2020.2.18 - 2020.6.2 9:50-12:15@6C102雨课堂+腾讯会议
课程号30250293

模式识别与机器学习
Pattern Recognition & Machine Learning

汪小我、张学工
xwwang@tsinghua.edu.cn
zhangxg@tsinghua.edu.cn
清华大学自动化系

《模式识别与机器学习》听课方式要求

- 用腾讯会议系统在线听课，建议用电脑听课
  - 会议系统备用顺序：腾讯会议、会畅、Zoom、B站直播
- 最好用耳麦，不发言时麦克风静音，需发言时临时打开
- 保持荷塘雨课堂连接，但雨课堂静音
- 要求用雨课堂互动时使用雨课堂互动
- 如出现连接故障及时用雨课堂弹幕和微信告知

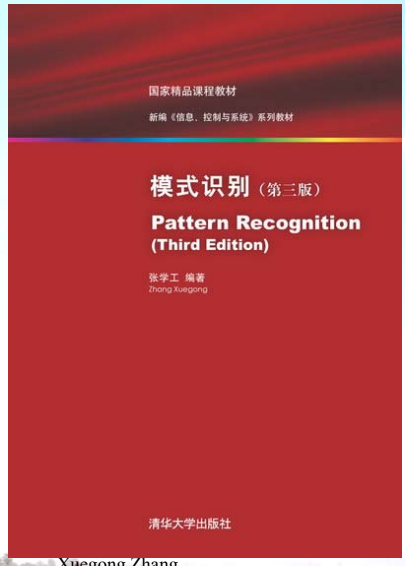
疫情期间课程组织方式

- 按课表时间上课，课间休息教师根据内容掌握
  - 2月18日起，每周二9:50-12:15，请提前15分钟连线
- 课堂纪律：
  - 按时上下课，上课期间不做与上课无关的事
  - 认真听讲，积极互动
  - 出现技术故障立即反馈
- 上课技术方案
  - 主方案：腾讯会议 + 雨课堂 + 微信
  - 备用方案A：其他远程会议 +雨课堂 + 微信
  - 备用方案B：B站直播 + 微信

本课程主要内容

- 模式识别与机器学习的概念
- 模式识别与机器学习问题的数学表达和系统基本构成
- 模式识别与机器学习主要流派的基本思想、基本理论与代表性方法
- 前沿讨论

教材和参考书



课程安排

- 要求：
  - 听课、互动
  - 课外阅读教材
  - 作业（小实验）
- 成绩构成：
  - 课堂成绩
  - 平时作业
  - 期末考试

板块	内容
1、基础与经典确定性方法	模式识别与机器学习概念，线性方法，分类器评价方法，人工神经网络，支持向量机，近邻法，决策树，集成学习
2、概率学习	贝叶斯决策理论，概率密度估计，贝叶斯网络与隐马尔可夫模型
3、特征工程	模型选择，特征选择，特征提取，降维与流形学习
4、非监督学习	聚类分析，混合模型估计，非监督神经网络
5、深度学习	卷积神经网络，循环神经网络，深度神经网络，表示学习
6、前沿探索	迁移学习，半监督学习，生成模型，机器学习历史回顾，机器学习的伦理问题

与另一门课的关系

《模式识别与机器学习》

- 中文授课
- 汪小我、张学工
- 春季学期
- 本科生课
- 每周3学时
- 有指定教材
- 自动化系本科生必修课

Machine Learning

- English instructions
- Xuegong Zhang
- Fall semester
- Graduate/undergraduate
- Three hours per week
- No textbook yet
- Can be used to replace the required course for DAU undergraduate students

有没有问题？

弹幕时间

弹幕时间

弹幕时间

弹幕时间

2020.2.18 - 2020.6.2 9:50-12:15@6C102雨课堂+腾讯会议
《模式识别与机器学习》

第一章 绪论

2020.2.18

Xiegong Zhang

讨论

- 什么是模式识别（Pattern Recognition）？
- 什么是机器学习（Machine Learning）？

看图说话

看图说话

看图说话











看图说话



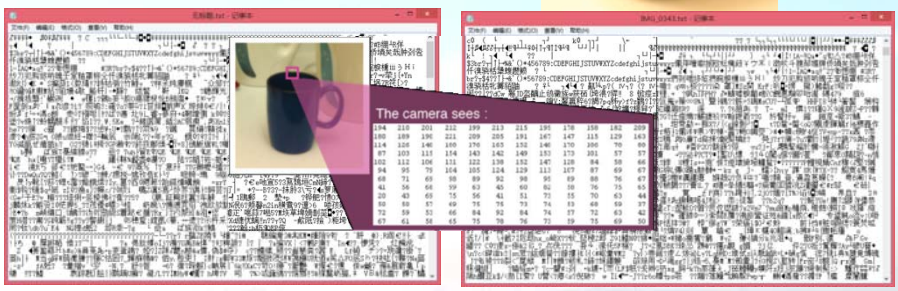








思考题：
- 你怎么认出来它是杯子？
- 能不能让机器也认出来？


The camera sees :

