学习的算法都能解决，粗略地列举一下这些方法如下：

- ✓ 处理分类问题的常用算法包括：逻辑回归(工业界最常用)，支持向量机，随机森林，朴素贝叶斯(NLP 中常用)，深度神经网络(视频、图片、语音等多媒体数据中使用)。
- ✓ 处理回归问题的常用算法包括：线性回归，普通最小二乘回归（Ordinary Least Squares Regression），逐步回归（Stepwise Regression），多元自适应回归样条（Multivariate Adaptive Regression Splines）
- ✓ 处理聚类问题的常用算法包括：K 均值（K-means），基于密度聚类，LDA 等等。
- ✓ 降维的常用算法包括：主成分分析（PCA）,奇异值分解（SVD） 等。
- ✓ 推荐系统的常用算法：协同过滤算法
- ✓ 模型融合(model ensemble)和提升(boosting)的算法包括：bagging，adaboost，GBDT，GBRT
- ✓ 其他很重要的算法包括：EM 算法等等。

通常，机器学习里所说的“算法”与程序员所说的“数据结构与算法分析”里的“算法”略有区别。

前者更关注结果数据的召回率、精确度、准确性等方面，后者更关注执行过程的时间复杂度、空间复杂度等方面。

当然，实际机器学习问题中，对效率和资源占用的考量是不可或缺的。

第 10 页：

人类学习的一般步骤是：

通过以往的经验，总结归纳出一系列规律，在遇到一个新的问题时，会利用掌握的的规律方法进行处理，以达到对未来的预测。

机器学习的步骤也是类似，通过历史数据来训练出一个好的模型，当输入新的数据时，也可以通过此模型来进行结果预测。

第 11 页：

最简单的步骤可以表示为：数据收集、数据清洗、特征工程和数据建模

假如我们的目标是想做一道番茄炒蛋，首先我们要需要的食材就是番茄和鸡蛋，有了材料之后，我们需要对这些材料进行清洗，清洗完之后我们要对这些食材进行处理，提取重要的部分，最后把打好的鸡蛋和切碎的番茄放到锅里炒成一盘番茄炒蛋。

如果我们真的想做一盘好菜，哪些方面是需要重点关注的呢？

第 12 页：

首先是数据，**成功的机器学习，不是拥有最好的算法，而是拥有最多的数据！数据和特征决定了机器学习的上界，而模型和算法只是逼近这个上界。**

各种不同的算法在输入的数据量达到一定的级数后，都有相近的高准确度。当数据足够大的时候，算法的的性能优势就没有太突出了。

第 13 页：

我们刚刚已经说过，数据和特征决定了机器学习的上界，那么特征工程是机器学习中不可或缺的一部分，在机器学习领域中占有非常重要的地位。

特征工程，是指用一系列工程化的方式从原始数据中筛选出更好的数据特征，以提升模型的训练效果。好的数据和特征是模型和算法发挥更大的作用的前提。特征工程通常包括数据预处理、特征选择、降维等环节。

通常原始数据需要经过：数据预处理、特征提取和特征转换等特征处理操作，以获得更有效的特征，再输入到模型进行学习并输出结果。

- **预处理：**经过数据的预处理，如去除噪声等。比如在文本分类中，去除停用词等。
- **特征提取：**从原始数据中提取一些有效的特征。比如在图像分类中，提取边缘、尺度不变特征变换特征等。
- **特征转换：**对特征进行一定的加工，比如降维和升维。

  特征抽取（Feature Extraction）： PCA、LDA

  特征选择（Feature Selection）：互信息、TF-IDF

第 14 页：

图中画出了机器学习的一般步骤，

当我们手中已经拥有了数据，并且确认好目标，那么我们首先需要对原始数据进行特征工程，选择出好的特征后在输入模型。在此之前，我们需要对数据集进行划分，将数据分为训练集和测试集，在训练过程中只能用训练集的数据，如果在训练过程中没有学到好的特征和规律，我们通常会重新选择模型或调参，当我们得到了一个训练好模型之后，再用测试集进行测试，此时得到的预测结果是最终的预测结果。

