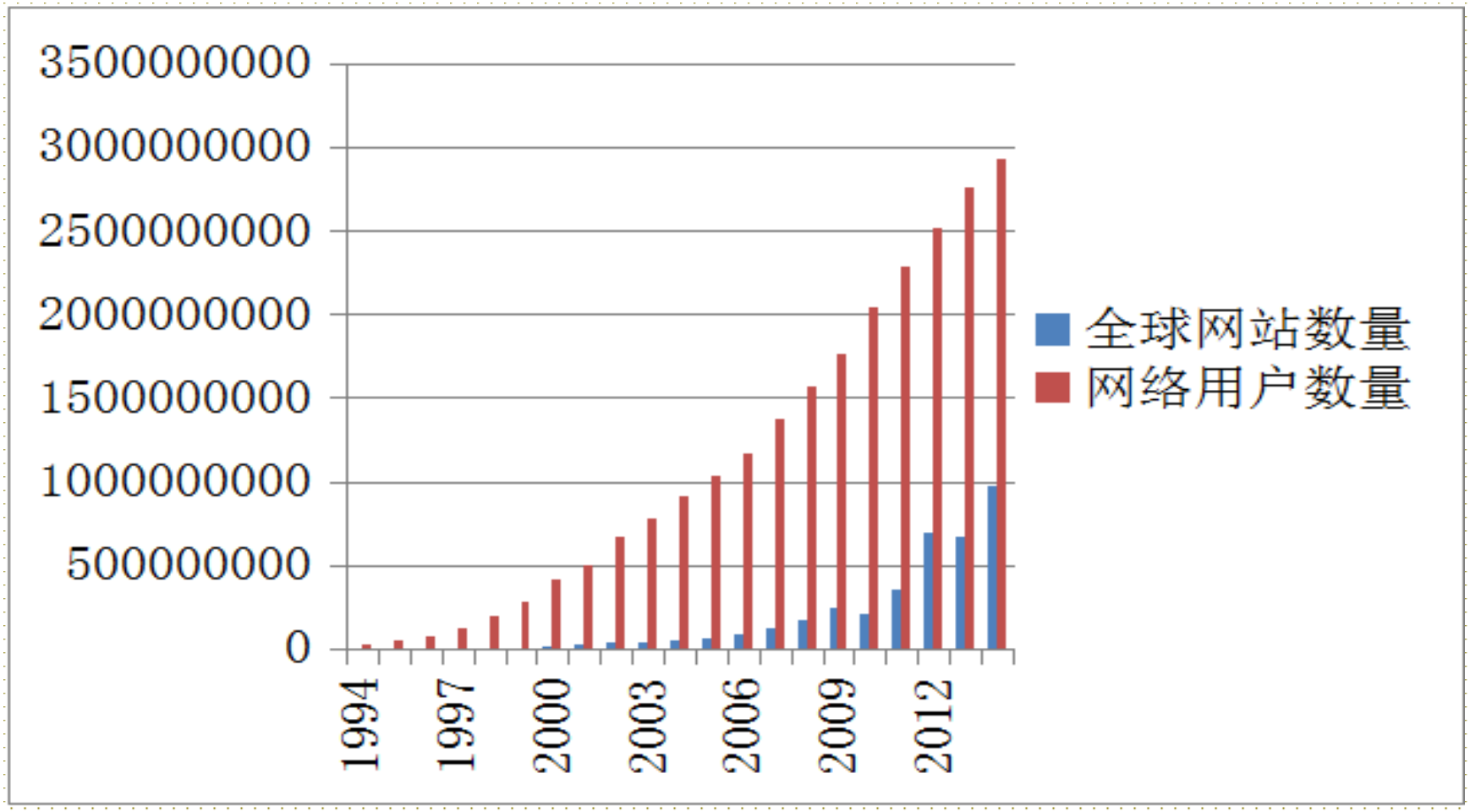
机器学习: 统计与计算之恋

张志华

May 2016

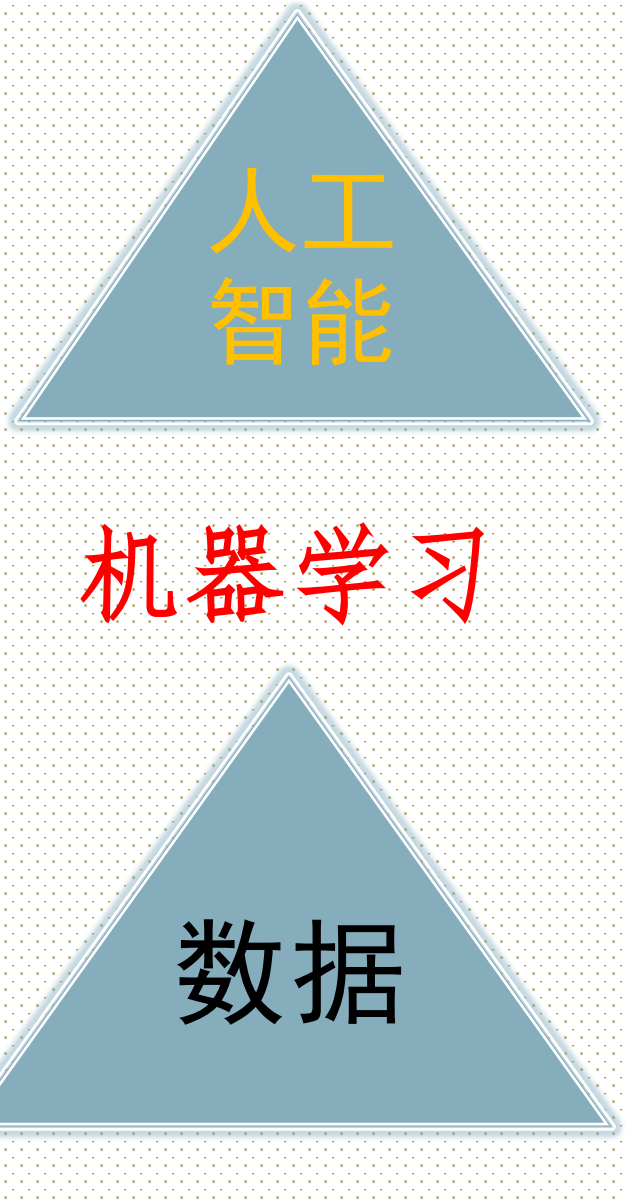
机遇与挑战：快速增长的数据规模



Alpha-Go

- **简单监督神经网络**: 快棋手
- **复杂监督神经网络**: 慢棋手
- **强化学习神经网络**：左右互博：最大化最后的奖励
- **值网络**（棋感）：两名强化棋手赛一局后抽一个[状态，下步棋位]和得分组成[数据，标签]
- **下棋策略**：
  - **模拟**：通过下一步棋，使得最终获胜的概率尽可能大
  - **更新**：更新值网络

