机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

信息与通信工程学院 网络搜索教研中心

北京邮电大学

本学期的授课主题

- **机器学习与数据科学**

- 机器学习
  - 围绕有监督学习问题展开

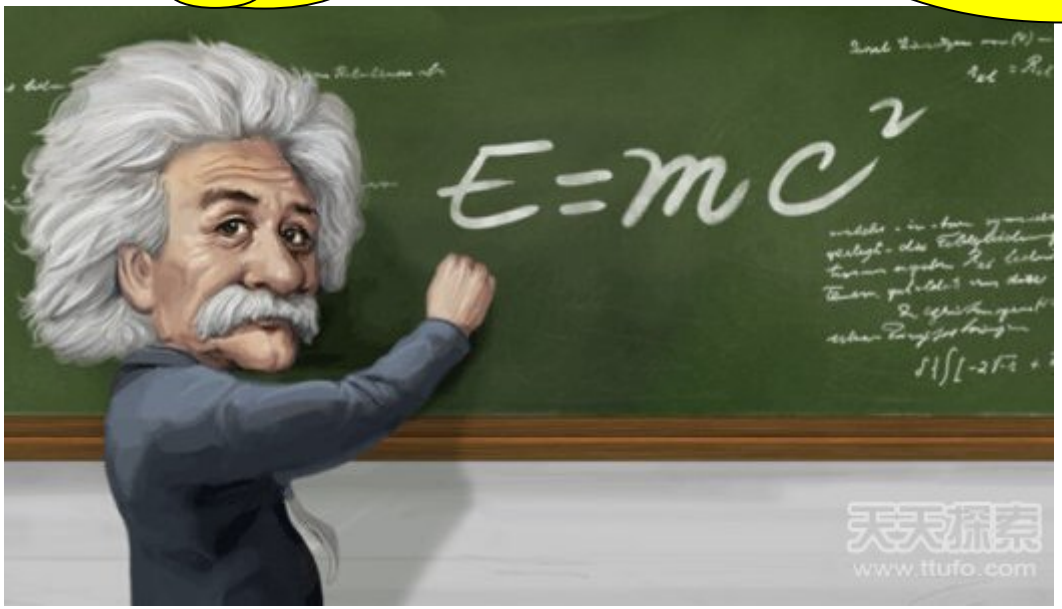
- 数据科学
  - 包括数据的感知与获取、（探索性）分析与处理、计算与存储等，主要对应于机器学习的无监督学习部分（聚类/降维/异常检测等）

内容提要

- 课程基调
- 课程定位与目标
- 学习建议
- 授课内容与安排
- 考核方式
- 参考书目

知识 vs. 想象力

- **Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand.**



爱因斯坦(Einstein)说：“想象力比知识更重要。因为知识是有限的，而想象力概括着世界的一切，推动着进步”。

知识 vs. 想象力

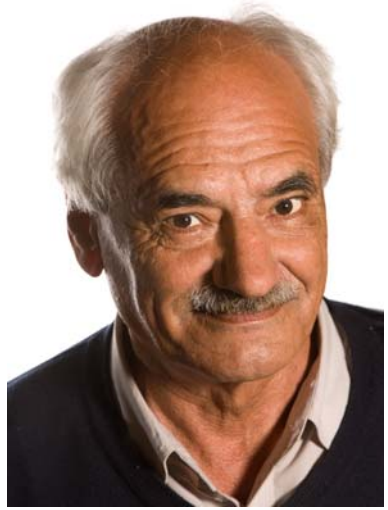
- 爱因斯坦的论断有其特定适用范围，并非普遍成立
  - 爱因斯坦所说的是，当你行走在现有理论的边界，只能借助想象力去开拓方向。
  - 而对于还在学习和训练中大部分人来说，坚实的知识和理论储备更重要，因为真正富有远见、深刻的洞察力和颠覆性的想象往往需要建立的坚实的理论基础之上。
  - 在大多数时候，理论就好比登山杖，是推进研究工作的重要工具。
    - “没有金刚钻、别揽瓷器活儿”
    - “工欲善其事、必先利其器”
    - “把水加热到沸点”

理论 vs. 实用

- 基于VC维的统计学习理论之父

Nothing is more practical than a good theory.