什么是模式识别？

“看”到东西 ➔ 认出东西、产生想法

观察 ➔ 判断

观测 ➔ 分类决策

量化观测 ➔ 类别标签

$x \in R^d$ ➔ $y \in \{\omega_1, \omega_2, \cdots\}$

模式识别

什么是模式识别？

人

机器

“看”到东西 ➔ 认出东西、产生想法

观察 ➔ 判断

观测 ➔ 分类决策

量化观测 ➔ 类别标签

$x \in R^d$ ➔ $y \in \{\omega_1, \omega_2, \cdots\}$

模式识别

如果机器通过实例学会识别，而不是在程序里写好怎样识别，那就是**机器学习**。

什么是模式识别？

人

机器

“看”到东西 ➔ 认出东西、产生想法

观察 ➔ 判断

观测 ➔ 分类决策

量化观测 ➔ 类别标签

$x \in R^d$ ➔ $y \in \{\omega_1, \omega_2, \cdots\}$

模式识别

如果机器通过实例学会识别，而不是在程序里写好怎样识别，那就是**机器学习**。

基于数据的机器学习

什么是模式识别？

人 → 机器

“看”到东西 ➔ 认出东西、产生想法

观察 ➔ 判断

观测 ➔ 分类决策

量化观测 ➔ 类别标签

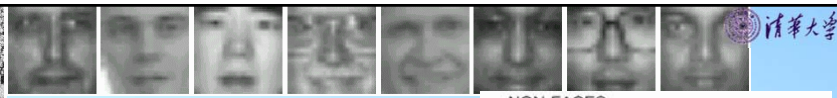
$x \in R^d$ ➔ $y \in \{\omega_1, \omega_2, \cdots\}$

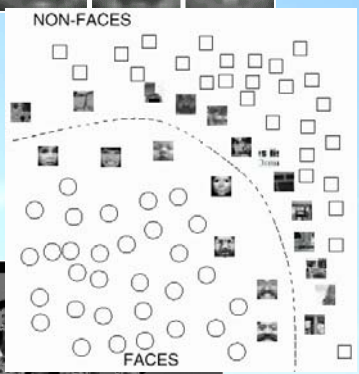
模式识别

如果机器通过实例学会识别，而不是在程序里写好怎样识别，那就是**机器学习**。

基于数据的机器学习

Xuegong Zhang





Souroce:
MIT AI Lab

Xuegong Zhang

22


Initial Training Set
Classifier
False Positives
Pedestrian Detection System
Souroce:
MIT AI Lab
Xuegon
23

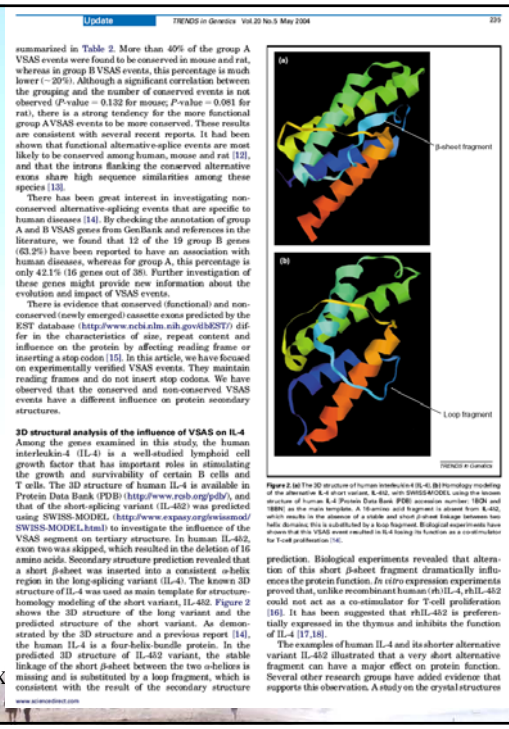

之前的选做大作业：用计算机检测照片中的人脸并识别其性别
Xuegong Zhang
24


Xuegong Zhang
25


Xuegong Zhang
26


清華大學
Xuegong Zhang
27


清華大學
X



Update TRENDS in Genetics Vol.20 No.5 May 2004 229

summarized in Table 2. More than 40% of the group A VSAS events were found to be conserved in mouse and rat, whereas in group B VSAS events, this percentage is much lower (~20%). Although a significant correlation between the grouping and the number of conserved events is not observed (*P*-value = 0.132 for mouse; *P*-value = 0.081 for rat), there is a strong tendency for the more functional group A VSAS events to be more conserved. These results are consistent with several recent reports. It had been shown that functional alternative-splice events are most likely to be conserved among human, mouse and rat [12], and that the introns flanking the conserved alternative exons share high sequence similarities among these species [13].

