机器学习入门

Todd 2016

提要

- 机器学习介绍
- 典型算法一：K均值聚类
- 典型算法二：神经网络
- 机器学习应用思考

例一：新闻聚合


Baidu 新闻
百度一下
帮助 高级搜索 设置
新闻全文 新闻标题
首页 百家 财经 娱乐 体育 互联网 时尚 汽车 国内 国际 军事 视频 传媒 个性推荐 房产 名站 更多
热点要闻 个性推荐
登录更懂你
习近平：推进军队规模结构和力量编成改革
重要讲话精神 图解 会见基辛格 向上价值观 治国 国平
李克强：切实把职业教育摆在更加突出的位置
塞拉利昂总统诚谢 总理点赞的赣南脐橙被抢购一空 刘云山出席会议
内蒙古一煤矿爆炸事故致32死 相关责任人被控
媒体：两月内矿难频发 勿让矿难随煤炭市场一起反弹 现场
特朗普与蔡英文通电话 中方提出严正交涉
济南厂房坍塌事故致5人遇难 相关责任人被控制
意大利公投在即 伦齐作最后"拉票" 专题
借贷宝称"裸条"并非平台产生 设百万"打裸基金"
春运火车票本月15日起开售 学生票已开始预售
女子手机疑在商店丢失调监控被拒 商家称监控权限在美国
外媒:中国去年380万对夫妇离婚 婚外情是头号杀手
央视曝九江彭泽瞒中央环保督查：来我停走我生产
法媒称个人征信将渗透中国人生活：逃票将影响就业
外媒：中国生育水平全球最低 二孩政策影响甚微
小偷将赃物藏派出所隔壁 民警守株待兔将其抓获
华北重霾能见度不足3公里 仍有单位违规排污
习近平：推进军队规模结构和力量编成改革
热搜新闻词 HOT WORDS
更多
习近平：推进军队规模结构和力量编成改革
习近平会见美国前国务卿基辛格
网络谣言中食品安全占45%
五战区司令集中出镜
万里与夫人合葬八宝山
春运火车票本月15日起开售
北京今明达五级重度污染
2017年春运1月13日开始
青岛现贝壳滩
韩国在野党提出总统弹劾案
百度百家 BAIJIA 财经 | 文化 | 娱乐 | 体育
加入百家号
HOT

例一：新闻聚合

实现方法	特点
人工筛选	工作量太大，不现实。
传统程序处理	没有清晰的规则。
机器学习	自动聚类，热点发现。

例二：图像识别

例二：图像识别

实现方法	特点
人工筛选	工作量太大，不现实。
传统程序处理	没有清晰的规则。
机器学习	由人类告诉程序一些样本，程序自动学习出规则。

机器学习定义

- 定义一：利用经验来改善计算机系统自身的性能。（周志华）
- 定义二：机器学习研究的是如何赋予计算机在没有被明确编程的情况下仍能够学习的能力。（Arthur Samuel，1959 ）
- 定义三：对于某个任务 T 和表现的衡量 P，当计算机程序在该任务 T 的表现上，经过 P 的衡量，随着经验 E 而增长，我们便称计算机程序能够通过经验 E 来学习该任务。（Tom Mitchell，1998）

机器学习应用

- 1．数据库挖掘
- 2．一些无法通过手动编程来编写的应用：如自然语言处理，计算机视觉
- 3．一些自助式的程序：如推荐系统
- 4．理解人类是如何学习的

大数据、数据挖掘与机器学习

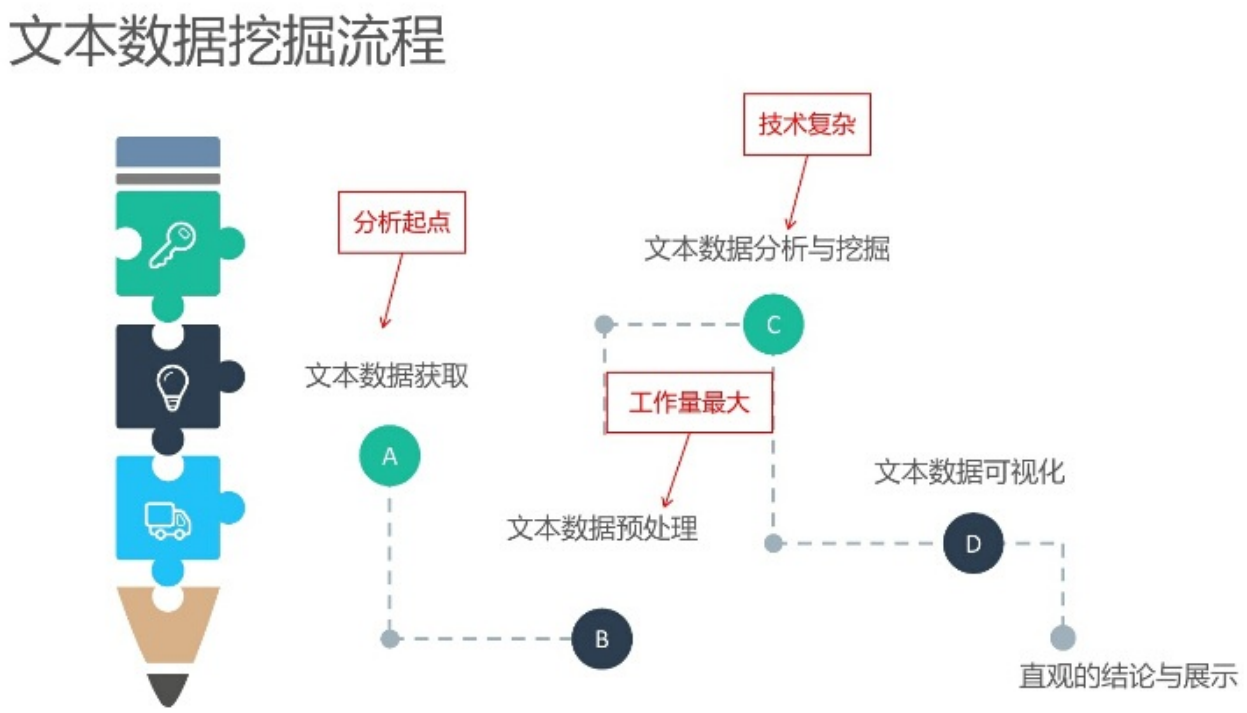
- 大数据≠大价值
- 数据挖掘包含数据管理技术与数据分析技术
- 机器学习是数据挖掘最常见数据分析技术
- 大数据可能改善机器学习的效果，但也是相对的。