“没有什么比一个好的理论更实用。” --- V. Vapnik

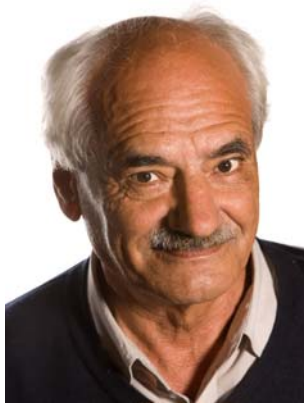


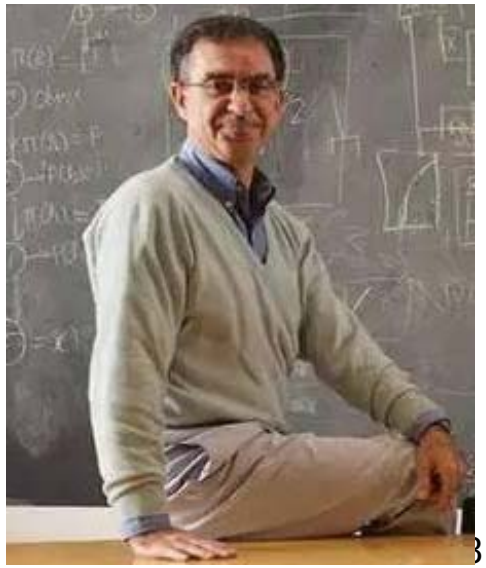
打好理论基础的理由

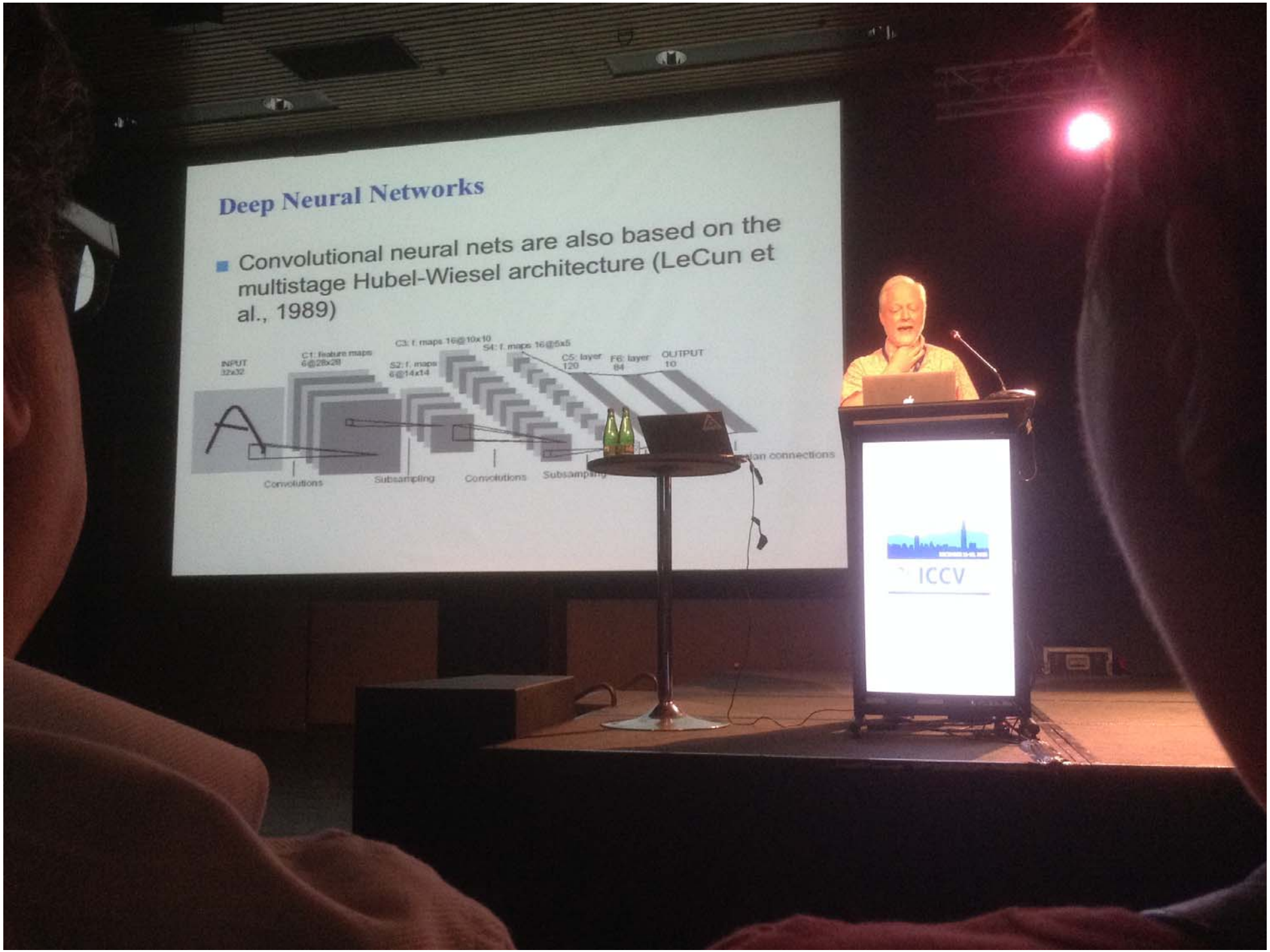
- 站在前人的肩膀上开展研究工作
  - 为了能够读懂研究领域内的精要文献，做出更好的研究
- 理论可以指导实践
  - 为了更好地指导实践
    - “没有理论的实践是盲目的实践”
  - 为了更有效率地找到创新性点子或创新方向