- 回顾与思考
- 几个简单的研究思路


人工
智能
机器学习
数据
统计
计算

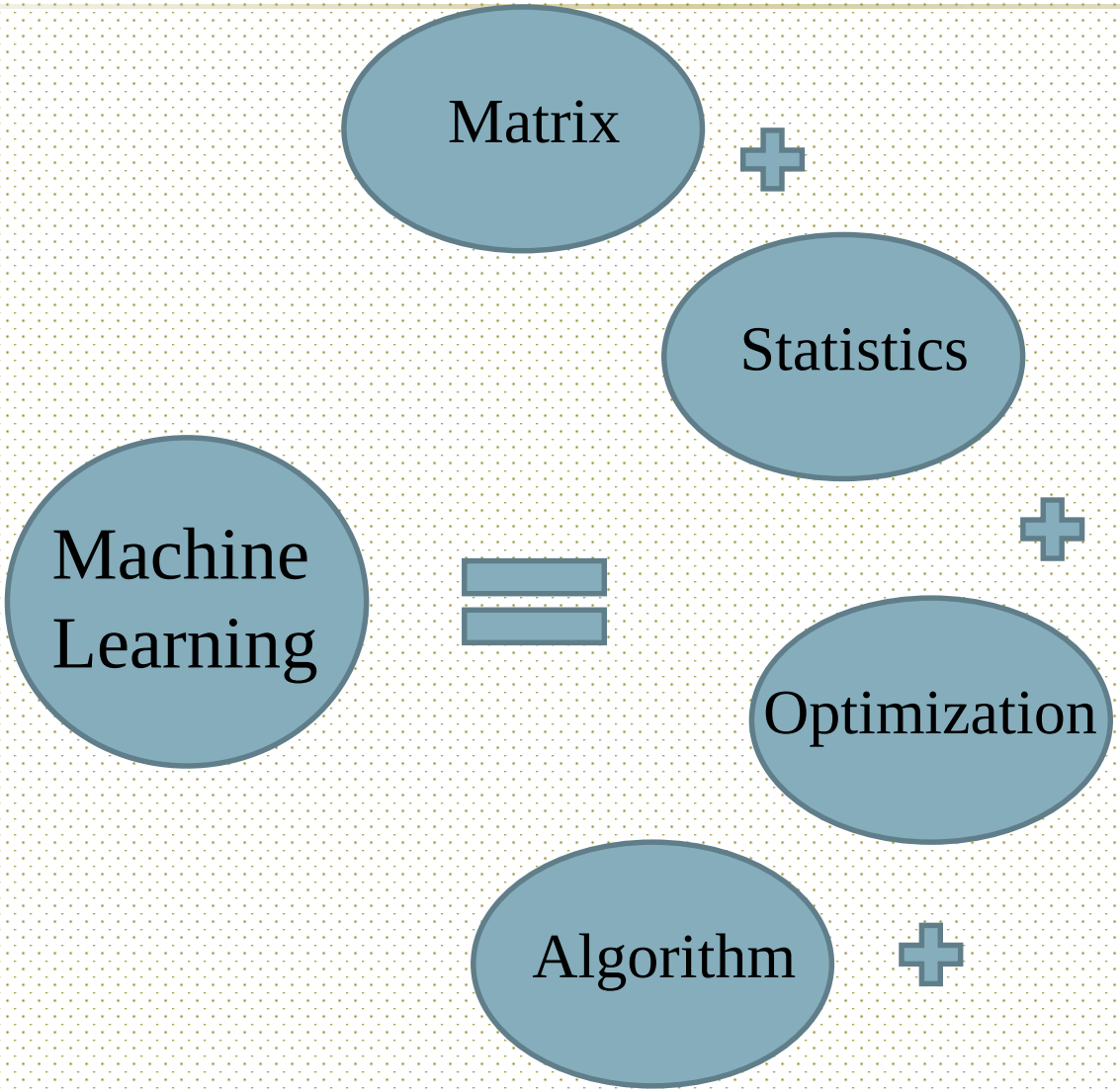
机器学习

“It is one of today’s rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science.”

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 2015.

机器学习

机器学习或统计机器学习



机器学习 三个层次

初级(Low Level)	• 数据获取和特征提取

中级(Middle Level):	• 数据处理与分析
	  • 应用问题导向(Data Mining)
	  • 方法与算法 (Machine Learning)
	  • 推理

高级(High Level):	• 智能与认知

机器学习发展几个关键期

上世纪90年代平淡时期。

1996年 --- 2006年黄金期：

SVM，Boosting, Kernel Method,Lasso, etc.

手写数字识别

点击率预估

股票预测

2006年 --- 2009年徘徊期。

机器学习发展几个关键期

- 成为计算机科学和人工智能的主流学科。

美国院士

- Mike Jordan
- Tom Mitchell
- Jerome Friedman
- Daphne Koller
- Robert Tibshirani
- Robert Schapire
- Bin Yu
- Larry Wasserman
- Stephen Boyd

图灵奖

- 2011年 Judea Pearl
- 概率和因果性推理演算法

当下热点

- 深度学习
- Alphago
- 无人驾驶汽车
- 人工智能助理

工业界的视角

- 微软模式到谷歌模式：从制造到服务。
- “Big Data”和数据驱动。
- 深度学习在计算机视觉、自然语言理解、语音识别、智力游戏等领域的颠覆性成就。

数据科学

- 数据科学 Data Science
  - 计算机科学、数值分析、现代数据分析等的交叉学科；
  - 目的是从数据中获得知识，获得有价值的信息，服务社会；
  - Data Scientist or Data Engineer应具备三个条件：底层架构开发或使用能力 (Spark, MapReduce or Hadoop); 程序开发能力；数学建模和解决问题能力。

统计学界的视角

- Statisticians thought that computer scientists **were reinventing the wheel**.
- Computer scientists thought that statistical theory **didn't apply** to their problems.
- Statisticians now recognize that computer scientists are **making novel contributions**
- Computer scientists now recognize the **generality** of statistical theory and methodology

Larry Wasserman
美国科学院院士
All of Statistics

统计学的视角

- “Using fancy **tools** like neural nets, boosting, and support vector machines **without** understanding basic **statistics** is like doing brain surgery before knowing how to use a band-aid.” **Larry Wasserman**
- 这是为什么学术界对深度学习仍存疑虑？

计算机科学与统计

- Statisticians and Computer Scientists
  - Computer Scientists: Computing and Intuitive Ability
  - Statisticians: Modeling and Theoretical Analysis
- Examples
  - Boosting, Support Vector Machines (SVMs) and Sparse Modeling
  - Kernel Principal Component Analysis (KPCA)and Multidimensional Scaling (MDS); ISOMAP, LLE, NMF published in Science and Nature

统计与计算机科学

- 成为统计学的一个主流方向；许多著名统计系纷纷招聘机器学习领域的博士为教员。
- 计算在统计已经变得越来越重要：传统多元统计分析是以**矩阵**为计算工具，现代高维统计则是以**优化**为计算工具。
- 计算机学科开设高级统计学课程，比如统计学中的核心课程“经验过程”。

计算机科学界的思考

- 计算机科学发展的三个阶段（Foundation of Data Science, by Avrim Blum, John Hopcroft and Ravindran Kannan）
  - 早期：让计算机可以工作。发展重点在于程序语言、编译原理、操作系统、以及支撑它们的数学理论；
  - 中期：让计算机变得有用。发展重点在于算法和数据结构；
  - 当今：让计算机具有多的应用。发展重点从离散类数学转到概率和统计。