深度学习、机器学习与人工智能

- 人工智能发展的三个阶段：推理期，知识期，学习期，其代表技术分别为自动定理证明系统，专家系统（clips），机器学习。
- 神经网络是机器学习中的典型技术。
- 深度学习是神经网络的变种或者加强版。
- 强人工智能 与 弱人工智能
- 重复性与创造性

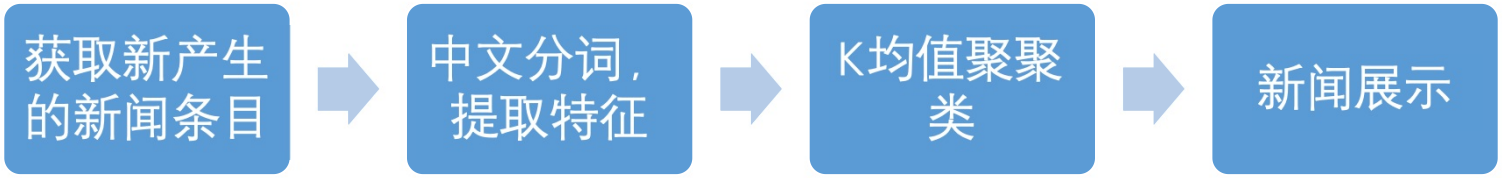
K均值聚类

- 百度新闻是如何实现聚类的？



K均值聚类

- 一个简化的流程



K均值聚类

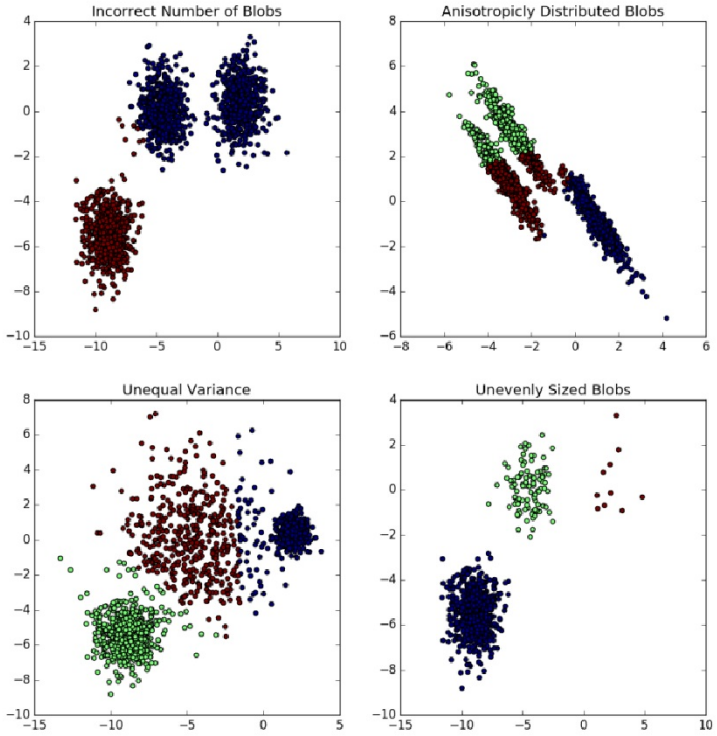
- 假设现在采集了24小时内产生的10000条新闻
- 中文分词后，提取这些新闻有意义的高频词，假设工涉及高频词300个，那么现在可以构建一个维度为300，样本数为10000的矩阵

	习近平	军队	改革	内蒙古	煤矿	……（共1000维）
新闻1	5	10	10	0	0	
新闻2	4	8	8	0	0	
新闻3	0	0	0	4	4	
新闻4	0	0	0	4	4	
……（共10000条）						

K均值聚类

- 自动提取最热门的10条新闻
- 300维空间中有10000个点，试图从中找出10个聚类结构来

简化：二维数据的聚类


Incorrect Number of Blobs
Anisotropicly Distributed Blobs
Unequal Variance
Unevenly Sized Blobs

基本思想

- 以空间中k个中心点进行聚类，对最靠近他们的对象归类，通过迭代的方法，逐次更新各聚类中心

具体步骤

- K-均值是一个迭代算法，假设我们想要将数据聚类成 n 个组，其方法为：
  1．首先选择 K 个随机的点，称为**聚类中心**（cluster centroids）
  2．对于数据集中的每一个数据，按照距离 K 个中心点的距离，将其与距离最近的中心点关联起来，与同一个中心点关联的所有点聚成一类
  3．计算每一个组的平均值，将该组所关联的中心点移动到平均值的位置
  4．重复步骤 2-4 直至中心点不再变化