"What I am most fond of are beautiful and simple theoretical ideas that can be translated into something that works".

- "我最喜欢的，是那些能够被变成特别好用的东西的、漂亮而又简单的理论点子。"
  - 这是谁说的？六选一


Deep Neural Networks
Convolutional neural nets are also based on the multistage Hubel-Wiesel architecture (LeCun et al., 1989)
INPUT 32x32
C1: feature maps 6@28x28
S2: f. maps 6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
C5: layer 120
F6: layer 84
OUTPUT 10
Convolutions
Subsampling
Convolutions
Subsampling
ICCV


PAMI Distinguished Researcher Award
Yann LeCun
Sponsored by
intel
ICCV

课程定位与目标

- 课程定位
  - 专业基础课
    - **面向对机器学习与数据科学方向感兴趣的博士研究生、硕士研究生、以及理论基础较好的高年级本科生**

- 课程目标
  - **为在下列方向开展深入研究工作打下必要基础**
    - 机器学习 / Machine Learning
    - 数据科学 / Data Science
    - 模式识别 / Pattern Recognition
    - 机器视觉 / Computer Vision
    - 数据挖掘 / Data Mining
    - …

学习建议

- 1. 记笔记、课后要尽可能搞懂每个细节
  - 记一些笔记(要点/关键**idea**)
    - 两耳听课终觉浅，绝知课后要推导
- 2. 阅读学习
  - 阅读论文和专著，不能停留于“速食”产品
    - 建立自己的专业知识树
- 3. 做实验
  - 典型算法要在相应**Benchmark**数据上做实验
    - 在具体的实验中，通过观察和思考，发现问题

授课内容的几个线索

- 1. 有监督学习 → 无监督学习
  - 回归和分类 ...
  - 降维和聚类(以及密度估计、异常值检测) ...
- 2. 算法由简单到复杂
  - 基于实例的学习方法、线性模型、非线性模型 ...
- 3. 对学习过程的认识和理解的由浅入深：
  - 从基于记忆的学习开始、模型中参数的优化、学习过程的统计性质、基于VC维的统计学习理论、正则化理论

YES !

这些理论，可以拿来解释生活、指导人生么?

