机器学习与TensorFlow简介

Derek

GDG Suzhou

西交利物浦大学博士生，GDG Suzhou 组织者
国际电气电子工程师协会计算机分会，中国计算机学会，ACM数据挖掘分会会员

agenda

- 机器学习及其重要性
- 机器学习的内容
- 统计机器学习
- 如何用于机器学习解决问题
- TensorFlow简介

机器学习及其重要性

什么是机器学习？

学术定义：
机器学习Machine Learning利用**一些方法**来使机器（计算机）**实现人的学习行为**，以便获取**新的知识和技能**，重新组织已有知识结构使之不断**改善自身的性能**。

机器学习的主要元素

实现人的学习行为:学习任务
一些方法:　　学习策略
新的知识和技能:　学习到的模型
改善自身的性能:　模型的应用效果

一个简单的例子

任务:识别或分类图像、文字、语音
策略: 优化函数(测试误差最小化)
模型:条件概率分布或决策函数
性能标准:分类准确度
一句话:机器学习根据一个任务，**选定一个模型**，**使用一个学习策略**，不断提高模型在任务上的性能

机器学习在当下的重要性

机器学习是人工智能AI的一个重要分支；
Google无人驾驶汽车，AlphaGo，西部世界等技术都基于机器学习。Google等大企业全面转型AI公司；
没有机器学习的大数据公司就是一个数据库和云计算公司。
没有机器学习的公司将丧失核心竞争力。

机器学习的一些误解

机器学习不是简单的数据分析和可视化。(很多人分不清数据挖掘和机器学习)
大数据公司不仅只是能对数据进行简单的统计分析和可视化。
机器学习需要在已有的数据上构建模型，模型具有判断和决策能力
数据分析和挖掘关注是什么，有什么意义；机器学习关注如何在已有的数据中提取经验，并运用这些经验应对和完成后续的任务，能面对未知事物。

机器学习的内容与常见模型

机器学习的内容

- 监督学习:分类，回归预测
- 无监督学习(聚类)
- 半监督学习
- 强化学习
- 迁移学习(新热点)

监督学习

监督学习是指利用一组已知类别的的样本调整分类器的参数，使其能够对未知类别的样本进行类别判断。

如：提供一组已知正常贷款和不良贷款的用户信息，训练模型。使其能够根据年龄，学历，工作状况等信息预测新增用户是否有不良贷款倾向。（芝麻信用，蚂蚁花呗，蚂蚁借呗）

美国ZestFinance，给没用信用记录的年轻人放贷（银行不敢放贷），坏账率很低。

监督学习

监督学习也可以应用已有数据进行回归预测
如股票，汇率，天气预测
预测近期波动
美国大多数股票交易由机器学习，而非操盘手完成

无监督学习

无监督式学习目的是去对原始资料进行分类，以便了解资料内部结构。无监督式学习在学习时并不知道其分类结果是否正确。其特点是仅对此种网络提供输入范例，而它会自动从这些范例中找出其潜在类别规则。

无监督学习就像没有老师指导的学习，自己去根据数据的特征，探索数据中潜在的类别。

无监督学习

聚类(Clustering)目的在于把相似的东西聚在一起，而我们并不关心这一类是什么。
当我们得到大量的数据，如电商用户购买记录时。我们关心的是我的用户里面有哪些人群，他们有哪些消费习惯。所以，我们通过聚类，将数据中行为相似的人归为一类，便于进一步分析和预测。
例如蚂蚁金服案例：经常购买紧身牛仔裤的年轻女性也经常买手机屏幕。所以当其买手机时推销碎屏险。

半监督学习

半监督学习是今年来模式识别和机器学习领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类问题。半监督学习对于减少标注代价，提高学习器性能具有非常重大的实际意义。

1. 有标记样本难以获取.需要专门的人员,特别的设备,额外的开销等等.
2. 无标记的样本相对而言是很廉价的.