There has been great interest in investigating non-conserved alternative-splicing events that are specific to human diseases [14]. By checking the annotation of group A and B VSAS genes from GenBank and references in the literature, we found that 12 of the 19 group B genes (63.2%) have been reported to have an association with human diseases, whereas for group A, this percentage is only 42.1% (16 genes out of 38). Further investigation of these genes might provide new information about the evolution and impact of VSAS events.

There is evidence that conserved (functional) and non-conserved (newly emerged) cassette exons predicted by the EST database (http://www.ncbi.nlm.nih.gov/dbEST/) differ in the characteristics of size, repeat content and influence on the protein by affecting reading frame or inserting a stop codon [15]. In this article, we have focused on experimentally verified VSAS events. They maintain reading frames and do not insert stop codons. We have observed that the conserved and non-conserved VSAS events have a different influence on protein secondary structures.

3D structural analysis of the influence of VSAS on IL-4

Among the genes examined in this study, the human interleukin-4 (IL-4) is a well-studied lymphoid cell growth factor that has important roles in stimulating the growth and survivability of certain B cells and T cells. The 3D structure of human IL-4 is available in Protein Data Bank (PDB) (http://www.rcsb.org/pdb/), and that of the short-splicing variant (IL-4δ2) was predicted using SWISS-MODEL (http://www.expasy.org/swissmod/SWISS-MODEL.html) to investigate the influence of the VSAS segment on tertiary structure. In human IL-4δ2, exon two was skipped, which resulted in the deletion of 16 amino acids. Secondary structure prediction revealed that a short β-sheet was inserted into a consistent α-helix region in the long-splicing variant (IL-4). The known 3D structure of IL-4 was used as main template for structure-homology modeling of the short variant, IL-4δ2. Figure 2 shows the 3D structure of the long variant and the predicted structure of the short variant. As demonstrated by the 3D structure and a previous report [14], the human IL-4 is a four-helix-bundle protein. In the predicted 3D structure of IL-4δ2 variant, the stable linkage of the short β-sheet between the two α-helices is missing and is substituted by a loop fragment, which is consistent with the result of the secondary structure prediction. Biological experiments revealed that alteration of this short β-sheet fragment dramatically influences the protein function. *In vitro* expression experiments proved that, unlike recombinant human (rh)IL-4, rhIL-4δ2 could not act as a co-stimulator for T-cell proliferation [16]. It has been suggested that rhIL-4δ2 is preferentially expressed in the thymus and inhibits the function of IL-4 [17,18].

The examples of human IL-4 and its shorter alternative variant IL-4δ2 illustrated that a very short alternative fragment can have a major effect on protein function. Several other research groups have added evidence that supports this observation. A study on the crystal structures

www.sciencedirect.com

Figure 2. (a) The 3D structure of human interleukin-4 (IL-4). (b) Homology modeling of the alternative IL-4 short variant, IL-4δ2, with SWISS-MODEL using the known structure of human IL-4 (Protein Data Bank (PDB) accession number: 1BON and 1BBN) as the main template. A 16-amino acid fragment is absent from IL-4δ2, which results in the absence of a stable and short β-sheet linkage between two helix domains; this is substituted by a loop fragment. Biological experiments have shown that this VSAS event resulted in IL-4 losing its function as a co-stimulator for T-cell proliferation [16].