计算机科学界的思考

机器学习起着核心作用，戏称为全能学科(universal discipline)。

机器学习 三个层面

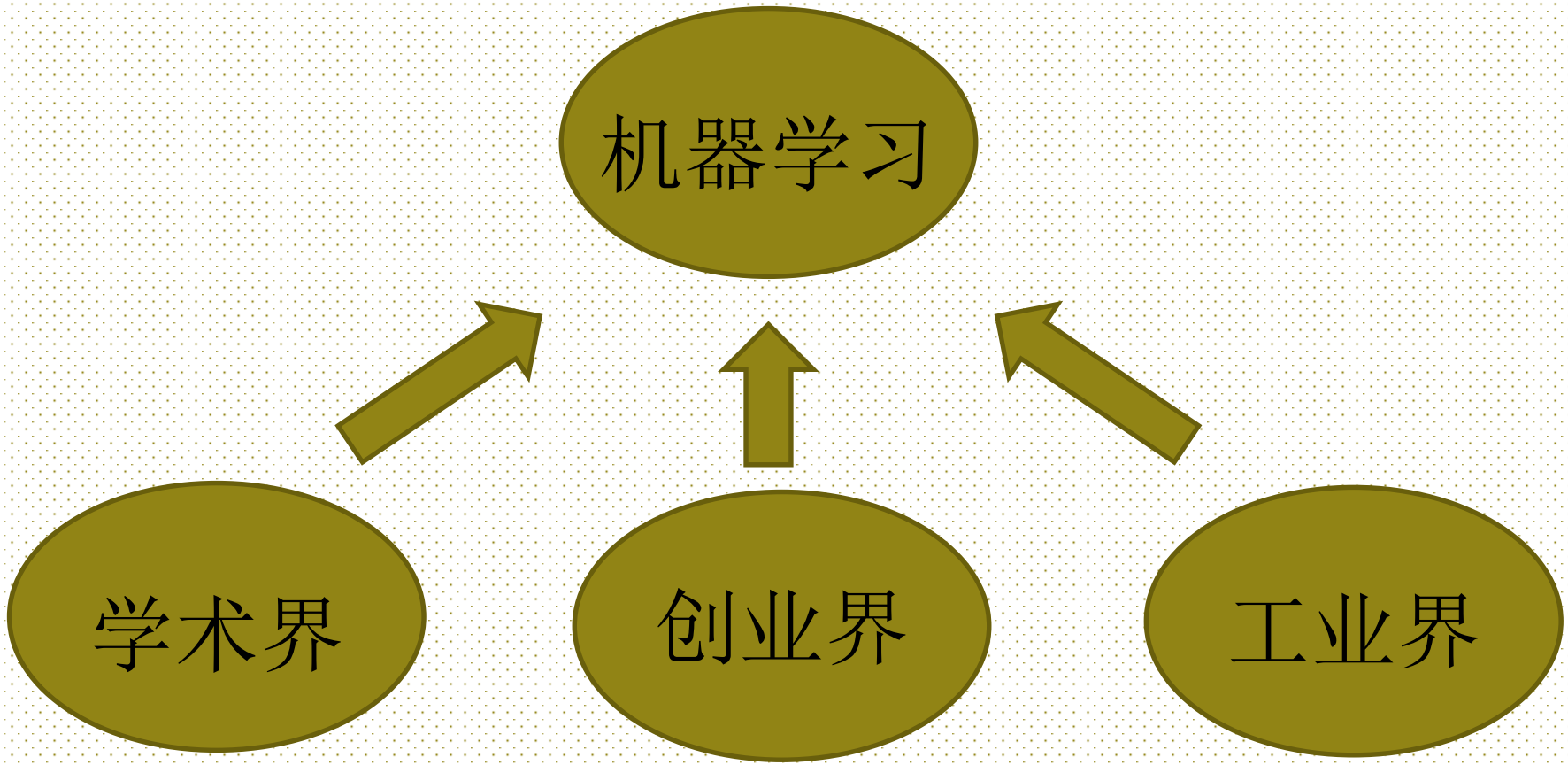
独立学科构架	• 研究自身的理论基础
研究工具	• 为应用学科提供解决问题方法与途径。
研究源泉	• 为统计、理论计算机、运筹优化等学科提供新的研究问题。

机器学习发展启示

- 理想与务实
  - 时髦概念转为理论基础的建立，特别是树立了统计学作为学科的理论基础。
  - 以数据和问题为驱动 ---- 对工业界发展的支撑。
- 多学科的交叉：

  是深度融合，而不是名词堆砌。
- 兼容并收。

机器学习发展启示



国际发展现状

Berkeley	Stanford	CMU	MIT	University of Toronto
• 统计系和计算机系共同构成的AI研究 • 强调统计和计算机的融合 • 统计机器学习理论方面的研究处于国际领先地位 • Professor Mike Jordan --- 统计机器学习主要推动者、建立者和传播者	• 不确定领域 • 概率图模型 • 概率机器人 • 统计学习	• 第一个建立Machine Learning系的学校； • 贝叶斯统计学是世界的研究中心 • 贝叶斯与计算科学的交叉研究处于国际领先地位 • Professor Tom Mitchell --- 早期机器学习主要建立者	• CSAI和Media实验室在信息学科的理论和应用方面的创新无与伦比的 • 统计学与信息理论一直是这两个实验室创新的理论源泉。 • JMLR杂志的发起地	• 机器学习发表在 “Science” and “Nature” 的论文大多来自该校Machine Learning Group。 • Professor Geoffrey E. Hinton --- 伟大思想家，神经网络学派 建立者之一。

国内现状－个人看法

- 统计学处于两个极端
  - 作为数学的一个分支，主要研究理论问题；
  - 作为经济学的一个分支，主要研究经济分析中的应用。
- 面向于数据处理、分析的工业统计学的深度研究有巨大潜力。

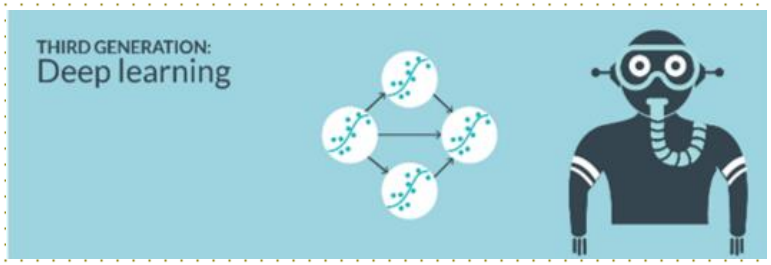
国内现状－个人看法

- 机器学习的现状
  - 得到了广泛的关注，也取得一定的成就。但更缺少高品质的成果。
  - 许多高校都开设机器学习课程。但大多讲座性质；系统性、原理性的课程缺少。
- 统计学与计算机科学还处于各自为战。
- 面向于大数据的统计学与计算机科学的交叉学科统计机器学习是机遇也是挑战。

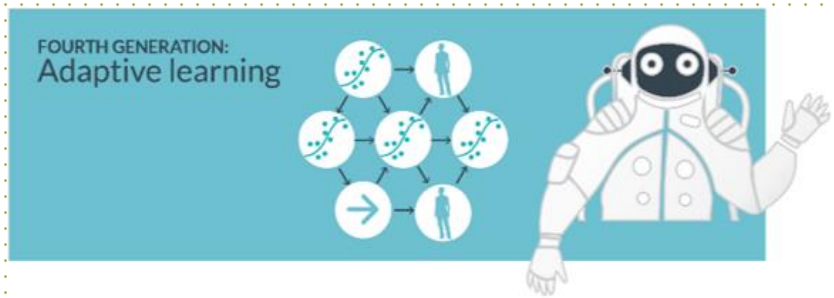
- 回顾与思考
- 几个简单的研究思路

基于规则

SVM、决策树、 kNN等



深度学习



自适应学习

LDA

CUR Decomposition

Alternating Least Squares

Shortest Path

Conditional Random Fields

K-core Decomposition

Dirichlet Process

SVD

Max-Product Linear Programs

新算法的图结构建模和并行化

PageR

pagation

Noise elimination

Non-negative MF

Proximal Message Passing

Triangle Counting

ADMM

Gibbs Sampling

Graph Coloring

Deep Learning

Hierarchical
多级

隐含数据模型

- 隐含数据模型(Latent Data Models)
  - 作为概率图模型的一种延伸，它是一个重要的多元数据分析工具。
  - 隐含变量有三个重要的性质：
    - 比较弱的条件独立相关代替较强的独立相关；

隐含数据模型

- de Finetti Representation Theorem:

  A set of random variables are exchangeable if and only if they can be expressed as a mixture of a set of independent and identically distributed random variables, conditioned on some parameter.