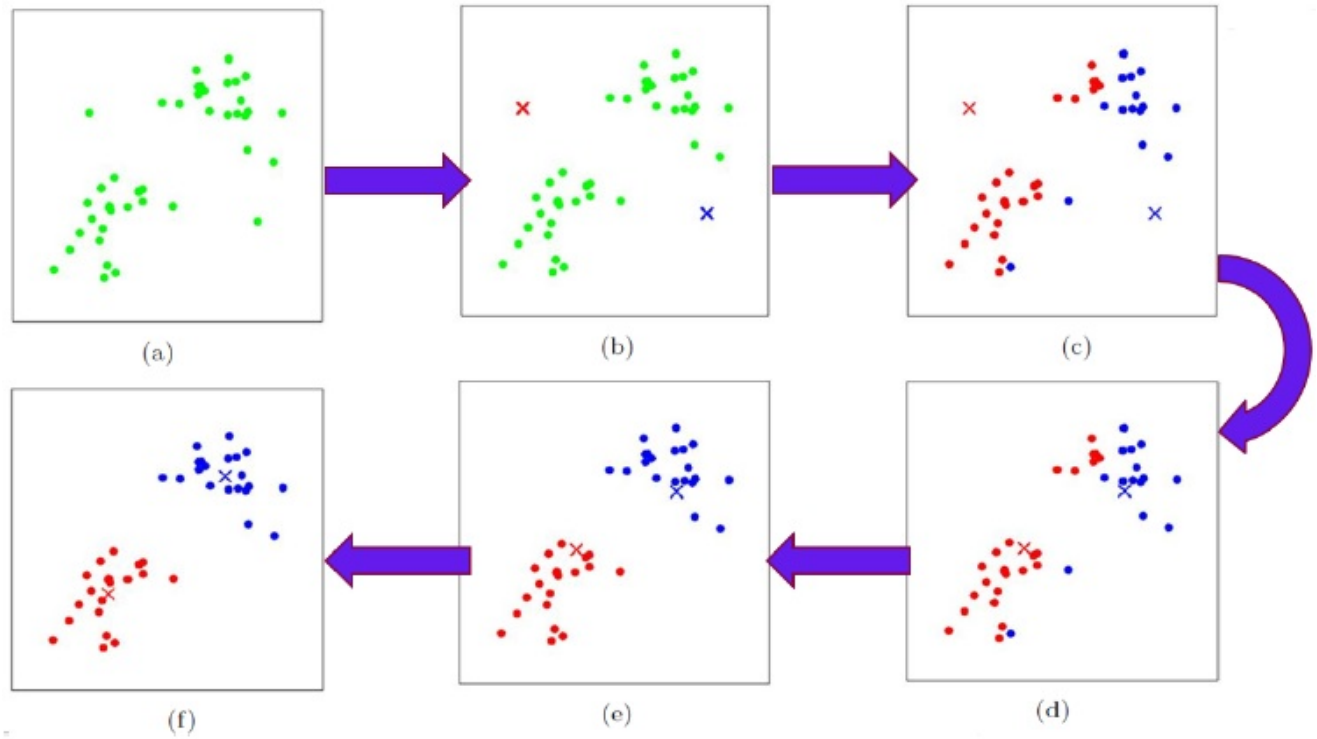
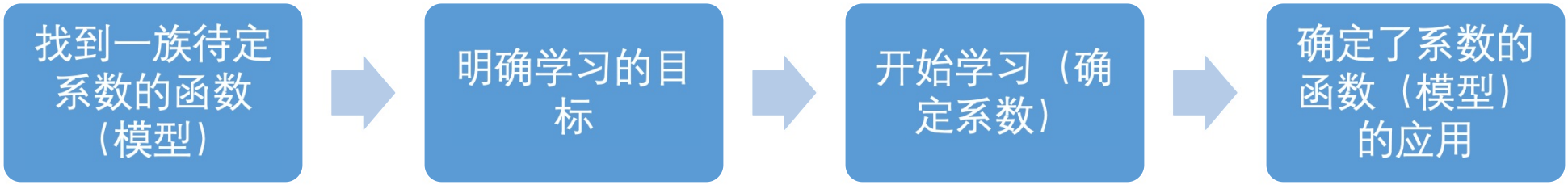


Figure 2: Procedure of K-means

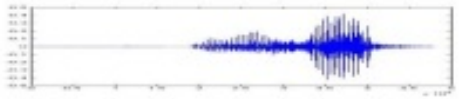
K均值聚类

- 需要人工设定的参数：K值的选取，距离函数
- 典型的无监督学习

典型的机器学习步骤



- Speech Recognition

$f^*($  $)=$“你好”

- Handwritten Recognition

$f^*($  $)=$“2”

- Playing Go

$f^*($  $)=$ “5-5” (step)

- Dialogue System

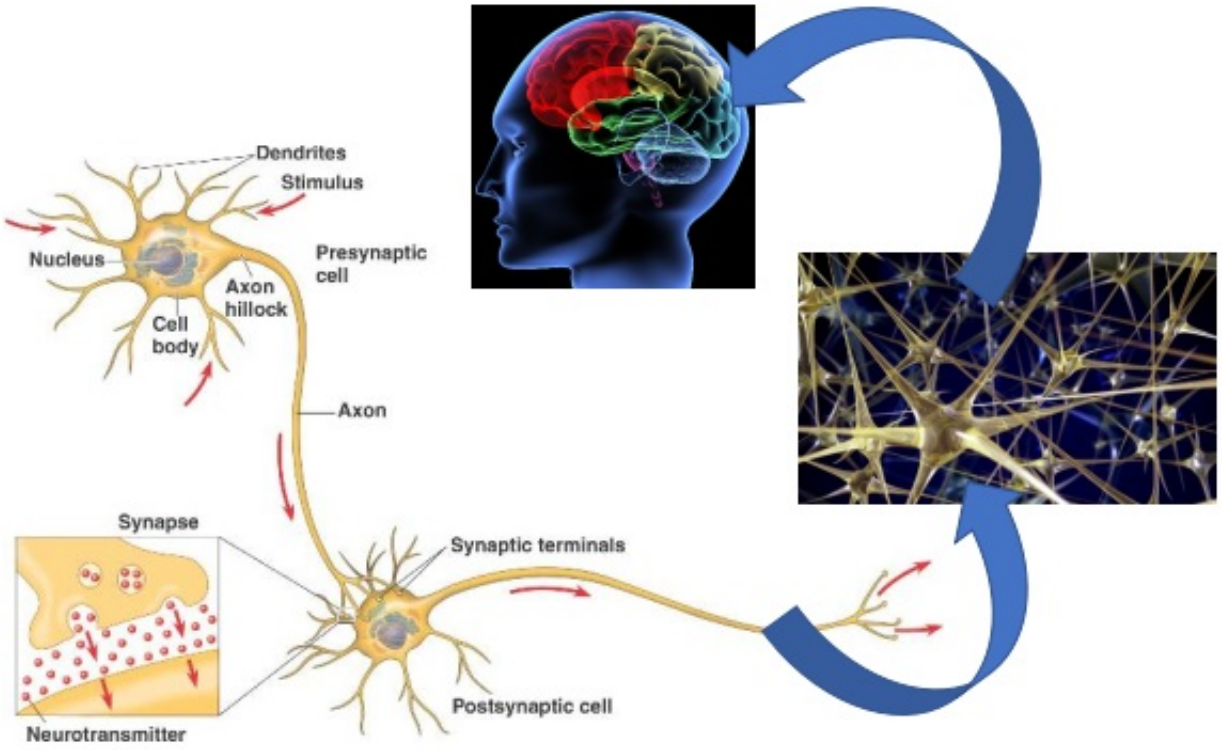
$f^*($ “Hi” $)=$ “Hello”