人生轨迹的最优解

- 正则化理论
  - 在学习专业知识的同时，要学习如何做人、做事、做学问的道理，要结识良师益友
    - 经验误差项：
      - 专业知识
    - 通用的正则化项：
      - 做人做事做学问的道理
    - 专用的正则化项
      - **良师益友**

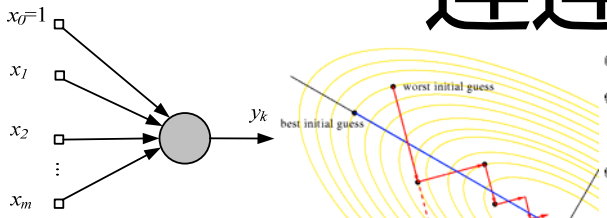
$$f^* = \arg\min_{f \in H} \varepsilon(f) + \lambda \cdot \Omega(f) + \gamma \cdot \omega(f)$$

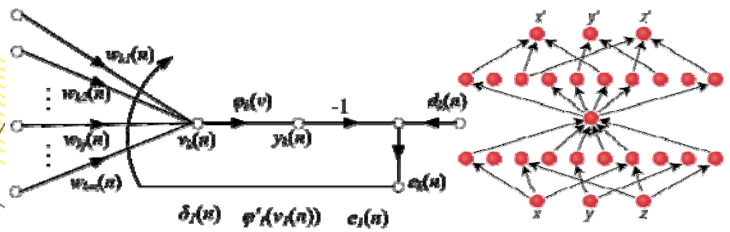
《机器学习&数据科学》2018年春季 授课安排

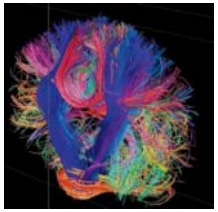
- **专题0 引论 3 (1+2) 课时**
  - 学习问题发展简史 / 必要数学基础
- **专题 1 基于实例的学习 3 课时**
  - NN \ k-NN \ Pazen窗 \ 理论渊源
- **专题 2 线性模型 3 课时**
  - 线性回归 \ 线性分类 \ 感知器 \ 从统计视角看最小二乘 → 稀疏编码?
- **专题 3 线性模型的扩展 3 + 3 课时**
  - 核方法 \ 多层感知器& BP 算法 \ 最优化理论 → 深度网络?
- **专题 4 深度学习 3 课时**

