

我们一起开始机器学习吧



王晋东不在家

<http://jd92.wang>

2016.12.22 20:30-21:30

1



目录

机器学习简介

- 什么是机器学习?
- 为什么现在会火?
- 为什么机器学习有必要?

迁移学习和深度学习

- 机器学习的未来: 无监督学习
- 什么是迁移学习?
- 深度学习的本质是什么?

机器学习快速入门建议

- 研究者、学生
- 工业界、开发者



知乎 Live

背景知识



下围棋



玩游戏

千秋明月照幽窗，一夜西风满院凉。山寺钟鸣惊宿鸟，水边芳草自生香。
一枕相思夜未休，春山秋雨惹离愁。凭栏望断江南月，花落无声水自流。
春到江南草更青，胭脂粉黛玉为屏。无端一夜西窗雨，吹落梨花满地庭。
百万兵戈战阵前，楚歌声里起狼烟。旌旗蔽日烽连塞，鼓角惊城血染关。
一夜秋风扫叶开，云边雁阵向南来。清霜渐染梧桐树，满地黄花坡上栽。
梨花落尽柳絮飞，雨打芭蕉入翠微。夜静更深人不寐，江头月下泪沾衣。
雨打芭蕉滴泪痕，残灯孤影对黄昏。夜来无寐听窗外，数声鸡鸣过晓村。
孤舟一叶泊江头，雁去无声送客愁。莫道春来芳草绿，人间万里尽风流。

写诗



谱曲



画画

- 机器好像越来越像人了。。。



什么是机器学习(1)

- 让计算机从数据中自动学习知识,并运用学到的知识来服务以后的任务。

自动挡!

培训主体

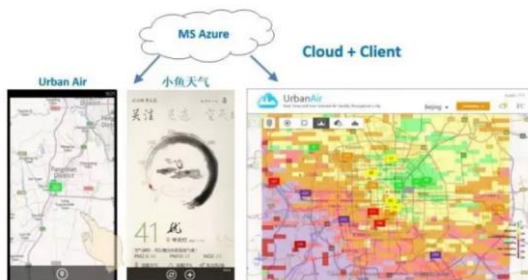
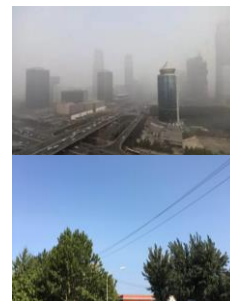
从哪里学

最终目的: 预测未来

■ 例子1: 预测雾霾

红色预警尚未结束, 北京启用机器学习预测空气污染

2016-12-22 THU数据派

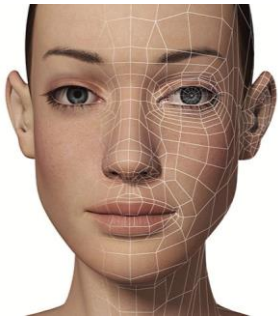


IBM绿色地平线计划
微软城市计算小组

什么是机器学习(2)

例子2：人脸识别

传统的方法：基于手动特征



- 鼻子多大
- 双眼距离
- 眼睛多大
- 肤色白不白?
- 嘴巴对鼻子距离
- 嘴巴多大
- 脸多宽
- . . .



扎克伯格



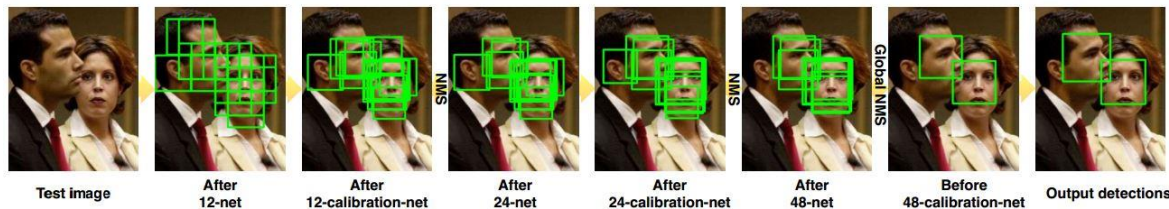
扎克伯格



- 有几张脸能完全一样?
- 我们白着呢!
- 这么多人怎么量距离?