(what the user said) (system response)

Step 3: Learn!

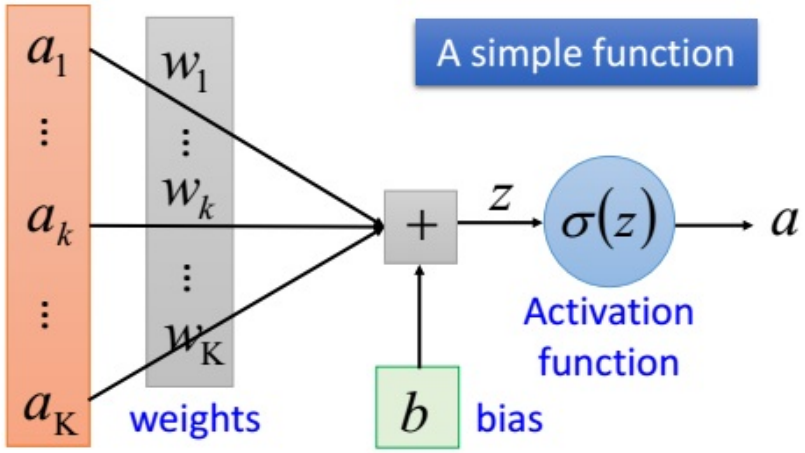
Pick the best function f*

人脑


Dendrites
Stimulus
Nucleus
Presynaptic cell
Axon hillock
Cell body
Axon
Synapse
Synaptic terminals
Postsynaptic cell
Neurotransmitter