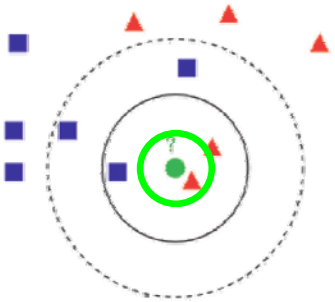
有监督学习

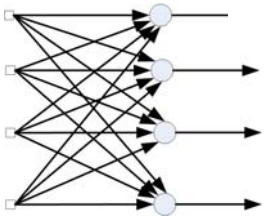
无监督学习

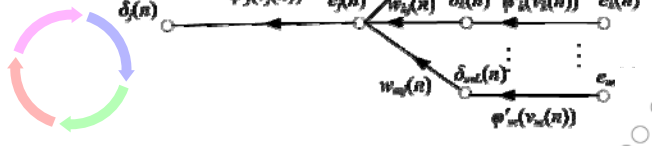


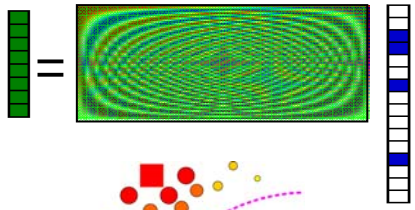


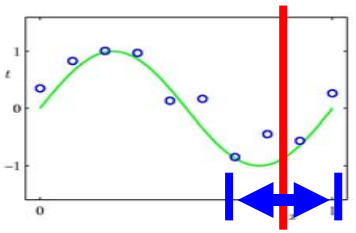


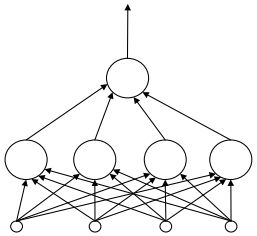


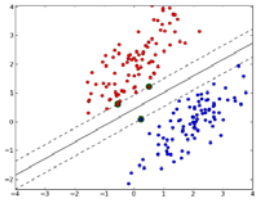


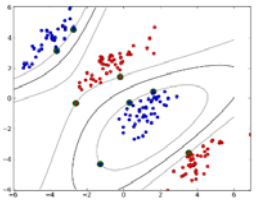


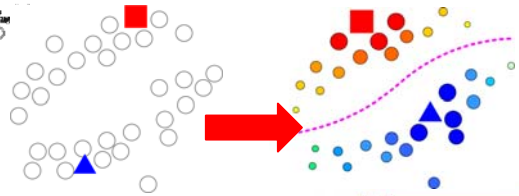


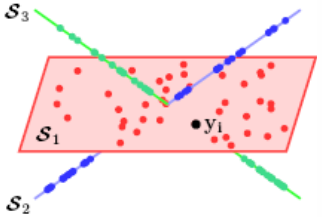


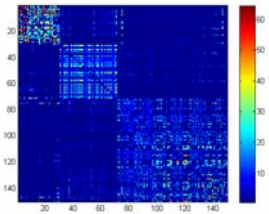


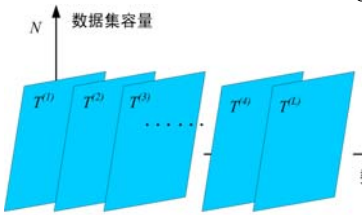


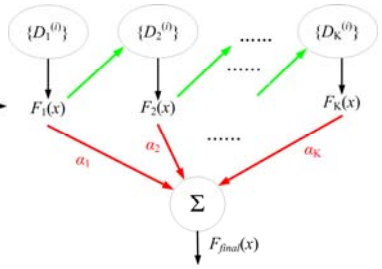


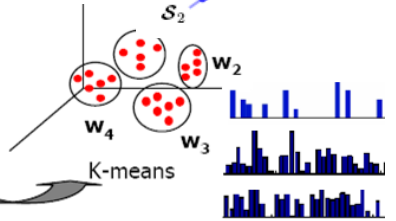


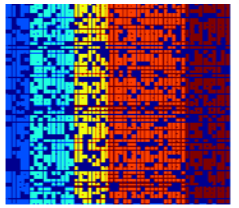


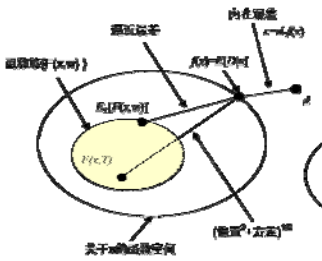


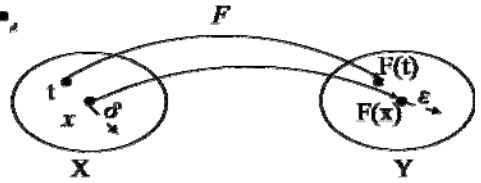


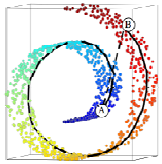


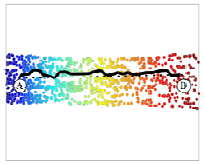


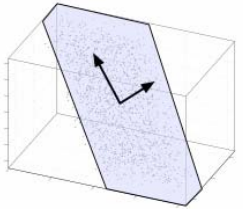






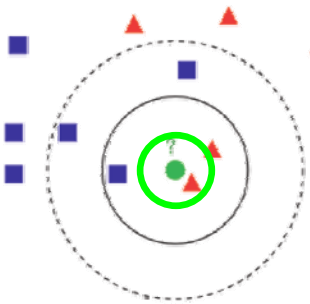


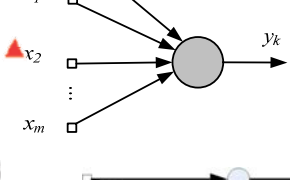


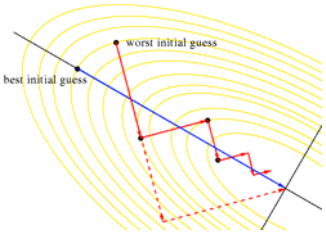


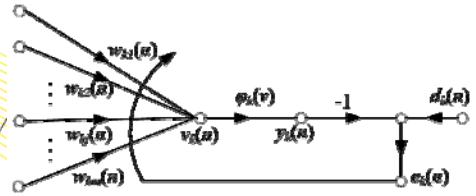
连连看?