“前世的五百次回眸换得今生的一次擦肩而过”

现在的方法：机器自动提特征





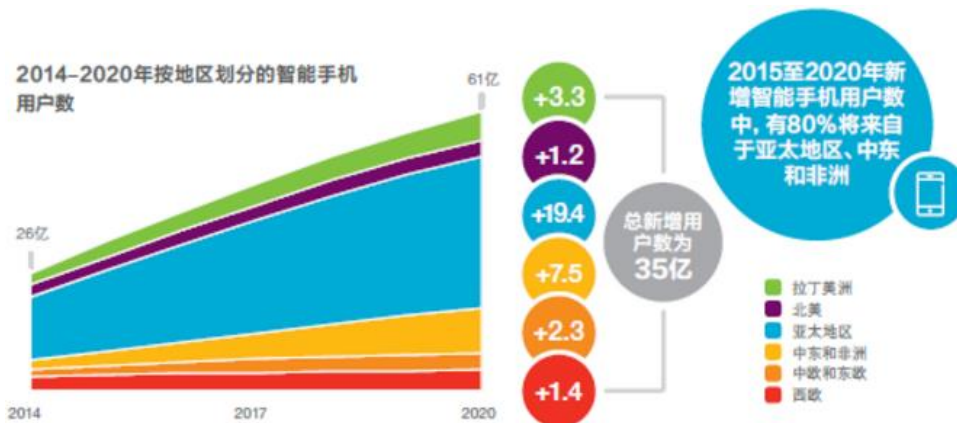
什么是机器学习(3)

- 前机器学习的时代
 - 基于经验：薄皮西瓜好吃
 - 基于规则推理：“我连窥天河，有云如蛇”——东风要来
- 这样的弊端：
 - 厚皮西瓜也有好吃的
 - 老夫夜观天象。。。年轻人不行
- 现实的情况是无限的
- 吾生也有涯，而知也无涯。以有涯随无涯，殆已！



什么是机器学习(4)

- 为什么机器学习现在火了?
 - 数据爆炸的时代! 一见多识广
 - 计算机更强大了!



所以, 机器学习是必要的!



无监督学习

■ 机器学习常用分类

- 监督学习
- 半监督学习
- 无监督学习
- 增强学习

■ "Pure" Reinforcement Learning (cherry)

- ▶ The machine predicts a scalar reward given once in a while.

▶ **A few bits for some samples**

■ Supervised Learning (icing)

- ▶ The machine predicts a category or a few numbers for each input
- ▶ Predicting human-supplied data
- ▶ **10→10,000 bits per sample**

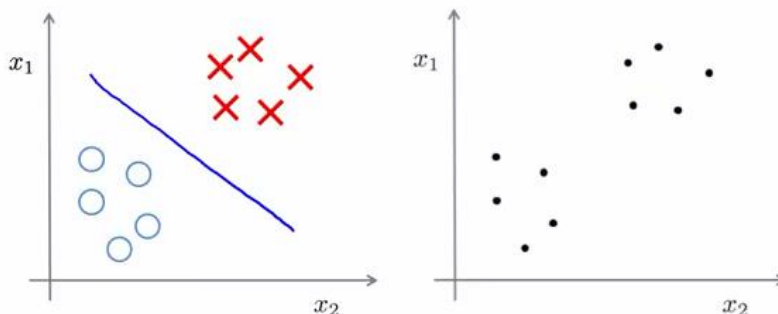
■ Unsupervised/Predictive Learning (cake)

- ▶ The machine predicts any part of its input for any observed part.
- ▶ Predicts future frames in videos
- ▶ **Millions of bits per sample**



■ (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

Living area (feet ²)	Price (1000\$)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



无监督才是世界的本源，没有标注怎么做？



迁移学习(1)