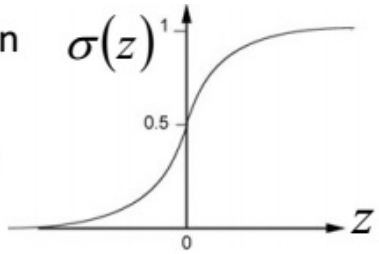
神经元

$$z = a_1 w_1 + \cdots + a_k w_k + \cdots + a_K w_K + b$$

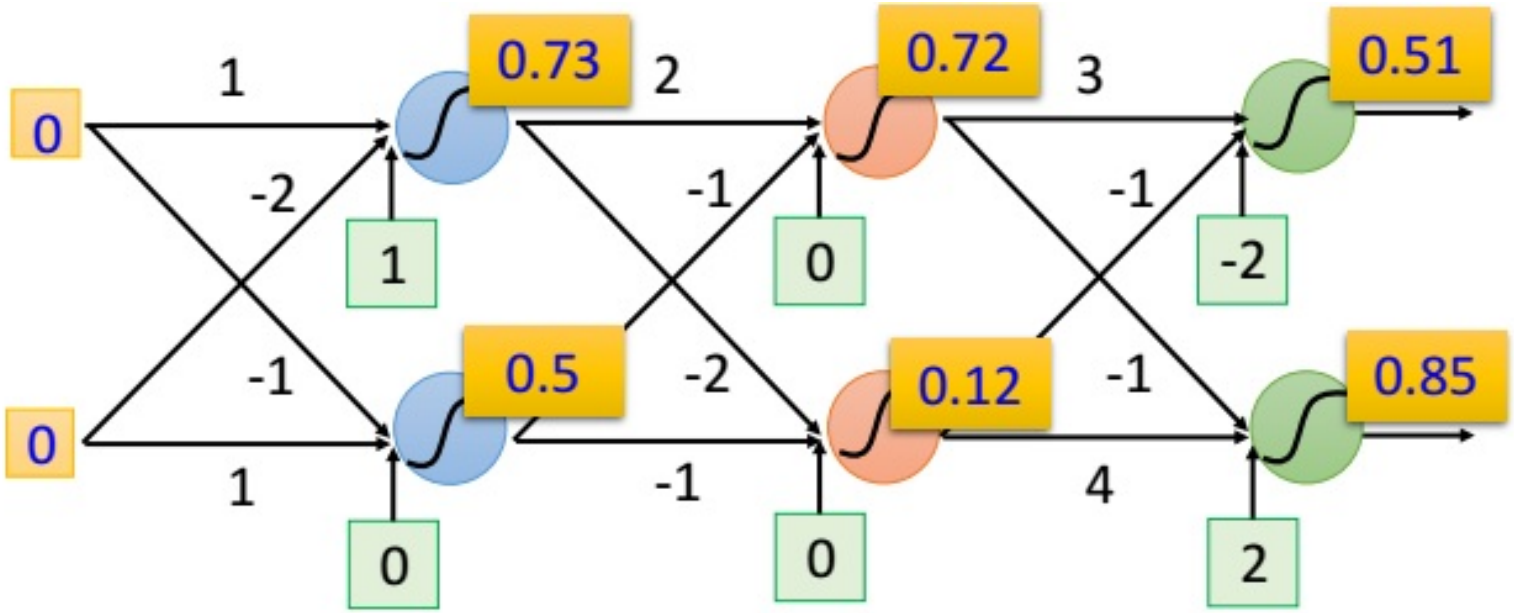


Sigmoid Function

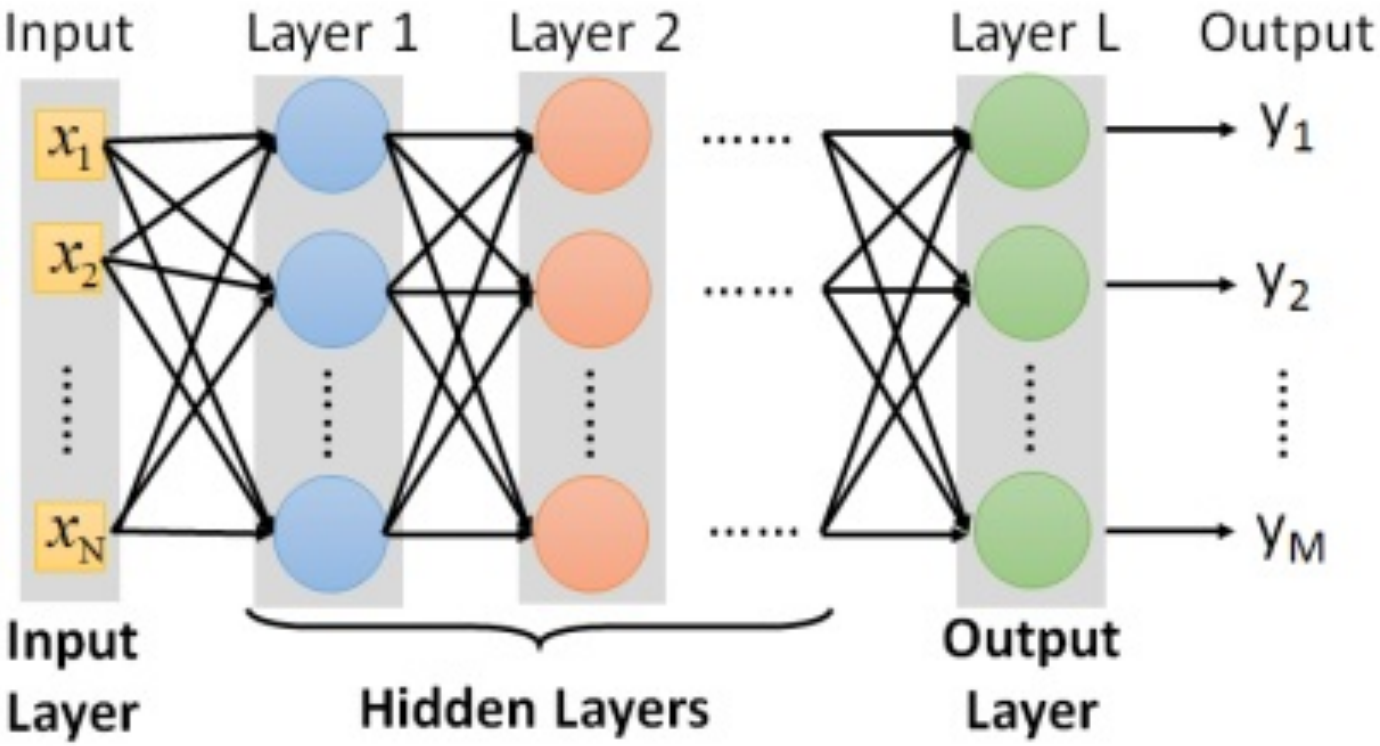