$$\theta \sim p(\cdot);$$

$$[x_1, \ldots, x_n] \sim p(\cdot \mid \theta).$$

- Gaussian Scale Mixture

隐含数据模型

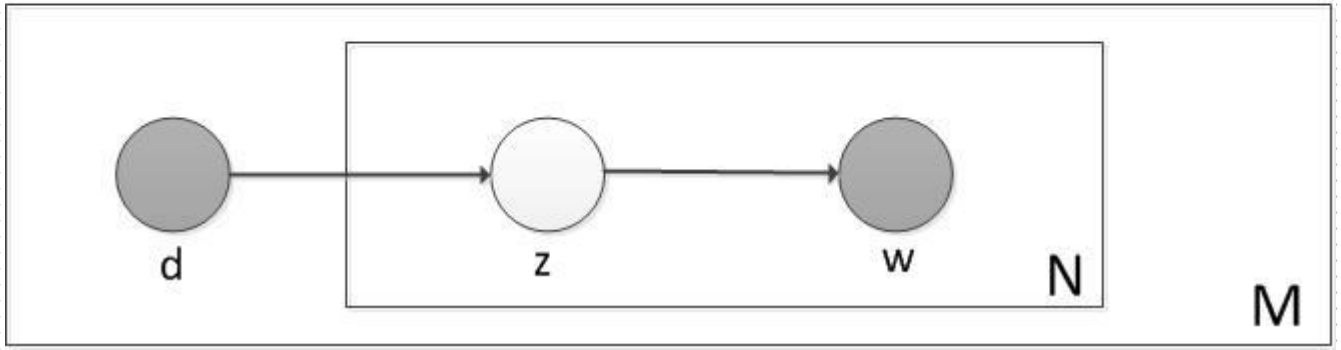
- 隐含数据模型(Latent Data Models)
  - 作为概率图模型的一种延伸，它是一个重要的多元数据分析工具。
  - 隐含变量有三个重要的性质：
    - 比较弱的条件独立相关代替较强的独立相关；
    - 蕴含某种好的物理意义；

隐含数据模型

Probabilistic latent semantic indexing (pLSI)

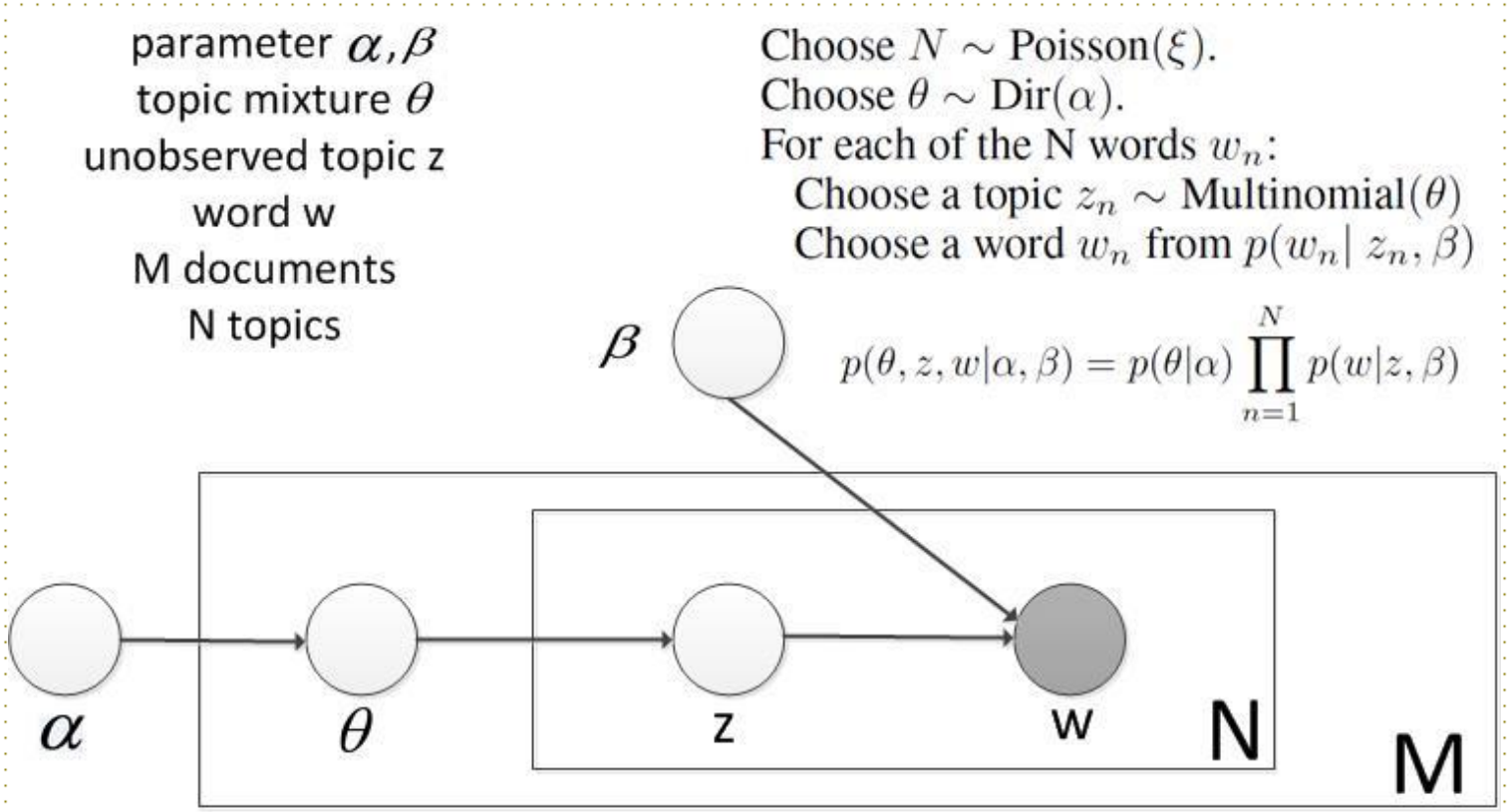
document label d
unobserved topic z
word w
M documents
N topics

$$p(d,w) = p(d)\sum_z p(w|z)p(z|d)$$



隐含数据模型

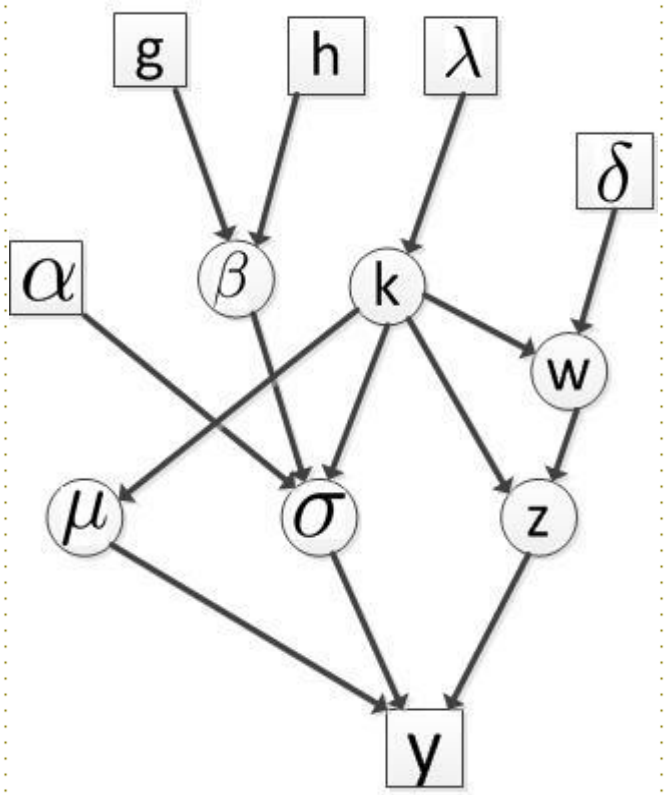
Laten Dirichlet Allocation (LDA)

