第01课 2019-11-10

马克飞象

人工智能总览

- MLP 多层感知器（1957） DNN FC
- 专家系统 （1980）
- 3秒之内
- 深度学习 （辛顿）
- 受限波儿兹曼机
- AlexNet ImageNet CNN
- Alpha Go Res DQN
- 强化学习 **游戏** 交互
- BERT 2018 Goole 发布 NLP 迁移学习
- ERNIE 百度
- 算力的提升
- 单机 分布式PS All Reduce
- 大数据 收集用户数据 数据价值的挖掘

人工智能应用

- 智能客服&助手 Siri Cortana 小艾 语音识别 NLP 语音合成
- 图像识别 百度识图 视屏解析 广告推广
- 实时视频解析 -交通拥堵预测 帧 CNN
- 个性化推荐： 千人千面的电商 线性 决策树 Deep and W
- 征信 转账 支付宝 银行 信用卡 风险管理
- 黑盒 SVM 深度学习的局限

> **机器学习算法**

- scikit-learn (numpy) python libarary
- 分类 聚类 降维 回归
- numpy(c c++) pandas matlib

- **分类和回归算法 离散是分类 连续是线性回归**
- $y = f(x) = ax + b$
- KNN
- 决策树，随机森林，GBDT
- SVM
- Linear Regression, Logistic Regression

- KNN 懒惰学习 **优点：冷启动**
- *超参 参数*

- 决策树 随机森林 GBDT SVM
- 数据集 样本 样本空间
- 二叉决策树
- 信息熵

> 如果某个事物的确定性越高，表示它需要的信息越少。
> 反之，越是随机，表示它需要的信息就越多。
> 信息的基本作用就是消除人们对事物的不确定性。
> 香农把它称为"信息熵" (Entropy)，一般用符号 H 表示，单位是比特

$$\mathbf{H(x)} = -\sum_{i=1}^{n} \mathbf{p}(x_i)\mathbf{log}\left(\mathbf{p}(x_i)\right)$$

- 百度百科：信息熵
- 变长编码
- 信息论 平均需要的比特数
- ID3算法 信息、信息熵和信息增益

> ID3算法(Iterative Dichotomiser 3，迭代二叉树3代)是一种贪心算法，用来构造决策树。ID3算法起源于概念学习系统（CLS），以信息熵的下降速度为选取测试属性的标准，即在每个节点选取还尚未被用来划分的具有最高信息增益的属性作为划分标准，然后继续这个过程，直到生成的决策树能完美分类训练样例。

- 信息增益

- Linear Regression, Logistic Regression
- 线性回归
- 最小二乘法
- 前向 后向 梯度 导数
- SGD 动量优化器 Adam
- 能量函数(损失函数) BGD与SGD

> BGD与SGD BGD
> 优点:沿全样本最优的方向下降 缺点:当样本数目很多时，训练过程会很慢。 从迭代的次数上来看，BGD迭代的次数相对较少。
> SGD
> 优点:训练速度快 缺点:准确度下降，并不是全样本最优;不易于并行实现。
> 从迭代的次数上来看，SGD迭代的次数较多，在解空间的搜索 过程看起来很盲目。

- **聚类算法**

> 给定一个有N 个 对 象的数据集， 划 分聚类 技术将构造
> 数 据的k个划分，每一个划分代表一个簇，k≤n。也就是说 ， 聚类将数据划分为k个簇
> 1 基于划分; 2 层次聚类; 3 基于密度; 4 基于网格

- K-Means算法 无监督

- **深度学习**

> 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。
> 观测值(例如一幅图像)可以使用多种方式来表示，如每个像素 强度值的向量，或者更抽象地表示成一系列边、特定形状的区域等。而使用某些特定的表示方法更容易从实例中学习任务(例如， 人脸识别或面部表情识别)。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.

- BP神经网络 DNN Sigmoid Reload 函数
- CNN
- RNN LSTM

- 误差反向传播(BP)网络
- 单极性Sigmoid函数:
- $f(x) = 1/1 + e - x$

- **C N N 卷积神经网络- 局部感知，参数共享**
- MNIST（32*32） DNN 图像 **缺点： 参数量大 过拟合**
- 卷积核 奇数卷积核 平移不变性
- 区别： 全连接神经网络 局部连接神经网络
- CNN卷积神经网络-Pooling
- **相关概念**

> 样本
> 训练数据
> 测试数据
> 训练过程
> BGD和SGD之间的区别
> 神经网络基本结构
> BP算法
> 卷积神经网络卷积核

主流机器学习库和计算框架介绍

- 框架
- scikit-learn
- Spark MLlib

> MLib 是Spark的可以扩展的机器学习库，由以下部分组成:
> 通用的学习算法和工具类，包括分类、回归、聚类、协同过滤， 降维，当然也包括调优的部分

- Spark MLlib K-means
- Tensorflow

如何进行机器学习

- 机器学习基本工作流程

> 1 问题建模
> 2 获取数据
> 3 特征工程
> 4 模型训练与验证
> 5 模型诊断与调优
> 6 线上运行

- 数据摸底
- 缺失数据处理
- One-Hot编码
- 特征缩放
- 模型训练
- Learning Curve
- 交叉验证
- 模型融合