β-sheet fragment

Loop fragment

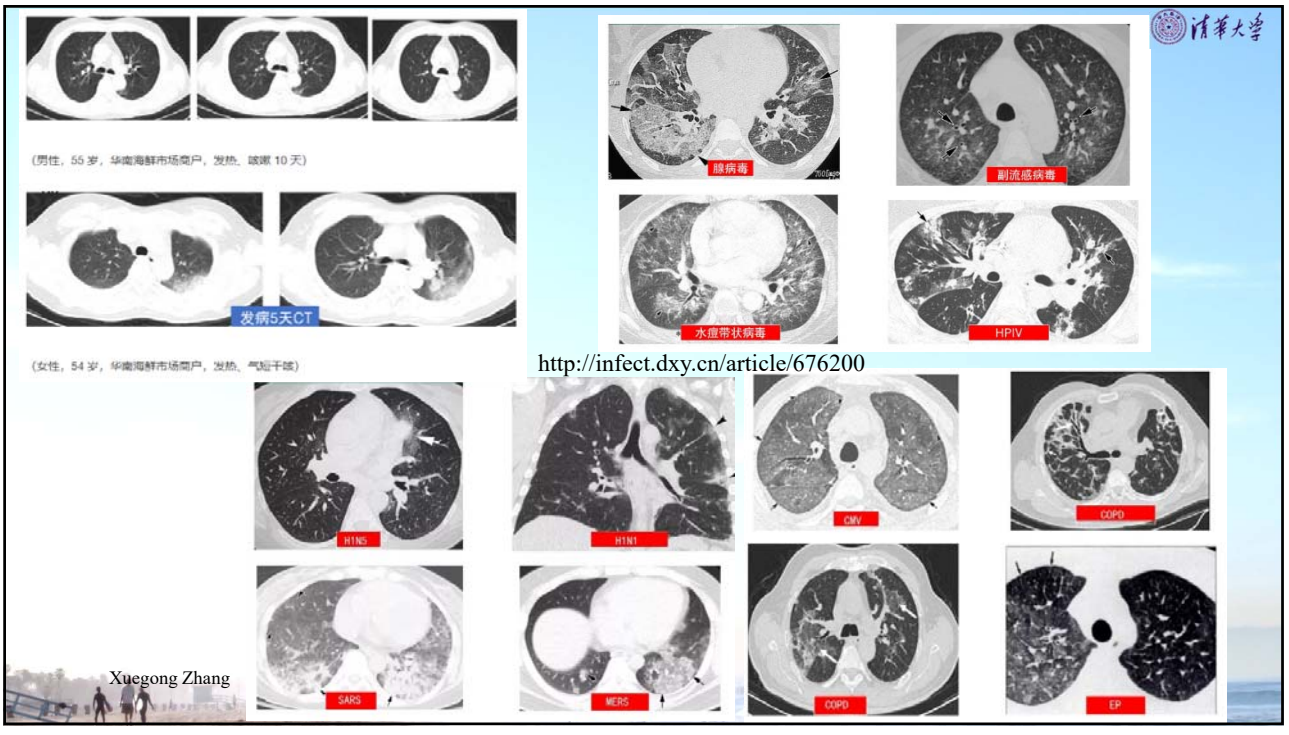
TRENDS in Genetics

OCR

summarized in Table 2. More than 40% of the group A VSAS events were found to be conserved in mouse and rat, whereas in group B VSAS events, this percentage is much lower (~20%). Although a significant correlation between the grouping and the number of conserved events is not observed …

X

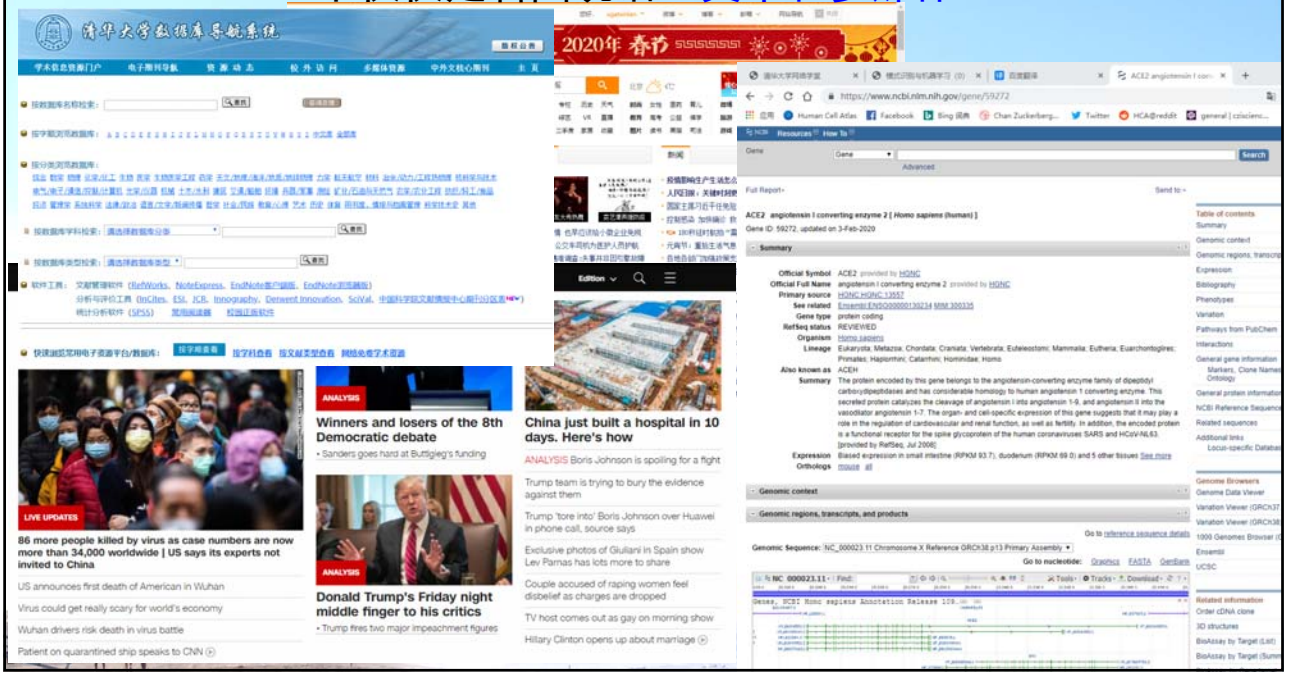
29



http://infect.dxy.cn/article/676200

Xuegong Zhang

不仅仅是看图说话：文本和多媒体



不仅仅是看图说话：电子病历

姓名：XXX　科室：肿瘤内科　病区：七病区　病案号：00067947

姓名：XXX　**性别**：女

年龄：43 岁　民族：汉族

婚姻状况：已婚　出生地：河北省衡水市

职业：农民　病史陈述者：患者本人

入院时间：2013-05-27 16:21　记录时间：2013-05-27 16:21

主　诉：间断咳嗽、咳痰 3 年，加重伴气短 1 月余

现病史：患者近 3 年来出现咳嗽，以干咳为主，咳痰费力，无明显气短，未予诊治。近 1 月多来感上述症状加重，伴气短，活动后明显，偶有喘鸣，行胸部 CT 示气管隆突上方占位，左锁骨上区淋巴结转移，行气管镜示气管下段肿物，活检病理为类癌。为进一步诊治收入院。