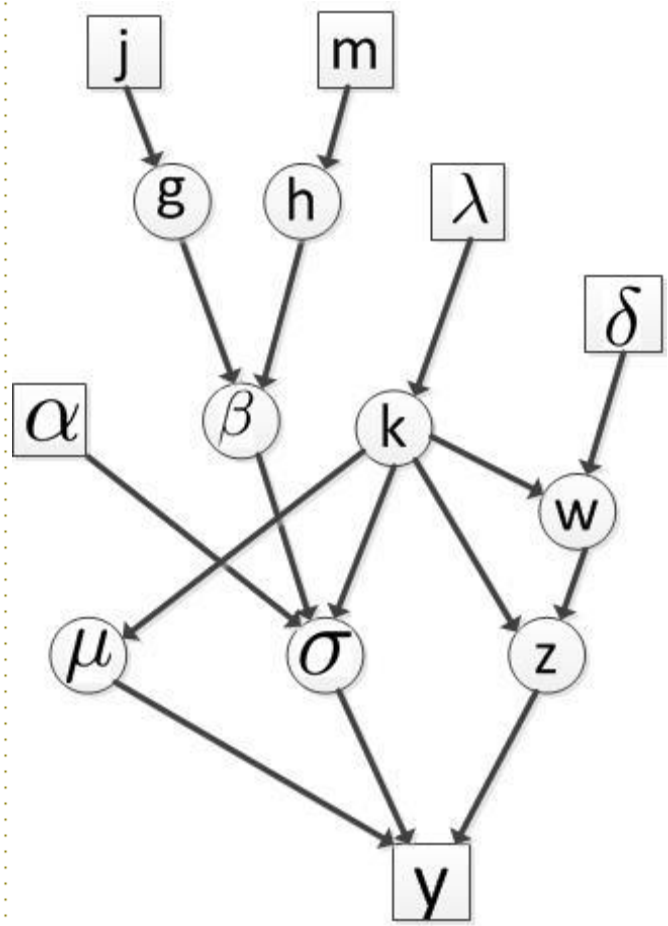


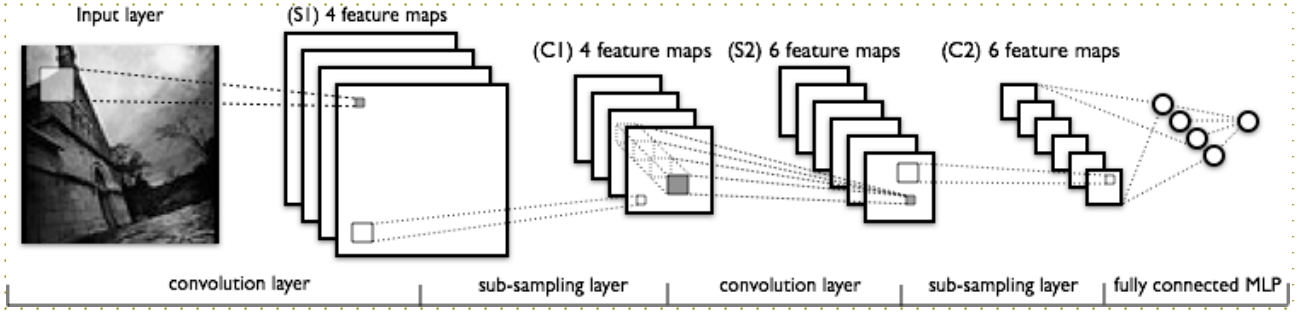
隐含数据模型

- 隐含数据模型(Latent Data Models)
  - 作为概率图模型的一种延伸，它是一个重要的多元数据分析工具。
  - 隐含变量有三个重要的性质：
    - 比较弱的条件独立相关代替较强的独立相关；
    - 蕴含某种好的物理意义；
    - 方便我们设计高效的计算方法(Expectation Maximization 算法和Data Augmentation思想)。