$$\sigma(z) = \frac{1}{1+e^{-z}}$$



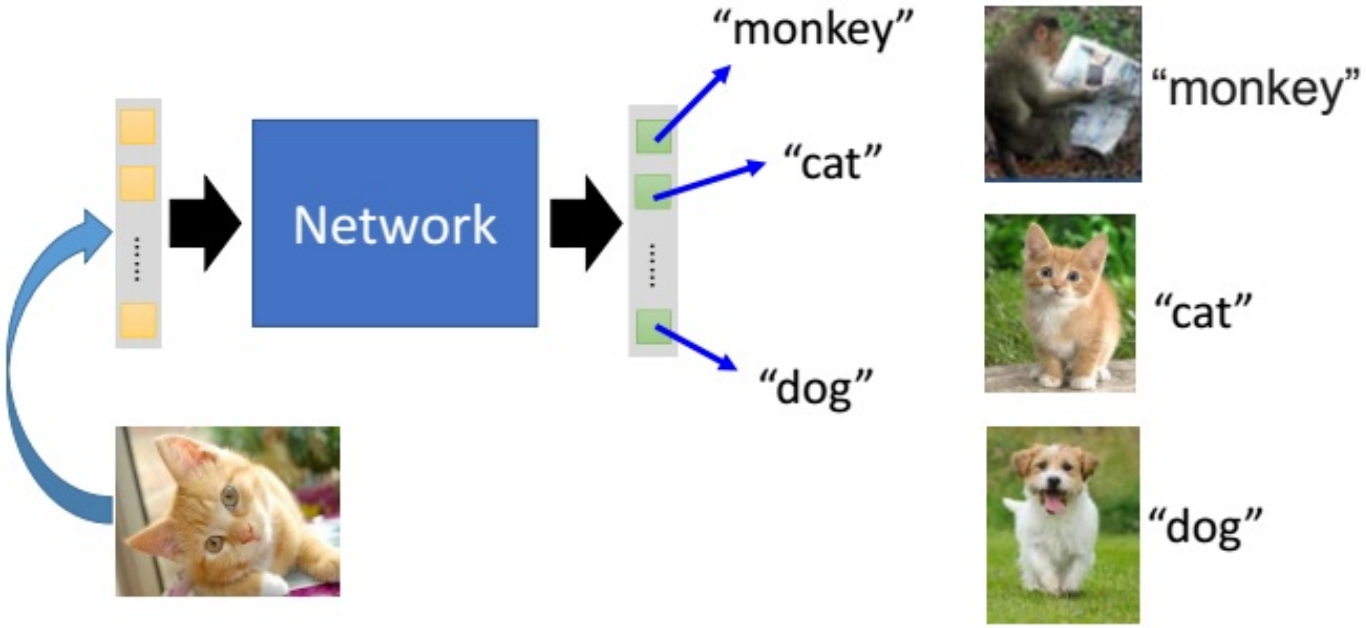
神经网络



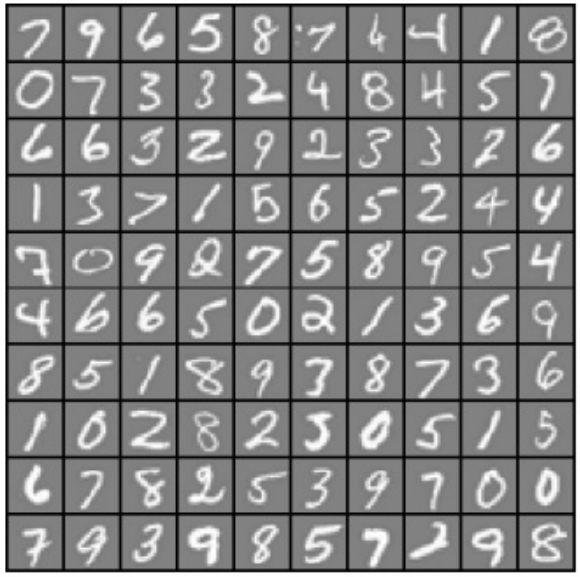
神经网络


Input
Layer 1
Layer 2
Layer L
Output
x_1
x_2
x_N
y_1
y_2
y_M
Input Layer
Hidden Layers
Output Layer