既往史：体健。否认明确的高血压、冠心病、糖尿病等慢性病史，否认肺结核、肝炎等传染病史，否认外伤史，有输血史，无明确食物、药物过敏史，预防接种史不详。

个人史：生于原籍，久居当地，近期无牧区及疫区接触史，无烟酒嗜好。

月经婚育史 21 岁结婚，有有 1 子 1 女，爱人与儿女均体健。

家族史：父母亲体健，有 2 弟均体健。否认家族遗传病史及恶性肿瘤家族史。

体 格 检 查

体温 36.6℃　脉搏 80 次/分　呼吸 22 次/分　血压 120/70mmHg

一般情况：　发育正常，营养中等，自主体位，神志清楚，对答切题，检查合作。

主诉：全身皮肤散在瘀点2天。

现病史：患者2天前无明显诱因出现全身皮肤散在瘀点、瘀斑，以四肢为甚，伴有牙龈出血，无血便、血尿及呕血，无畏寒、发热，无咳嗽。未就诊及治疗。今天仍有出现新鲜瘀点，为进一步检查、治疗拟收住院。四天前有腹泻。自发病以来，无畏寒、发热，无头痛、头晕，无咳嗽、咳痰，无胸闷、胸痛及气促。无腹胀、腹痛。大、小便正常，无血尿、血便。食欲、睡眠尚正常。双下肢无浮肿。

既往史：10年前有类似病史，经治疗后好转，具体诊断及治疗不详。否认有高血压病、糖尿病、肝炎、肺结核、血友病等病史。否认有外伤史、手术史及输血史。否认有药物及食物过敏史。预防接种史不详。

个人史：原籍出生、长大，近五年在ZZ工作，从事化妆品制造工作多年。未涉及疫水及传染病区。无嗜酒史。吸烟史6年，10支/天。

婚姻生育史:未婚、未育。

家族史：否认家族中有类似疾病患者，否认家族中有肝炎、肺结核、高血压病、糖尿病、血友病及肿瘤等疾病。

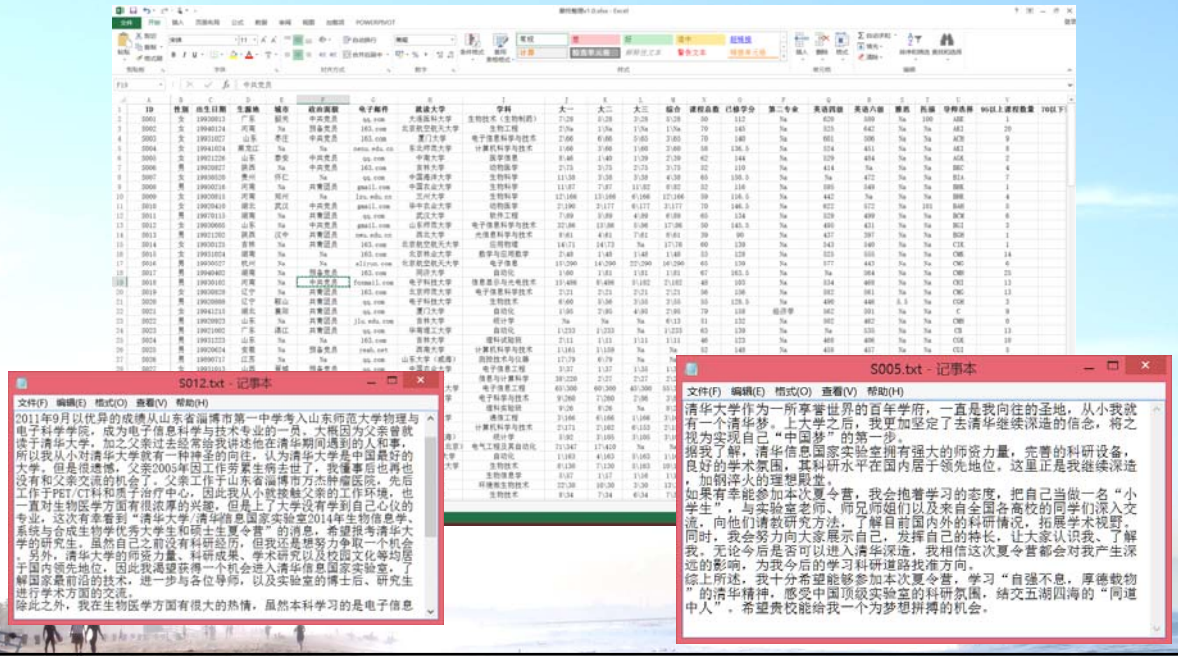
体格检查：T 36.5℃　P 80次/分　R 20次/分　BP 110/70mmHg