- 我们人，是有举一反三能力的：



会下中国象棋，
我们可以类比着学国际象棋



会骑自行车，
我们可以类比着骑摩托车



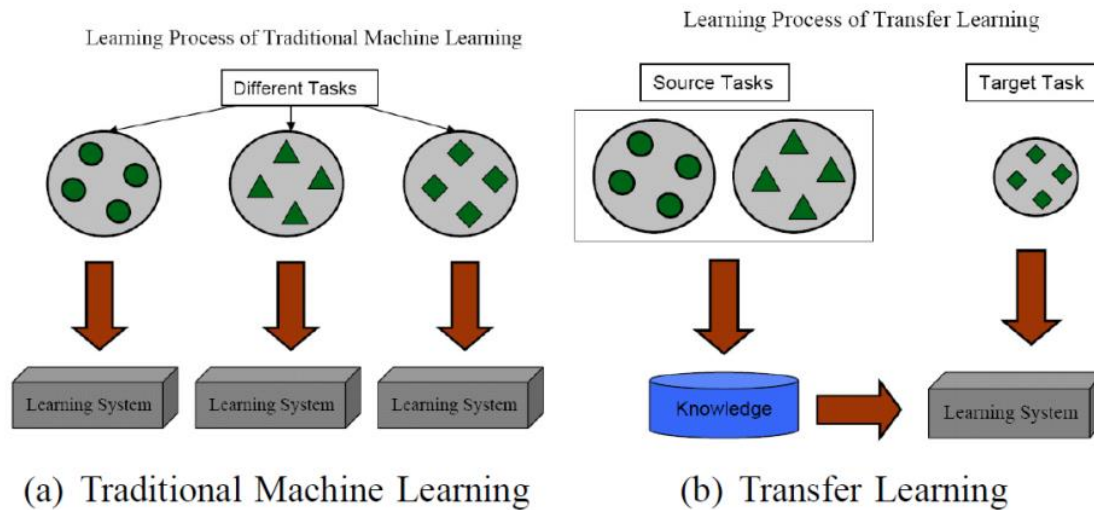
会写Java程序，
我们可以类比着学写C#

- 计算机可以吗？

迁移学习(2)

■ 迁移学习

- 将已经学习到的知识，应用于新的领域



- 节约新学习的成本
- 充分利用已有知识

负面迁移：

骑自行车→开汽车，能行不？

核心：找到相似性！

更多请参照我之前的报告：

http://jd92.wang/assets/files/l03_transferlearning.pdf

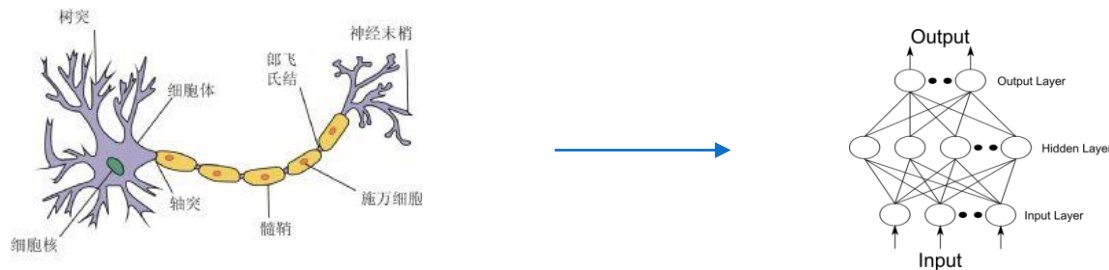
深度学习(1)