第 15 页：

机器学习的本质就是：学习输入和输出之间的函数关系，拿线性回归模型来说，

Y=wx+b，其中 y 是标签，w 是权重向量，x 是特征向量，也是我们输入的数据特征，b 是偏执

1）确定**损失函数**，（损失函数有很多种，MSE 均方误差损失函数、SVM 合页损失函数、Cross Entropy 交叉熵损失函数、目标检测中常用的 Smooth L1 损失函数。）
线性回归模型常用是 MSE(均方差) -> （Y(真实) – Y(预测)）的平方

2） 随机初始化: 权重向量 W 和 b

3） 将每个样本 X 的值带入 WX+b,计算一个 Y(预测)。

4） 根据 Y(预测)和 Y(真实)来计算 W 和 b 的梯度，根据梯度值来更新 W 和 b

5） 重复 3）到 4）的步骤，直到设置的训练轮数达到设定值。

第 16 页：

在人工智能机器学习中，很容易将“验证集”与“测试集”，“交叉验证”混淆。

一、三者的区别

训练集（train set） —— 用于模型拟合的数据样本。

验证集（development set）—— 是模型训练过程中单独留出的样本集，它可以用于调整模型的超参数和用于对模型的能力进行初步评估。 通常用来在模型迭代训练时，用以验证当前模型泛化能力（准确率，召回率等），以决定是否停止继续训练。

在神经网络中， 我们用验证数据集去寻找最优的网络深度（number of hidden layers)，或者决定反向传播算法的停止点或者在神经网络中选择隐藏层神经元的数量；

在普通的机器学习中常用的交叉验证（Cross Validation) 就是把训练数据集本身再细分成不同的验证数据集去训练模型。

测试集 —— 用来评估模最终模型的泛化能力。但不能作为调参、选择特征等算法相关的选择的依据。

一个形象的比喻：

训练集----------学生的课本；学生 根据课本里的内容来掌握知识。

验证集----------作业，通过作业可以知道 不同学生学习情况、进步的速度快慢。

测试集----------考试，考的题是平常都没有见过，考察学生举一反三的能力。

a)训练集直接参与了模型调参的过程，显然不能用来反映模型真实的能力（防止课本死记硬背的学生拥有最好的成绩，即防止过拟合)。

b)验证集参与了人工调参(超参数)的过程，也不能用来最终评判一个模型（刷题库的学生不能算是学习好的学生）。

c) 所以要通过最终的考试(测试集)来考察一个学(模)生(型)真正的能力（期末考试）。

但是仅凭一次考试就对模型的好坏进行评判显然是不合理的，所以接下来就要介绍交叉验证法

三、交叉验证法（模型选择）

a) 目的

交叉验证法的作用就是尝试利用不同的训练集/验证集划分来对模型做多组不同的训练/验证，来应对单独测试结果过于片面以及训练数据不足的问题。（就像通过多次考试，才通知哪些学生是比较比较牛 B 的）

如何确定模型调优的方向与思路呢？

需要对模型进行诊断的技术。

模型选择：拟合能力强的模型一般复杂度会比较高，容易过拟合。

如果限制模型复杂度，降低拟合能力，可能会欠拟合。

过拟合、欠拟合 判断是模型诊断中至关重要的一步。常见的方法如交叉验证，绘制学习曲线等。过拟合的基本调优思路是增加数据量，降低模型复杂度。欠拟合的基本调优思路是提高特征数量和质量，增加模型复杂度。

严格关注数据会**过拟合，**忽略数据会**欠拟合**。必须有一种方法来找到最佳平衡！幸运的是，数据科学中存在一种称为验证的完善解决方案。在我们的示例中，我们仅使用了训练集和测试集。这意味着我们无法提前知道我们的模型在现实世界中的作用。理想情况下，我们会设置一个"预测试"来评估我们的模型，并在真正的测试之前进行改进。**这种"预测试"被称为验证集，是模型开发的关键部分。**

误差分析 也是机器学习至关重要的步骤。通过观察误差样本，全面分析误差产生误差的原因:是参数的问题还是算法选择的问题，是特征的问题还是数据本身的问题……

诊断后的模型需要进行调优，调优后的新模型需要重新进行诊断，这是一个反复迭代不断逼近的过程，需要不断地尝试， 进而达到最优状态。