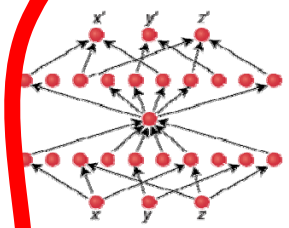
Since 1699, when French explorers landed at the great Bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-

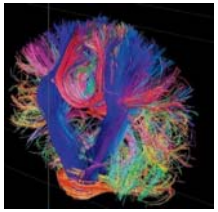


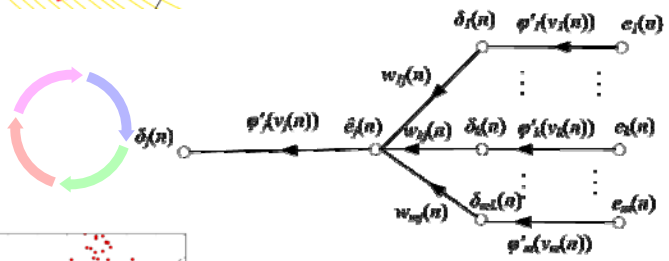


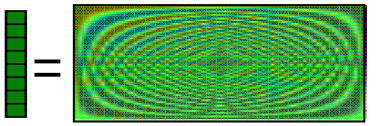


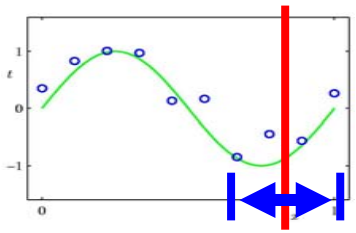


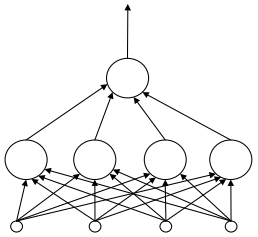


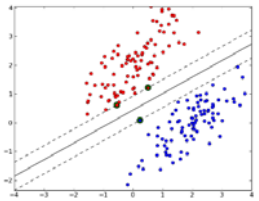


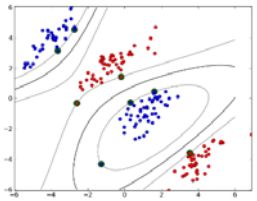


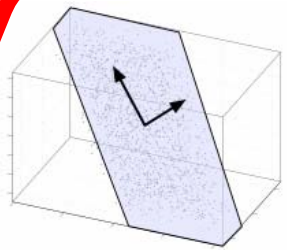


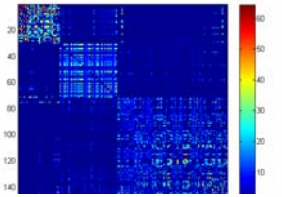


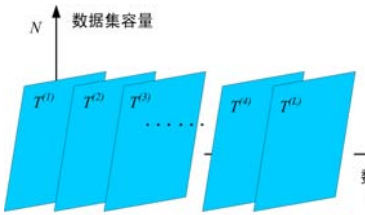


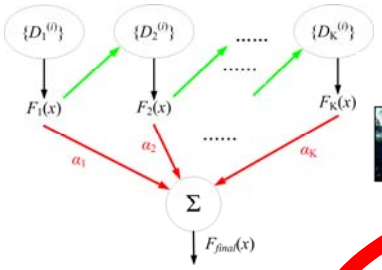


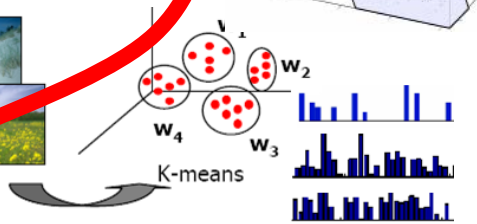


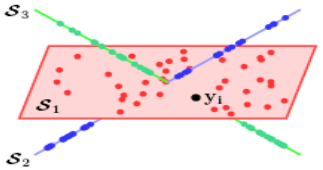


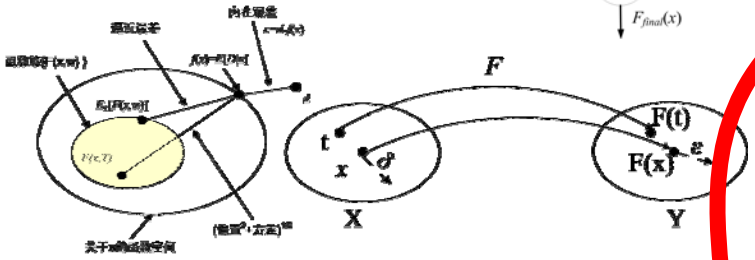


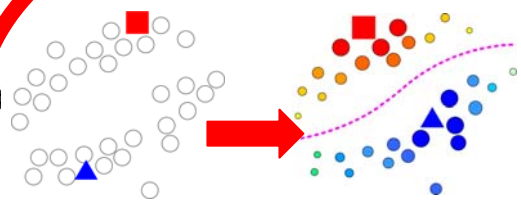


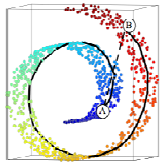


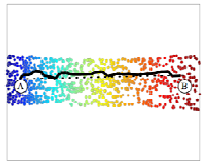


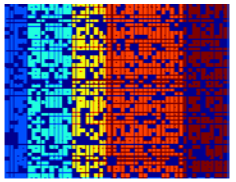












图说ML&DS2018

x_0
x_1
x_2
x_m
y_k
worst initial guess
best initial guess
$w_{kj}(n)$
$\varphi_k(v)$
$v_k(n)$
$y_k(n)$
-1
$d_k(n)$
$e_k(n)$
$\delta_j(n)$
$\varphi'_j(v_j(n))$