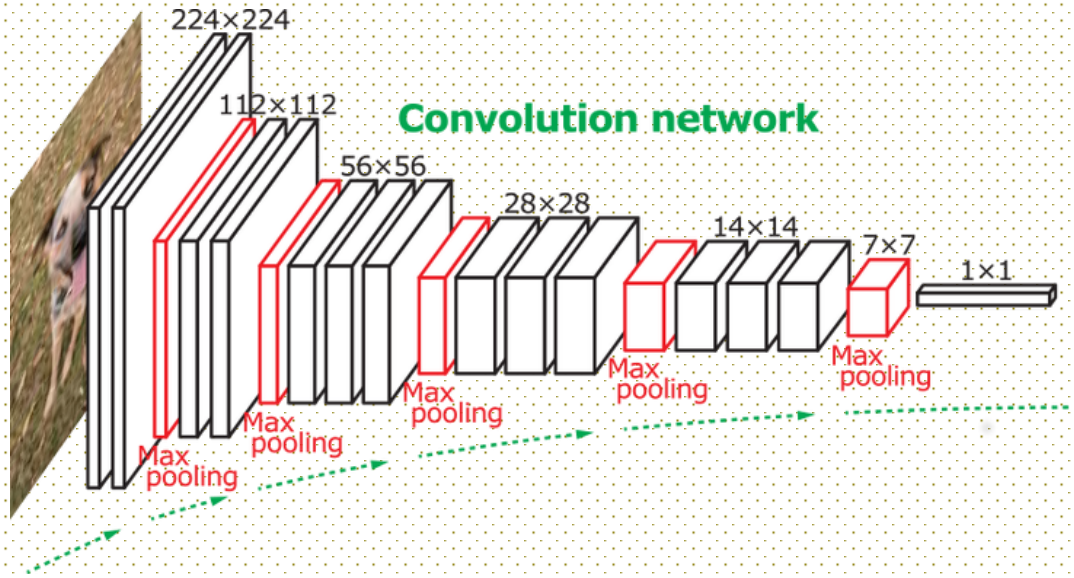
Hierarchical Bayesian Model



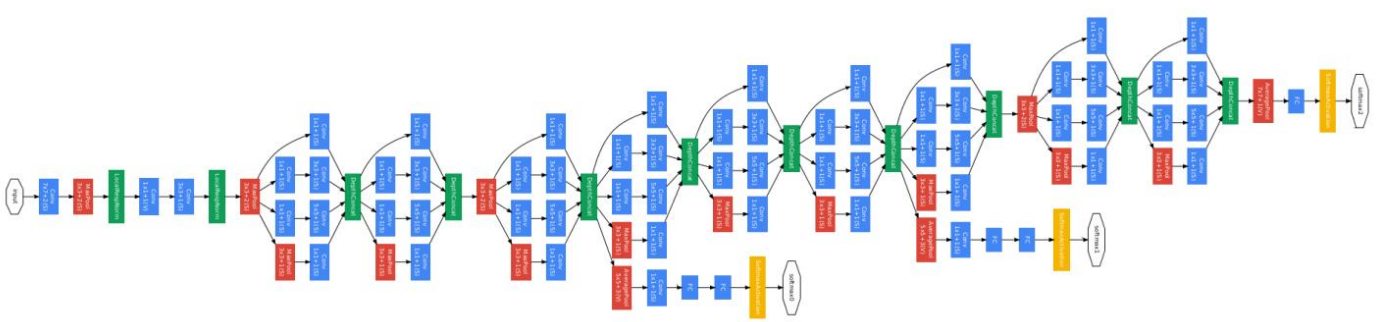




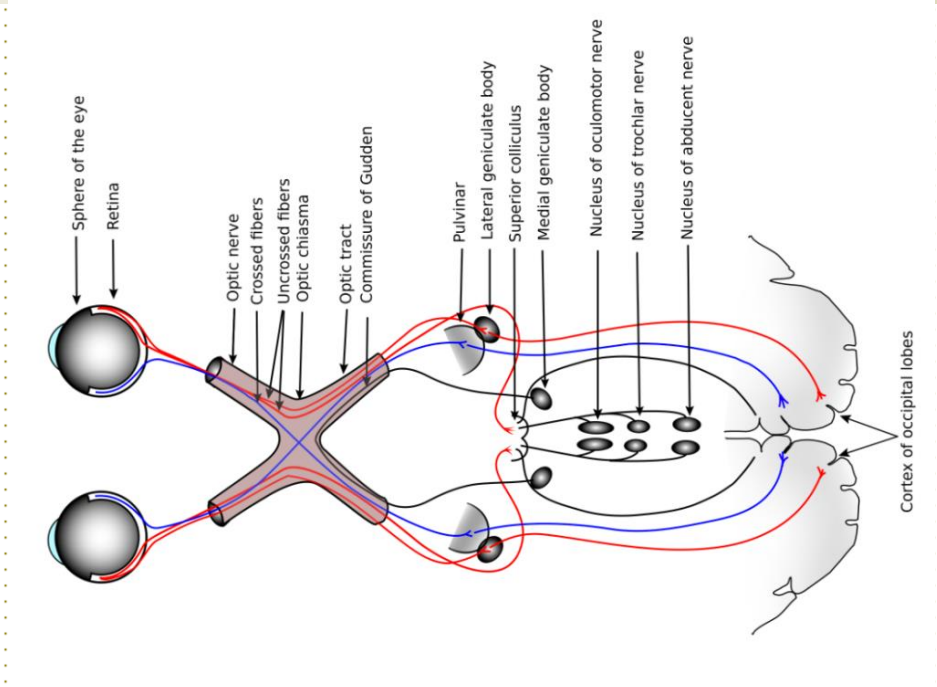
5 layer LeNet: digital classification



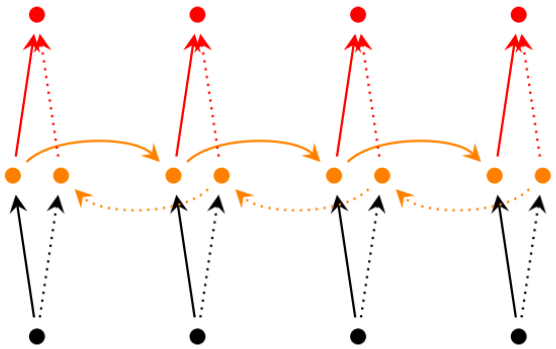
16 layer VGG: 7.4% top 5 classification error in ImageNet ILSVRC-2012



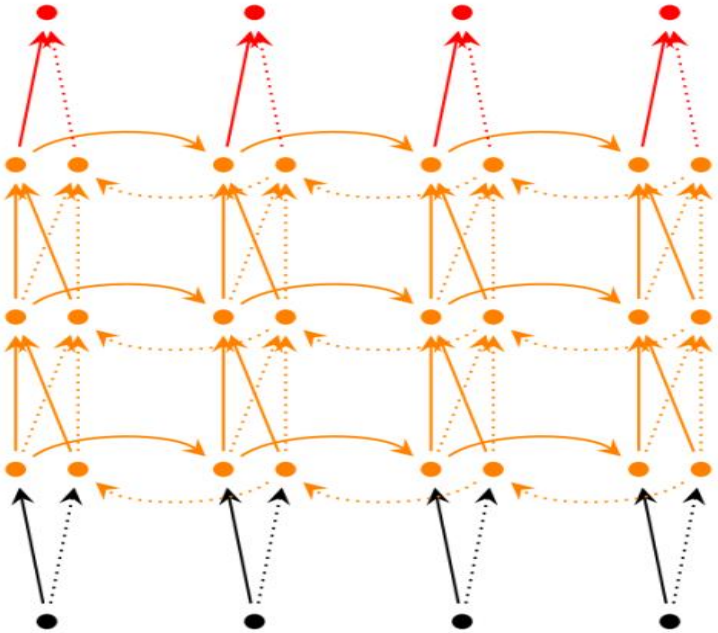
22 layer GoogleNet: 6.7% top 5 classification error ImageNet ILSVRC-2014



Mammal visual system



BiRNN



Deep BiRNN: boost translation performance 10%

自适应
Adaptive

Adaptive Importance Sampling

Algorithm 2 Population Monte Carlo Algorithm

1: **for** $t = 1 : T$ **do**
2: **for** $t = 1 : N$ **do**
3: generate proposal distribution $q_i^{(t)}(\theta) = K(\theta|\theta_i^{(t-1)})$.
4: sampling $\theta_i^{(t)} \sim q_t(\theta) = K(\theta|\theta_i^{(t-1)})$ and compute the weight $w_i^{(t)} = \pi(\theta_i^{(t)}|y)/q_i^{(t)}(\theta)$.
5: **end for**
6: Normalizing $w_i^{(t)}$ to sum up to 1.
7: Resample $\theta_i^{(t)}$ using $q_t(\theta) = w_i^{(t)} K(\theta|\theta_i^{(t)})$, create the sample $(\theta_1^{(t)}, \ldots, \theta_N^{(t)})$ with equal weight.
8: **end for**

Population Monte Carlo (JCGS, 2004)

自适应列选择(Adaptive Column Selection)

- 给定矩阵A，找其部分列构成矩阵 C,用$CC^{+}A$近似A。
- 先从矩阵A中采出部分列，组成C_1；
- 再从残差$A - C_1C_1^{+}A$中采出列，组成C_2
- $C = [C_1\ C_2]$

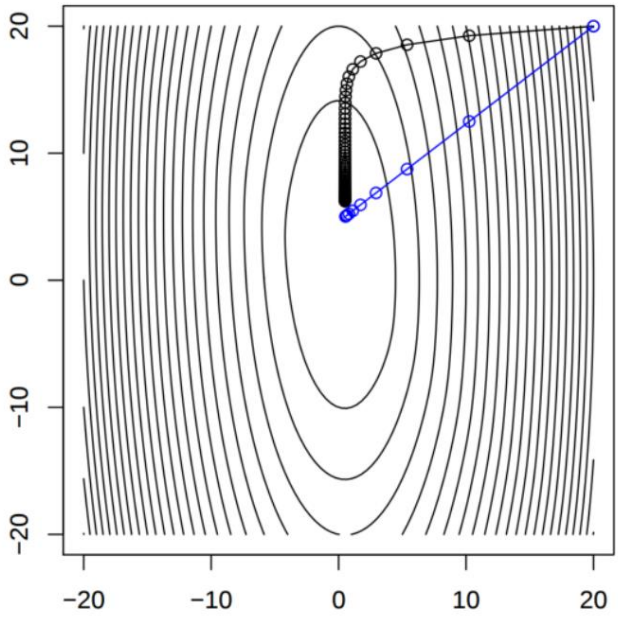
Matrix Approximation and Projective Clustering via Volume Sampling, Theory of Computing, 2006

AdaSDCA: 自适应的SDCA算法

- 解决带正则项的ERM（经验风险最小化）问题
- 固定的概率分布→ 跟dual residue 相关的概率分布
- Stochastic Dual Coordinate Ascent with Adaptive Probabilities, ICML 2015

Adaptive Regularization

- Adaptively adjusts to data geometry
- Adaptively determines learning rates for different dimensions



- Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, JMLR 2011

Adaptive Prediction

- 对在线凸优化问题加速
- Adaptive Regularization + Predictable Sequences = Adaptive Prediction
- Accelerating Online Convex Optimization via Adaptive Prediction, JMLR 2016