■ 深度学习的本质

- 模拟人脑中神经元的结构，拟合一个复杂函数 f
- 学习 f 的方法：复合形式

$$f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$$

- 把不同功能的函数按照一定层次进行复合



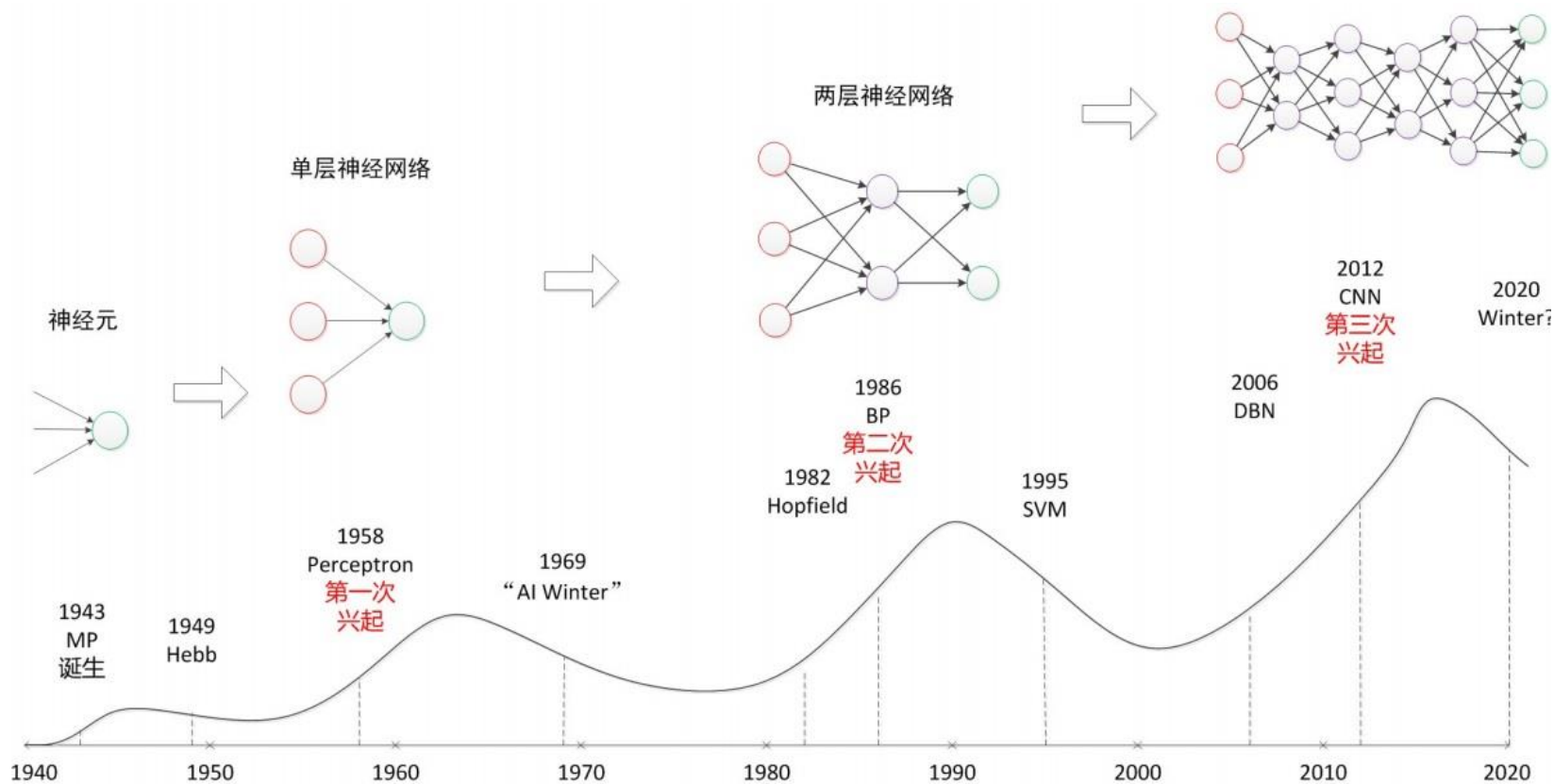
- 新瓶装旧酒!



知乎 Live

深度学习(2)

■ 深度学习的发展演变





深度学习(3)

■如火如荼中。。。.