发育正常，营养不良，……

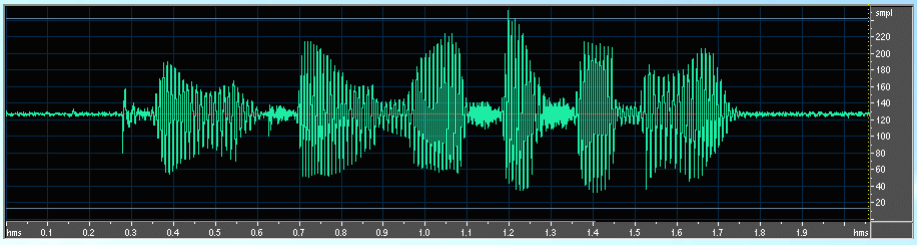
辅助检查

BLOOD RT：WBC：12.1×10*9/L, HGB:118g/L ,RBC:5.25×10*12/L ,PLT:3×10*9/L

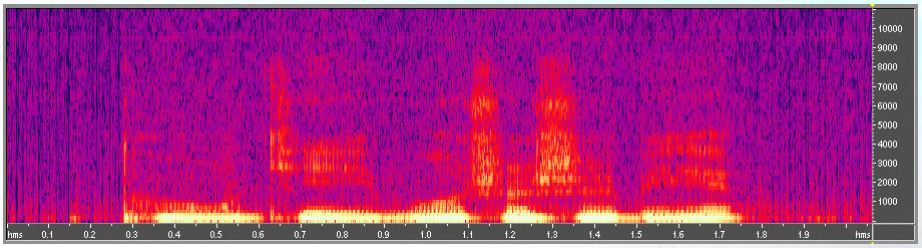
不仅仅是看图说话：研究生申请材料



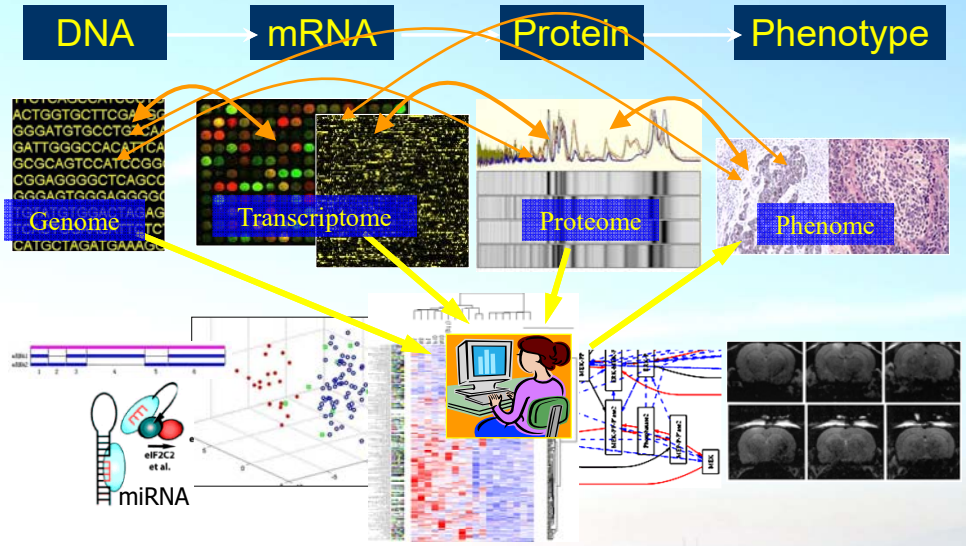
不仅仅是看图说话：语音



“统计模式识别”



不仅仅是看图说话： 基因组学数据



• • • • • •

再看什么叫模式识别

- People *recognize* things, from observations.

  ---- 识别
- People recognize things by *recognizing patterns*, rather than individual observations.

  ---- 模式识别

再看什么叫模式识别

- People *recognize* things, from observations.

  ---- 识别
- People recognize things by *recognizing patterns*, rather than individual observations.