半监督学习

思路:无标记的样本虽然不能直接用语监督学习，但结合无监督学习，我们能从其中发现数据的内在结构。

假设:与有标记样本相似的无标记样本和有标记样本应有相似的属性。可以将这些无标记样本视为有标记样本，应用到监督学习中作为辅助。这样可以使大量缺乏价值的无标记样本在少量有标记样本的帮助下发挥巨大价值。

在大数据(自然语言，图像与生物)中有巨大价值

强化学习

强化学习是机器学习中的一个领域，强调如何基于环境而行动，以取得最大化的预期利益。训练的方法是给定一些特定的输入和相应的奖惩机制，让机器学习在某种状况下采取何种行为能获得最大的奖励。

例如计算机学习西洋双陆棋。AlphaGo的策略网络就是运用了强化学习，通过输入的棋局和自我博弈，使得其知道在某一个地方落子的获胜比例。(价值网络负责搜索到获胜比例最高的落子点)。

迁移学习

迁移学习是近些年的研究热点。试图将在以往训练的模型中的参数直接运用于(迁移到)新的，训练中的新任务的模型。这样做的好处是能给给在训练中的模型一个好的初始参数，训练的速度更快。同样的数据集因为能够从以往的数据中获取经验，可以得到更优的性能。

如使用对imageNet训练得到的参数，作为初始参数，对新的数据训练模型。可大大缩短训练时间。

迁移学习是从弱人工智能到强人工智能的重要方式

一些常见的模型

决策树：一个树状的逐步决策过程，基于规则的模型。

k近邻算法(KNN): 通过相似的标记样本进行判断

支持向量机(SVM): 将线性不可分的低维空间映射到高维空间进行线性分割。将分类问题转化为一个高维空间上的平面分割最优化问题，基于结构风险最小化思想。

一些常见的模型

朴素贝叶斯估法:用贝叶斯估计代替极大似然估计的概率模型。求待测样本的最大后验概率类别。

人工神经网络: 基于神经科学的网络结构模型。本质上是一个输入到输出的函数，学习的内容是函数的参数。分为输入层，隐藏层(中间包括参数的层)和输出层。能够自动从数据中提取特征，优化特征空间。

一些常见的模型

深度学习：拥有2个以上的隐藏层的人工神经网络
卷积人工神经网络：一种结构较为复杂的多层人工神经网络。通过对数据进行特征提取，结合，在杂乱的数据中发现一些高层次的特征。卷积层中每个卷积核函数负责提取数据中的某种特征，较深的层组合这些特征成为高层次的特征。
例如在接近输入层的卷积层提取图像的线条轮廓信息，接近输出层的卷积层组合线条信息获得形状信息，最终能够判断是何种物体。

统计机器学习

统计学习的三要素

统计学习是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。
统计学习以计算机和网络为平台，以数据为研究对象，对数据进行分析和预测。
统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并在发展中逐步形成独子的理论体系和方法论。
统计学习是支撑机器学习发展的一个重要工具。

三要素:模型，策略，算法

统计学习的方法=模型+策略+算法

模型:统计学习首先要考虑的问题是选择什么样的模型。在监督学习中，模型是用于决策的函数或条件概率模型。决定用什么模型，也就决定了你可以用的函数的种类或者概率模型。我们的目标是在模型的范围内，找到一组能最好地完成任务地最优参数。

三要素:模型，策略，算法

策略:策略就是评价什么是最好地完成任务的标准。
评估一个参数下，模型的性能。
常见策略:损失函数和风险函数，经验风险最小化和结构风险最小化，最大后验概率估计。
不同的策略，将导致最终选定的最优参数不同。

三要素:模型，策略，算法

算法:如何基于数据集，根据学习策略，选择最优参数的方法。
由于假设空间过大，采用全局暴力搜索不可行且无法获得函数解析解，所以必须采取某些方法。
例如梯度下降法等。

确定了三要素，一个统计学习方法就确定了，所以三要素的选择极为重要。

如何运用机器学习解决现实问题

描述问题，明确任务

分析待解决的问题，明确问题对应的任务。例如是进行分类，预测，还是根据环境做出动态决策。

例如，利用医院已有体检结果数据进行学习，判断新病例是否患有癌症。
关键词:是否患有
任务:预测，分类

根据任务，确定问题对应的学习类别

确定任务对应的机器学习类别

是否患有癌症
分类问题
关键词：已有体检结果
全部数据有标注类别，在机器学习中的归类：监督学习

根据数据特征，选定合适的模型种类

因为模型基于一些统计学或其他理论假设，所以不同的模型适用与不同数据。需要选择假设能够满足的模型

是否患有癌症
在监督学习中选择一个分类模型
错误做法，对Iris鸢尾花卉数据集(四维，三类，150个数据)使用卷积人工神经网络，无从下手
卷积人工神经网络需要大量(上万，百万)的数据