2010年，美国国防部DARPA计划首次资助深度学习项目。

2011年，微软研究院和谷歌的语言识别研究人员先后采用DNN技术降低语音识别错误率20%-30%，是该领域10年来最大突破

2012年，Hinton将ImageNet图片分类问题的Top5错误率由26%降低至15%。同年Andrew Ng与Jeff Dean搭建Google Brain项目，用包含16000个CPU核的并行结算平台训练超过10亿个神经元的深度网络，在语音识别和图像识别领域取得突破性进展。

2013年，Hinton创立的DNN Research公司被Google收购，Yann LeCun加盟Facebook的人工智能实验室。

2014年，Google将语言识别的精准度从2012年的84%提升到如今的98%，移动端Android系统的语言识别正确率提高了25%。人脸识别方面，Google的人脸识别系统FaceNet在LFW上达到99.63%的准确率。

2015年，Microsoft采用深度神经网络的残差学习方法将Imagenet的分类错误率降低至3.57%，已低于同类试验中人眼识别的错误率5.1%，其采用的神经网络已达到152层。

2016年，DeepMind使用了1920个CPU集群和280个GPU的深度学习围棋软件AlphaGo战胜人类围棋冠军李世石。

国内对深度学习的研究也在不断加速：

2012年，华为在香港成立“诺亚方舟实验室”从事自然语言处理、数据挖掘与机器学习、媒体社交、人际交互等方面的研究。

2013年，百度成立“深度学习研究院”（IDL），将深度学习应用于语言识别和图像识别、检索，2014年，Andrew Ng加盟百度。

2013年，腾讯着手建立深度学习平台Mariana，Mariana面向识别、广告推荐等众多应用领域，提供默认算法的并行实现。

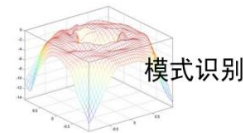
2015年，阿里发布包含深度学习开放模块的DTPAI人工智能平台。



深度学习(4)

■ 模型层面

- 深度信念网络 (DBN)
- 卷积神经网络 (CNN)
- 循环神经网络 (RNN)
- 深度+增强学习 (DLRL)
- 深度神经网络 (DNN)



模式识别

计算机视觉



数据挖掘



机器学习

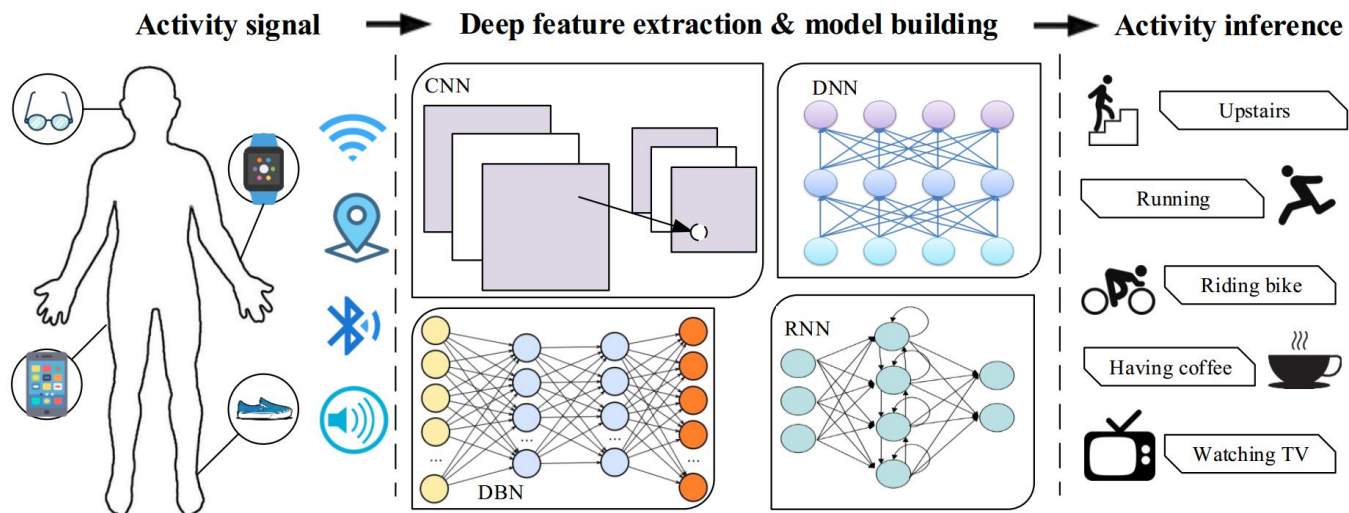
语音识别



统计学习



自然语言处理





深度学习(5)