图像识别

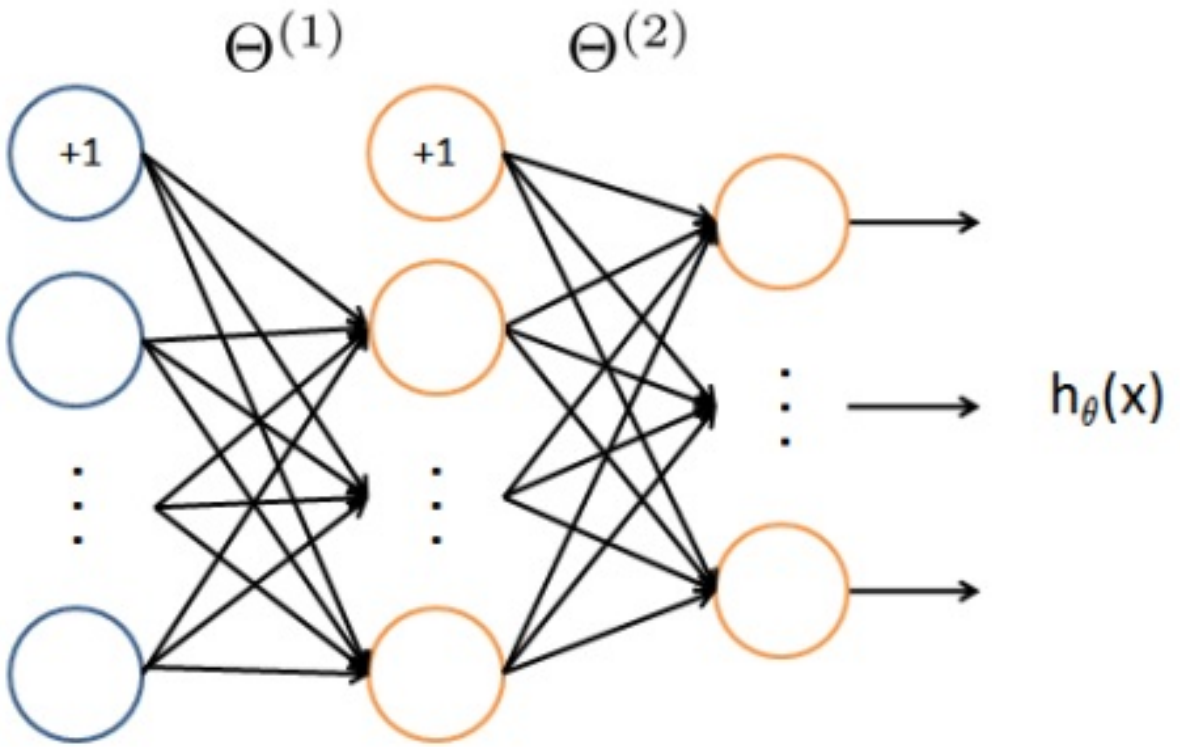


数字识别



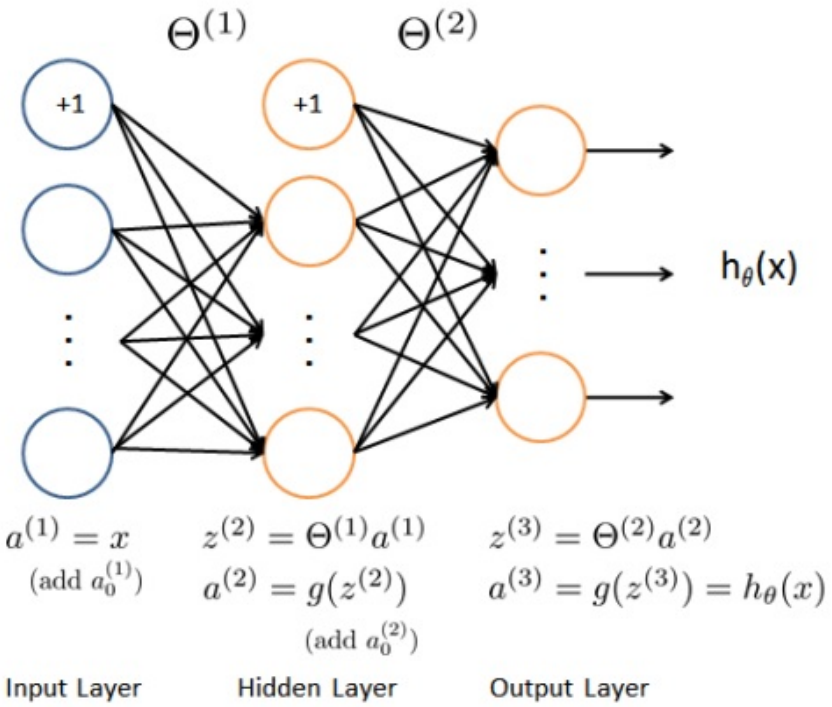
神经网络结构

- 输入层400(像素)
- 隐藏层25
- 输出层10



前向计算

- 随机初始化连接线的权值
- 对每个样本
- 从输入到输出计算，得到结果
- （结果可能是错的）



$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

优化目标

- 通过代价函数来观察算法预测的结果与真实情况的误差有多大
- 找到一组权值，使得误差J最小
- 基本方法，通过反向误差计算偏导数，转换为参数优化问题

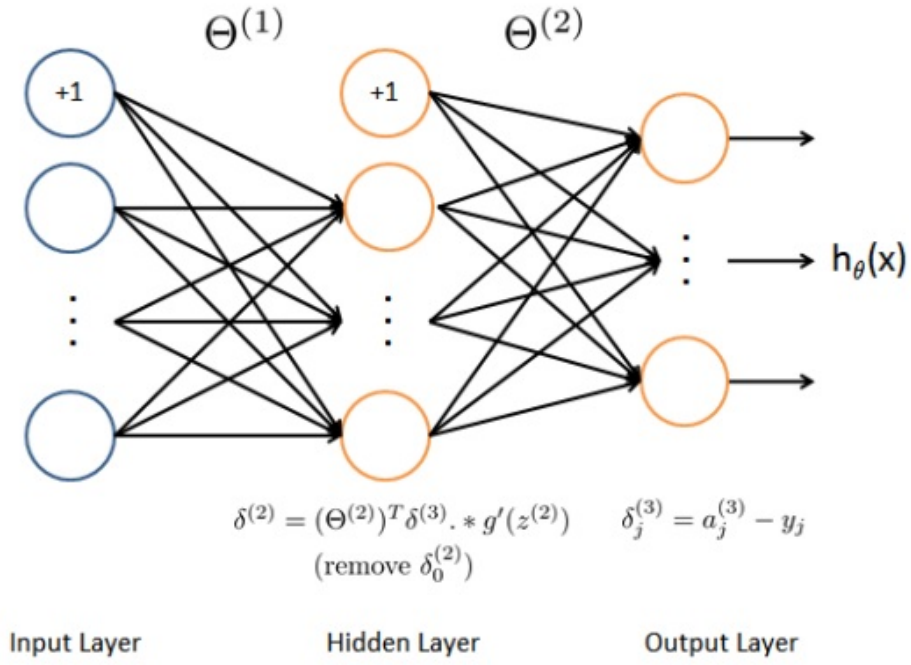
$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

$$D_{ij}^{(l)} := \frac{1}{m}\triangle_{ij}^{(l)} + \lambda\Theta_{ij}^{(l)} \text{ if } j \neq 0$$