$w_{kj}(n)$
$\delta_k(n)$
$e_k(n)$
x'
y'
z'
x
y
z
=
1
0
−1
t
0
数据集容量
N
$T^{(1)}$
$T^{(2)}$
$T^{(3)}$
$T^{(4)}$
$T^{(L)}$
数据集标号
$\{D_1^{(i)}\}$
$\{D_2^{(i)}\}$
$\{D_K^{(i)}\}$
$F_1(x)$
$F_2(x)$
$F_K(x)$
α_1
α_2
α_K
Σ
$F_{final}(x)$
w_1
w_2
w_3
w_4
K-means
s_1
s_2
s_3
y_1
F
t
x
F(t)
F(x)
X
Y

考核方式

- **平时成绩 60%**
  - 平时作业
    - 每个专题布置1-3道思考题，可选做其中部分题目
    - 题目涉及分析、计算与推导 或 编程实验

- **期末成绩 40%**
  - 期末作业或考试 (分析与推导类):
    - 给定一组题目，要求完成其中一部分

课程参考书目

- 《机器学习与数据科学》讲义
  - **2008 - (2018?) 准备中…**

- 中文参考书:
  - 周志华，机器学习，清华大学出版社，**2016.**
  - 于剑，机器学习——从公理到算法， **2017.**

- 影印版机器学习参考书
  - **Kevin Patrick Murphy: "Machine Learning: a Probabilistic Perspective", MIT Press, 2012.**
  - **Simon Haykin, Neural Networks and Learning Machines (Third Edition), Pearson Education, 2009.03. [神经网络与学习机器]**
  - **Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.**
  - **Simon Haykin, Neural Networks: A Comprehensive Foundation (2ed edition), Pearson Education.**
    - Simon Haykin著，叶世伟，史忠植译, 神经网络原理，机械工业出版社，2006年
  - **Rachel Schutt, Cathy O'Neil，数据科学(Doing Data Science)[影印版], 东南大学出版社, 2014年9月.**