选择学习策略和方法

根据所选模型的特征，数学性质，选择相应的学习策略和方法。

是否患有癌症
模型:朴素贝叶斯
策略:分类正确率
方法:结构风险最小化(等同于后验概率最大化)

学习建议

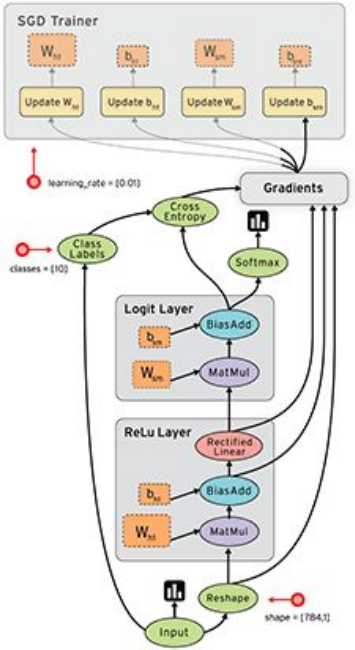
对于每种机器学习类别，学习一定量的模型
熟悉学习的策略和方法，掌握一定的数学工具
学会对数据进行分析和观察，对数据敏感。

常见教材:《机器学习》周志华
（俗称“西瓜书”）

TensorFLow简介

TensorFlow

一个用于机器学习的开源软件库



Data Flow

数据流图用“结点”和“线”的有向图来描述数学计算。“节点” 一般用来表示施加的数学操作，但也可以表示数据输入的起点/输出的终点，或者是读取/写入持久变量的终点。“线”表示“节点”之间的输入/输出关系。这些数据“线”可以输运“尺寸可动态调整”的**多维数据数组，即“张量”（tensor）。张量从图中流过**的直观图像是这个工具取名为“Tensorflow”的原因。一旦输入端的所有张量准备好，**节点将被分配到各种计算设备完成异步并行地执行运算**。

高度的灵活性

TensorFlow 不是一个严格的“神经网络”库。只要你可以将你的计算表示为一个数据流图，你就可以使用。你来构建图，描写驱动计算的内部循环。TF提供了有用的工具来帮助你组装“子图”（常用于神经网络），当然用户也可以自己在Tensorflow基础上写自己的“上层库”。当然万一你发现找不到想要的底层数据操作，你也可以自己写一点c++代码来丰富底层的操作。

用于但不局限于神经网络

真正的可移植性(Portability)

Tensorflow 在CPU和GPU上运行，比如说可以运行在台式机、服务器、手机移动设备等等。Tensorflow可以在CPU，GPU或者服务器上运行，灵活方便，可以不必纠结于平台，安心设计算法。

自动求微分

基于梯度的机器学习算法会受益于Tensorflow自动求微分的能力。作为Tensorflow用户，你只需要定义预测模型的结构，将这个结构和目标函数（objective function）结合在一起，并添加数据，Tensorflow将自动为你计算相关的微分导数。计算某个变量相对于其他变量的导数仅仅是通过扩展你的图来完成的，所以你能一直清楚看到究竟在发生什么。

多语言支持

Tensorflow 有一个合理的c++使用界面，也有一个易用的python使用界面来构建和执行你的graphs。你可以直接写python/c++程序，也可以用交互式的ipython界面来用Tensorflow尝试些想法，它可以帮你将笔记、代码、可视化等有条理地归置好。

性能最优化

比如说你又一个32个CPU内核、4个GPU显卡的工作站，想要将你工作站的计算潜能全发挥出来？由于Tensorflow 给予了线程、队列、异步操作等以最佳的支持，Tensorflow 让你可以将你手边硬件的计算潜能全部发挥出来。你可以自由地将Tensorflow图中的计算元素分配到不同设备上，Tensorflow可以帮你管理好这些不同副本。

学习建议

建议在Linux下使用GPU(只支持性能较好N卡)
推荐使用服务器版本，按时长付费，省去安装麻烦，性能强。
适合图像处理和自然语言分析等神经网络擅长的任务，并不是所有内容都适合用TensorFlow来做。
先学好机器学习在开始使用TensorFlow

学习资源

TensorFlow中文社区

Youtube

Bilibili 哲的王：TF Girls 修炼指南频道

12.14 GDG Suzhou TensorFlow专场，Google工程师现场解答