自适应增强（Adaptive Boosting）

- 预测错误的样本点 → 权重↑
- 预测正确的样本点 → 权重↓

- 多个弱分类器增强为强分类器

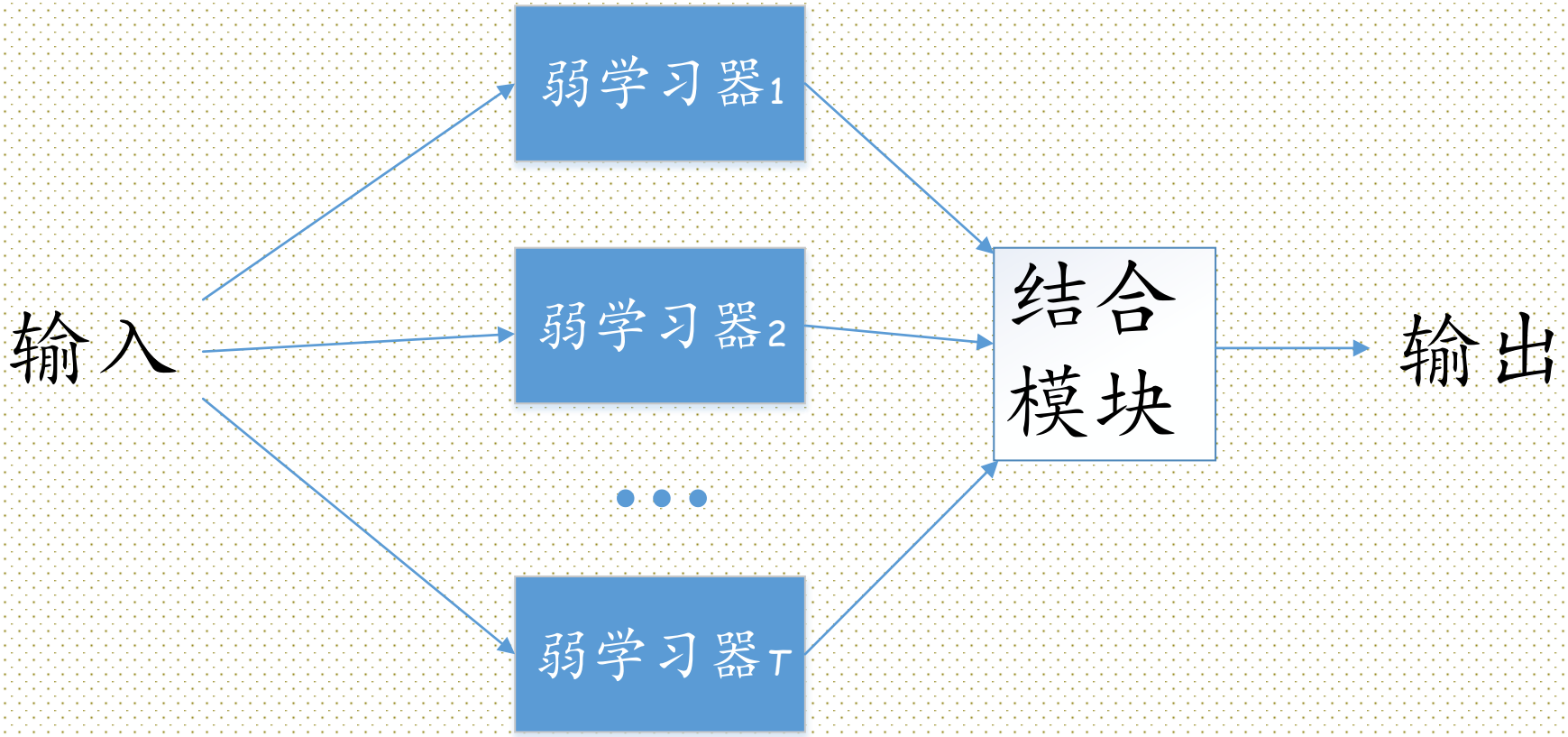
- Optimal and Adaptive Algorithms for Online Boosting, ICML 2015 best paper

Averaging
平均

基本思想

集成

- **集成学习(Ensemble Learning)**
  - 构建并组合多个学习器来完成学习任务

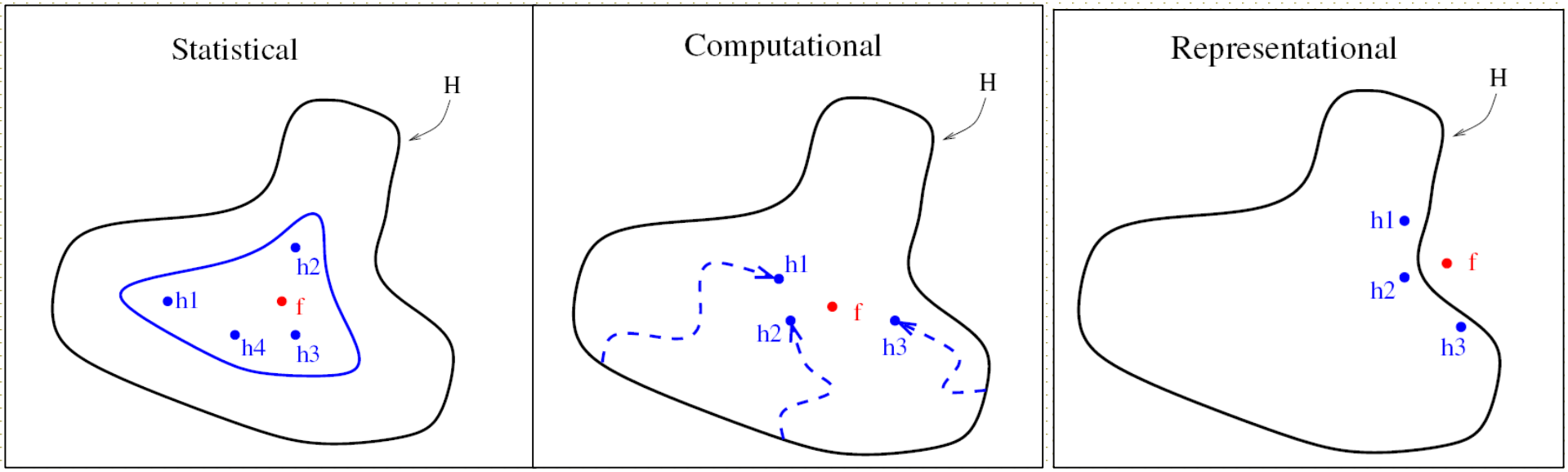


集成

优势

集成学习 vs. 单个学习器（Dietterich, 2000）

f表示真实的假设



统计—降低过拟合风险　计算—降低陷入局部极小点风险　表示—扩展假设空间

由于假设空间很大，可能存在多个假设在训练集上达到同等性能，使用单学习器可能导致泛化性能不佳，但多个学习器可以减小这一风险。

学习算法往往会陷入局部极小点，有的局部极小点对应的泛化性能不佳，而多次运行后进行结合，可以减小这个风险。

真实假设可能不在学习算法考虑的假设空间内，此时使用单学习器肯定无效，而结合多个学习器可以使假设空间扩大，可以学得更好的近似。

集成

生成策略

- **假设生成的弱学习器相互独立**
  - 对于二分类问题$y \in \{-1, +1\}$和真实函数f，假设弱学习器h_i的错误率为ϵ
  - 采用简单投票法

$$H(x) = sign(\sum_{i=1}^{T} h_i(x))$$

  - 集成的错误率

$$P\big(H(x) \neq f(x)\big) = \sum_{k=0}^{\lfloor T/2 \rfloor} \binom{T}{k} (1-\epsilon)^k \epsilon^{T-k} \leq \exp(-\frac{T}{2}(1-2\epsilon)^2)$$

  随着弱学习器数目T的增加成指数下降

- **然而独立不可能做到（因为是同一问题训练出的学习器）**
  - 通过依次调整样本分布，串行生成一系列弱分类器→AdaBoost
  - 通过采样的方法生成T个数据集，并行化生成弱分类器→Bagging