扩展参考书目

- 偏向统计和统计学习理论
  - **Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar: Foundations of Machine Learning, MIT Press, 2012.**
  - **Trevor Hastie et al., The Elements of Statistical Learning(影印版)，2009.01**
    - Trevor Hastie et al.著，范明等译，统计学习基础：数据挖掘、推理与预测，北京：电子工业出版社，2004年1月.
  - **L. Devroye, L. Gyorfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Springer, 1997.**
  - **V. N. Vapnik. Statistical Learning Theory. Wiley, 1998.**
  - **V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.**

Q / A

- Any Questions...

研究生培养模式

- 前2-3年 课业任务异常繁重
  - 打下坚实宽广的理论基础
    - → 有了金刚钻儿, 才能揽瓷器活儿
  - 协助导师从事部分研究工作
    - **RA：Research Assistant**
- 后2-3年 论文工作相对轻松
  - 重视开题环节
- 一般要5-6年左右毕业

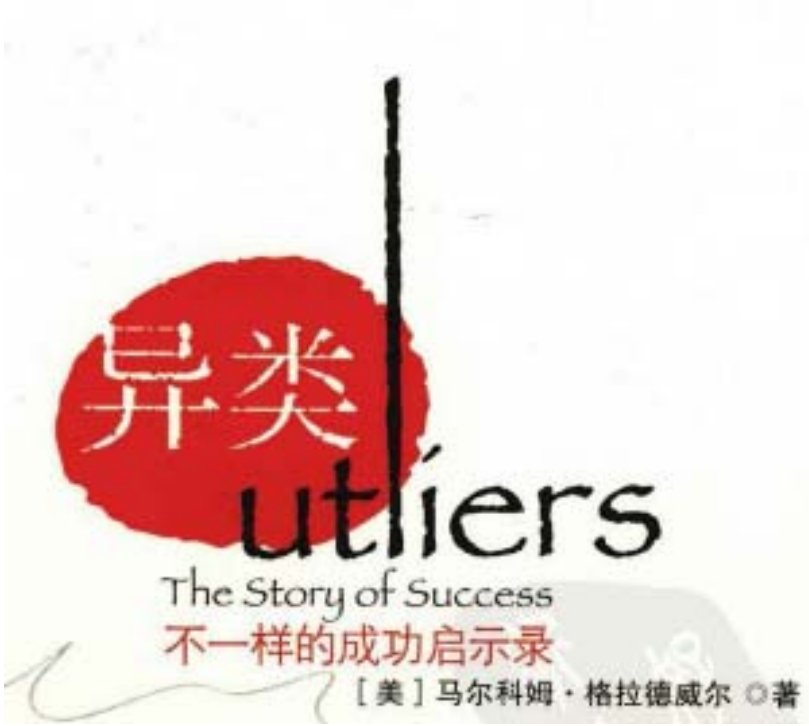
“基础，基础，还是基础!”

- 在研究生培养中，要重视理论基础
  - 扎实的理论基础是科研创新活动中永不枯竭之原动力，是酝酿突破性灵感的源泉
  - 施一公教授强调：“要想在科学研究上取得突破，批判性分析(critical analysis)是一种必需具备的素质”；而“严密的逻辑是批判性分析的根本”。
    - 这里所说的"严密的逻辑"不是别的，指的是理论基础；理论基础是当你行走在泥泞的研究之路上帮助你前行的拐杖。

天才的密码: 10000 hours

- 在任何专业领域，特别是科学和艺术领域，要达到世界一流水平，需要10000 小时的有效训练

  – **Outliers? (《异类》)**



态度决定高度

- No matter what you do, your attitude determines your altitude.
  - 无论做什么事情，你的态度决定你的高度。
    - By Margaret Thatcher (玛格丽特·撒切尔)
  - **The secrete ingredient of my secrete ingredient soup is nothing**
    - To make something special, you just believe in it's special.
      - From《功夫熊猫 I》


KUNG FU
PANDA
2
IN DIGITAL 3D AND 2D IN SELECTED THEATRES