  ---- 模式识别
- *Pattern Recognition with Machines*
  - People like to make machines that can do what we can, or what we can not.
    - Because we are curious
    - Because we are lazy
    - Because we are not so able

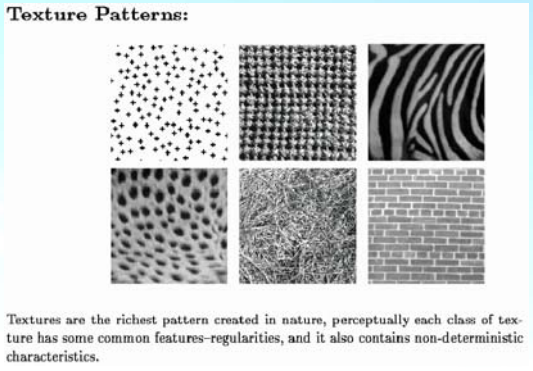
模式 与 模式识别
Pattern and *Pattern Recognition*

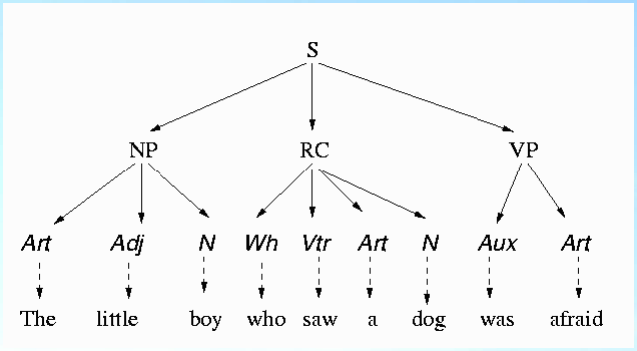
何谓“模式（Pattern）”？

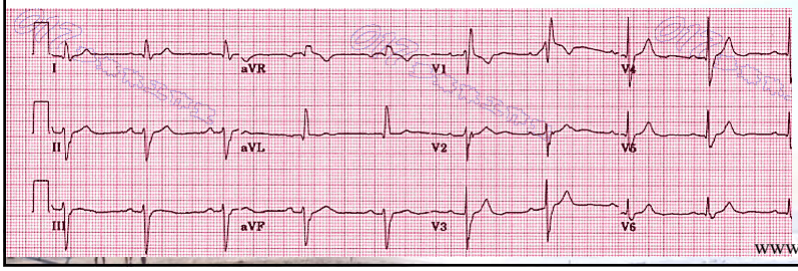
《说文》：
- 模，法也。
- 式，法也。

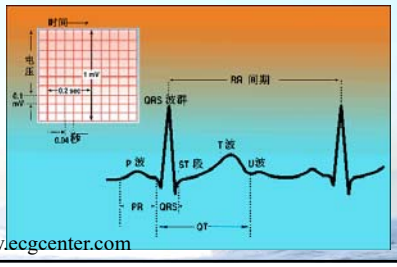
《现代英汉词典》：
- 拉丁词根 pat（父亲），父是子的“模型”
- 图案，花样；方式；样品；型，式样，纸样，模型；模范，典型；
- 模式
  - A physical arrangement of elements
  - Repeating; with some degree of correspondence in successive trials or observations

模式的例子









www.ecgcenter.com

模式的例子

- 社会模式：
  - 信用：良好 vs. 不良
  - 保险：驾驶习惯、出险风险
  - 信息服务：兴趣、喜好
  - 择偶：我的菜 vs. 不是我的菜
  - 文体：散文、小说、说明文、议论文、…
  - 性格：内向、外向、女汉子、萌妹子、娘炮、直男、…
  - 文化：东方、西方、…
  - 课堂风格：
  - 政治：国王制、君主立宪制、议会制、人民代表大会制、…

什么是“模式（Pattern）”？

- 对象的组成成分或因素之间存在的直接或间接的规律性关系

or

- 存在确定性或随机规律的对象、过程或事件的集合

什么是“识别（recognition）”？

- 《说文》
  - 识，知也。《回乡偶书》：儿童相见不相识,笑问客从何处来。
  - 别，分解也。《荀子·君道》：知国之安危臧否,若别白黑。
- 《现代英汉词典》
  - The act or process of identifying (or associating) an input with one of a set of known possible alternatives
- 《美国传统辞典》
  - An awareness that something perceived has been perceived before.

何为“模式识别”？

- Pattern Recognition
  - the recognition of patterns
- To see something 1 as something 2
- 通过对事物的观察对其**某种性质**的认识

尤指分类性质

模式识别研究的范畴

完全确定　　完全随机

解决模式识别问题的几类方法

- 基于知识的方法 (Knowledge-based)
  - AI、专家系统(Expert Systems)
  - 句法（结构）模式识别 (Syntax PR or Structural PR)
- 基于数据的方法 (Data-based)
  - 统计模式识别方法（Statistical PR）
  - 人工神经网络(ANN)、支持向量机(SVM)、…
  - 各种深度学习方法
- 混合方法(Hybrid Methods)