(引入随机属性选择→ 随机森林)

集成

结合策略

- **平均**
  - AdaBoost: 根据错误率对每个弱分类器赋权
  - 贝叶斯平均(BMA)：根据后验概率对不同模型赋权
- **相对多数投票**
  - Bagging与随机森林
- **自适应**
  - 在线学习(Online Learning)中的Hedge algorithm:每一轮随机选择一个专家(弱学习器)的意见，选择的概率取决于专家们前面的表现

经典集成学习算法 Bagging

- 并行
- Bagging (bootstrap aggregating)
  - 基本思路：利用不同的样本集合训练单个模型
  - 假设原始数据集为D（n个样本），步骤如下：

    for $i = 1$ to T

    从D中独立随机抽取m个数据（$m < n$），构成数据集S_i

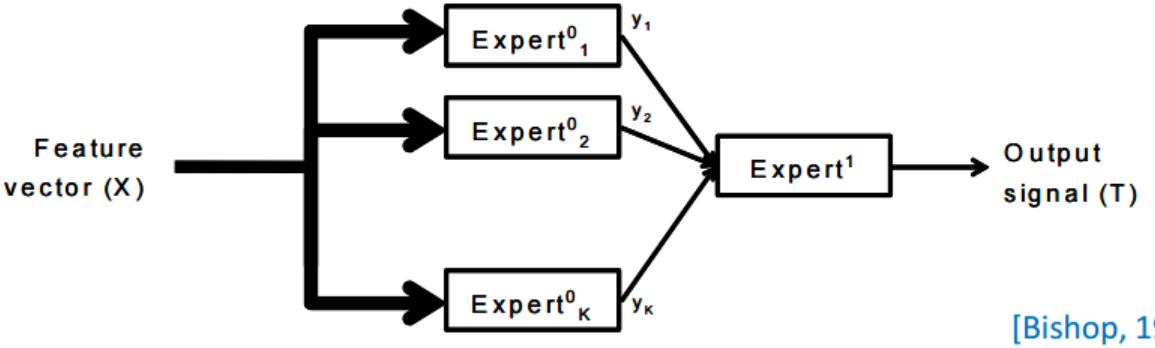
    用S_i训练得到一个模型

    end

    所有模型的输出结果投票决定最终的模型输出

经典集成学习算法 Stacking

- 层级
- Stacking (stacked generalization)
- 基本思路：两层的集成学习
  - 第一层：在原始数据集上训练多个不同的模型
  - 第二层：第一层模型的输出作为第二层模型的输入



[Bishop, 1995]

Anderson Acceleration (AA)

- Solve fixed point problems
  - $u = G(u)$
- Faster than Picard iteration
  - $u_{k+1} = G(u_k)$
- Motivation (Anderson 1965) in electronic structure computations

Basic Algorithms

anderson (u_0, G, m)

$u_1 = G(u_0);\ F_0 = G(u_0);$

for $k = 1, 2, 3....$

$m_k = \min(m, k)$

$F_k = G(u_k) - u_k$

Minimmize $\left\| \sum_{j=0}^{m_k} \alpha_j^k F_{k-m_k+j} \right\|$ subject to

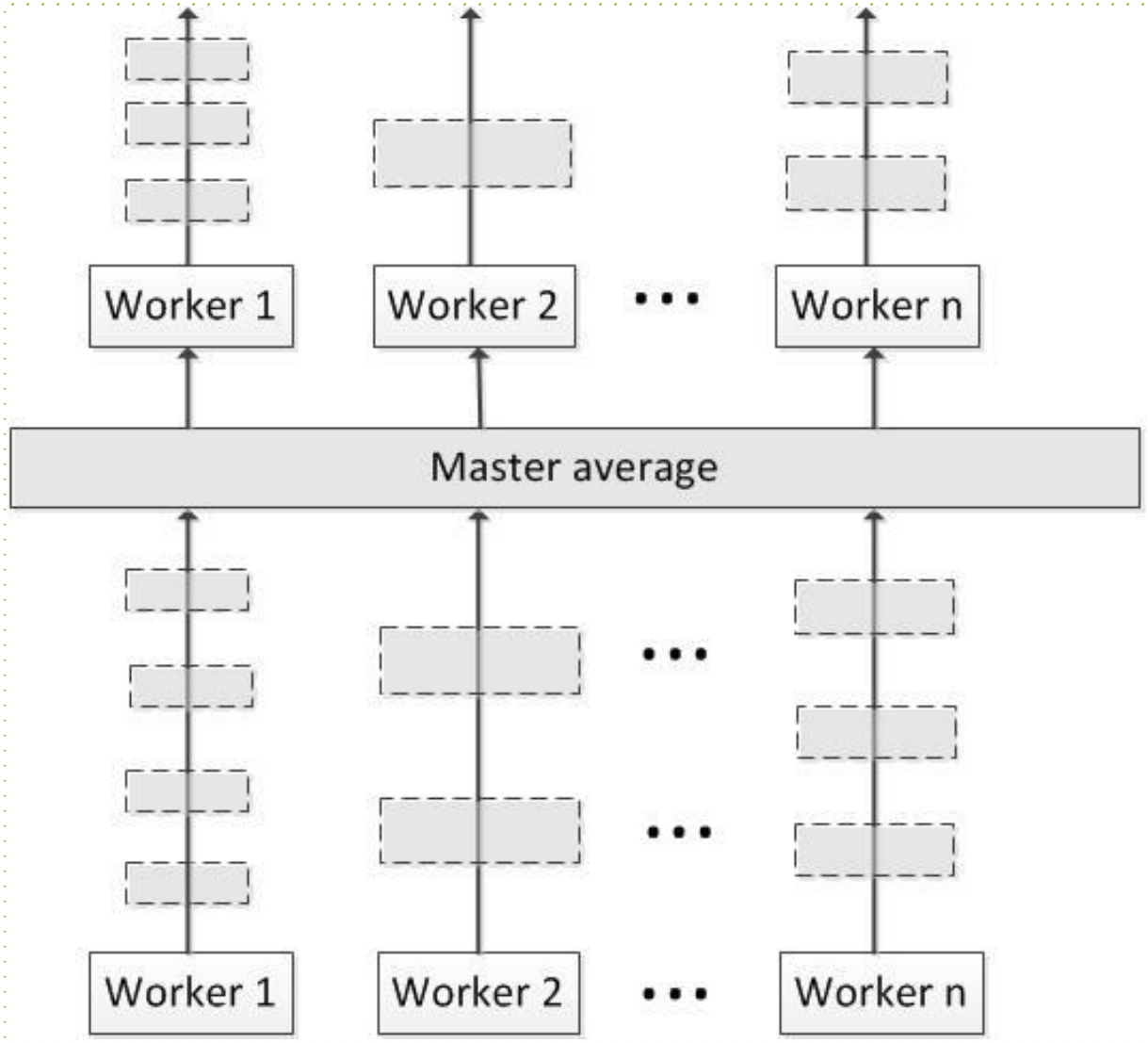
$\sum_{j=0}^{m_k} \alpha_j^k = 1$

$u_{k+1} = (1-\beta_k) \sum_{j=0}^{m_k} \alpha_j^k u_{k-m_k+j} + \beta_k \sum_{j=0}^{m_k} \alpha_j^k G(u_{k-m_k+j})$

end for

Why use AA?

- Convergence q-linearly for linear case and local r-linearly for local Lipschitz continuously differentiable G
- Does not require derivatives
- Implement efficiently by QR factorization
- Somewhat better conditioning


Worker 1
Worker 2
Worker n
Master average
Worker 1
Worker 2
Worker n

致谢

谢谢！