$$D_{ij}^{(l)} := \frac{1}{m}\triangle_{ij}^{(l)} \text{ if } j = 0$$

反向传播

- 比较计算结果与实际结果
- 反向计算各个连线权值的误差
- 更新权值



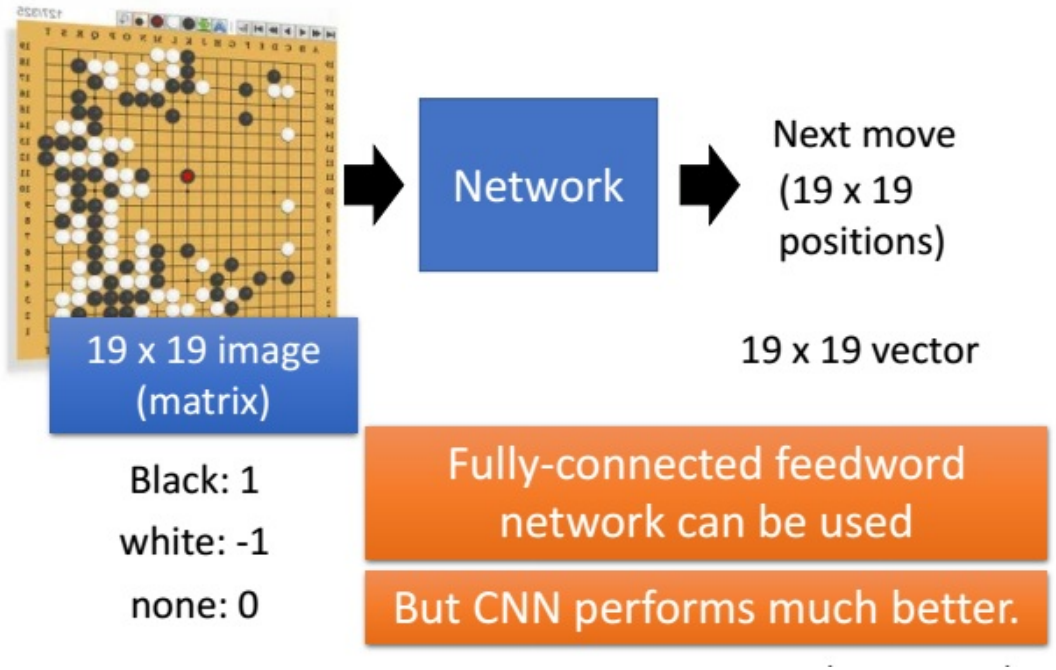
总结

- 1. 参数的随机初始化
- 2. 利用正向传播方法计算所有的 $h_\theta(x)$
- 3. 编写计算代价函数 J 的代码
- 4. 利用反向传播方法计算所有偏导数
- 5. 利用数值检验方法检验这些偏导数
- 6. 使用优化算法来最小化代价函数

总结

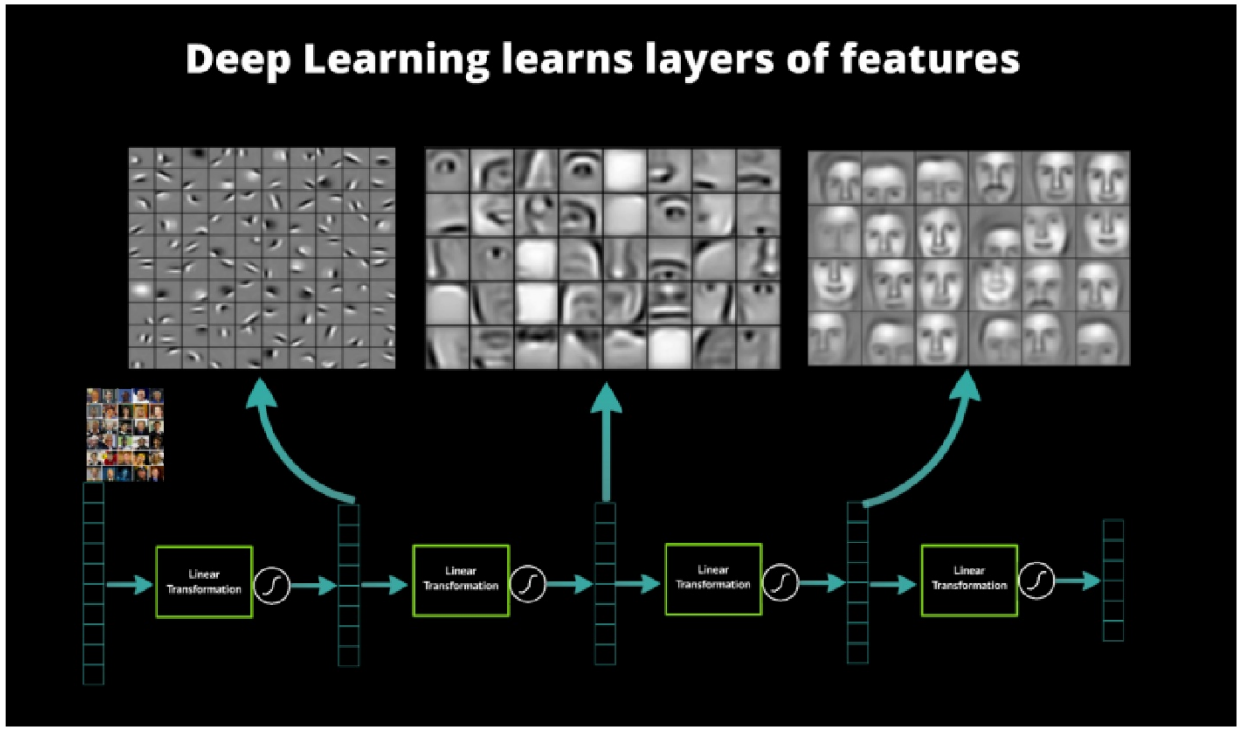
- 神经网络是典型的监督学习算法
- 最典型的应用是分类

Alpha Go


Network
Next move
(19 x 19
positions)
19 x 19 image
(matrix)
19 x 19 vector
Black: 1
white: -1
none: 0
Fully-connected feedword
network can be used
But CNN performs much better.

深度学习

- 更复杂（更多层）的神经网络
- 计算能力要求高
- 类似人的抽象过程？



总结

- 机器学习：利用经验来改善计算机系统自身的性能。
- 经验=数据？
- 非监督学习 与 监督学习
- 聚类与分类

机器学习应用思考

美军想让机器代替人类做这些事

- 美国国防部副部长罗伯特·沃克（Robert Work），2015.12
- 新型人工智能和自主能力的研发是美国国防部第三波“抵消战略”的关键，该战略旨在为美国创造出针对对手的战略性优势。
- 军方到底希望机器人在什么方面替代人类
  - 提醒情报分析人员要特别关注的信息。
  - 进行网络防御电子战以及超视距目标定位。
  - 帮助F-35飞行员确定瞄准和射击对象。
  - 改进飞行以及着陆方式。
  - 驾驶无人机和船只。
- 从2004年至今，美军无人机的轰炸已经在巴基斯坦的某个地区杀死了2500~4000人，其中绝大多数死者都被美国政府归类为“极端分子”。轰炸对象的选择背后，很可能是美国天网计划，采用机器学习算法对5500万民众进行打分。有研究认为这种算法会带来很大的误报率，而0.18%的误报率意味着99000名无辜民众被错误地标记为“恐怖分子”。

总结

- 机器学习并不神秘
- 需要分析数据的地方，机器学习就可能派上用场

- 谢谢