- 为什么深度学习会火?
 - 数据量更大：对模型简直天然加成
 - 计算能力增强：GPU计算，Google、微软计算集群
- 核心：理论+大多数工程
 - 调参，调参，调参。。。
- 最近：生成对抗网络、预测学习、对偶学习
 - 生成对抗网络：一个网络学习，一个网络判准
 - 增强学习：惩罚
 - 对偶学习：互相学习知识



知乎 Live

机器学习库对比(1)

名称	面向语言	支持平台	上手难度	推荐指数
Tensorflow	Python, C++	  	★ ★	★ ★ ★
Scikit-learn	Python	  	★	★ ★ ★
Caffe	C++, Python, Matlab	  	★ ★	★ ★
MXNet	Python, R, Julia	  	★ ★ ★	★ ★ ★
Keras	Python	  	★	★ ★ ★
Theano	Python	  	★ ★	★
torch	C, lua	 	★ ★ ★	★
CNTK	C++, C#, Python	  	★ ★ ★	★
WEKA	Java	  	★ ★	★ ★
Orange	Python	  	★ ★	★
Deeplearning4j	Java	  	★ ★	★



TensorFlow



scikit-learn

dmlc

mxnet



Caffe

theano



K



torch



Microsoft

CNTK



WEKA

The University of Waikato



orange

DATA MINING

FRUITFUL FUN

<https://linux.cn/article-7252-1.html>



机器学习库对比(2)

- 机器学习、深度学习强烈推荐：
 - Scikit-learn+Tensorflow
 - Tensorlayer, tflearn
- 有GUI界面的(基本做不了深度学习):
 - WEKA, Orange
- 从零开始使用Python安装：
 - Anaconda包(<https://www.continuum.io>)
 - Pip(<https://pypi.python.org/pypi>)
 - Scikit-learn(<http://scikit-learn.org/>)



如何入门机器学习(1)

- 学习机器学习的目的：
 - 大学生、研究者→深入了解，发文章
 - 工业界、开发者→懂行情，会使用
- 工业界和学术界，界限并不明显





如何入门机器学习(2)

■ 大学生、研究者

准备工作

- 理论知识 (高数、概率、线性代数、随机过程)
- 编码能力 (Python, Matlab, Java)

基本入门

- 李航《统计学习方法》
- 周志华《机器学习》
- 吴恩达公开课
- Kaggle竞赛

进阶提高

- 《模式分类》、《PRML》
- ICML、NIPS等国际会议
- 做自己的研究工作

- 书籍资料整理: <https://github.com/ty4z2008/Qix/blob/master/dl.md>
- 入门资料: <http://www.cnblogs.com/subconscious/p/4107357.html>
- 公开课: <http://open.163.com/special/opencourse/machinelearning.html>
- Kaggle竞赛: <https://www.kaggle.com/>

更多请参照我的Github:

<https://github.com/jindongwang/MachineLearning>



如何入门机器学习(3)

■ 工业界、开发者

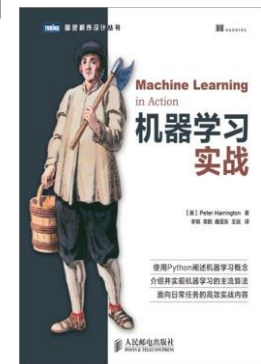


- 基本编程能力
(Python、Matlab)
- 常用框架了解

- 《机器学习实战》
- 框架基本使用

- Kaggle竞赛
- 做自己的项目

- 理论+实践，重在实践！
- 吴恩达写给工业界的新书：[Machine learning yearning](#)
- Kaggle竞赛：<https://www.kaggle.com/>



谢谢!

我们一起开始机器学习吧



王晋东不在家

<http://jd92.wang>

2016.12.22

21

路漫漫其修远兮
吾将上下而求索