机器学习

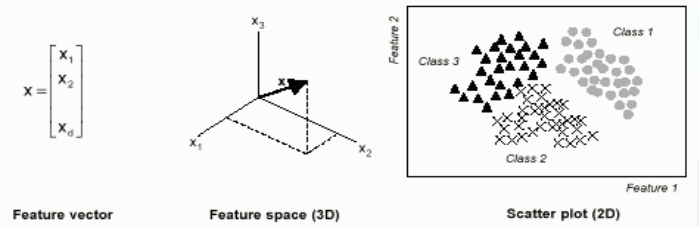
休息1分钟

概念和名词约定

- 样本sample：待研究对象的个体，包括性质已知或未知的个体
  （注意：统计学中有略微不同的用词习惯）
- 类别class：将所研究的样本性质离散化为有限的类别，认为同一类的样本在该性质上是不可区分的
  - 习惯上，类别用ω_i表示，如ω_1、ω_2，有时也用$\{-1,1\}$或$\{0,1\}$等表示
- 已知样本known samples：类别情况已知的样本
- 未知样本unknown samples：类别情况未知的样本
- 样本集sample set：若干样本的集合，分已知样本集和未知样本集

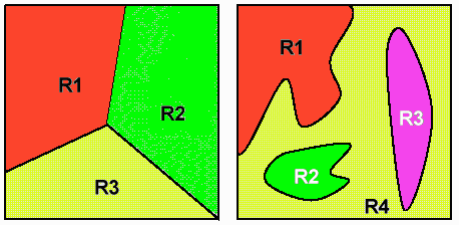
概念和名词约定

- 特征features：样本的任何可区分的（且可观测的）方面
  - 包括定量特征和定性特征，但通常最后转化为定量特征
- 特征向量feature vectors：样本的所有特征组成的 *n* 维向量
  是样本在数学上的表达，因此也称作**样本**
- 特征空间feature space：特征向量所在的 *n* 维空间，每一个样本（特征向量）是该空间中的一个点，一个类别是该空间中的一个区域



概念和名词约定

- 分类器classifier：能够将每个样本都分到某个类别中去（或者拒绝）的计算机算法
- Decision region: 分类器将特征空间划分为若干区域（决策域）
- Decision boundary: 不同类别区域之间的边界称作分类边界、决策边界或分类面、决策面



概念和名词约定

- 分类器/判别函数

$$\mathbf{y} = f(\mathbf{x})$$

预测 分类器 样本

$$X = [\mathbf{x}_1, \cdots, \mathbf{x}_m] = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}$$

特征

样本

- 当预期输出是离散分类时，机器学习就是模式识别

一个模式识别/机器学习系统的典型构成

objects

sensor

Feature extractor

Classifier Predictor Generator

OUTPUT (decision, action, …)

一个模式识别/机器学习系统的典型构成

原始观测获取

特征提取与选择
流形学习

监督学习方法：线性判别、感知器、罗杰斯特回归、ANN、SVM、决策树、贝叶斯决策、贝叶斯网络、HMM、…
非监督学习方法：C均值聚类、分级聚类、SOM、混合高斯模型、…

objects

sensor

Feature extractor

Classifier Predictor Generator

OUTPUT (decision, action, …)

经典方法：感知器、ANN、SVM、…
深度学习：CNN、RNN、LSTM、AE、DNN、RBM、GAN、VAE…

模式识别与机器学习系统举例

健康医疗数据的机器学习应用探索举例

闾海荣
lvhairong@tsinghua.edu.cn

健康医疗数据的机器学习应用探索举例

比如，用门诊病历预测MSA (multiple system atrophy)

主　诉：行走不稳3年
现病史：3年前自觉一次腹泻后行走不稳，下肢发沉，全身无力，头晕，下午明显。左侧轻度笨拙，逐渐加重。便秘1年多，尿频急1年多。3年前梦中喊叫，坠床，最近减少。睡眠呼吸暂停。
家族史：无类似疾病患者。
既往史：一直梦话多，喊叫，易醒。
查　体：神清，构音障碍。眼动充分，未见眼震。四肢肌力V级，肌张力可，腱反射对称未引出，右下肢外翻，病理征(+)。指鼻跟膝胫轻度不稳。姿势反射消失。

主　诉：走路不稳半年
现病史：患者近半年逐渐出现行走不稳，醉酒感，快走时出现头晕，不转。上下楼梯较前困难，无饮水呛咳。排便无力十余年，尿无力，有时梦中喊叫。1年前曾感觉无力，持续2个月，半年前出现类似无力，持续1个月。
家族史：无
既往史：颈椎病。近1年间断乏力感。一直血压偏低。
查　体：神清语利，眼动充分，未见眼震。四肢肌力V级，肌张力可，腱反射对称引出，病理征（-）。指鼻跟膝胫稳，步态基本正常，串联步态轻度不稳。


江瑞


顾卫红

NLP

12,000 dim features (0.00,0.86,0.58,…,0.75)

Learning Machine

Ensemble Learning: DT, LR, SVM, RF

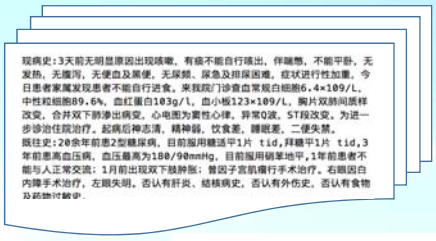
	ACC
DT	94.42
LR	81.81
RF	**97.53**
SVM	78.78
Ensemble	**98.78**

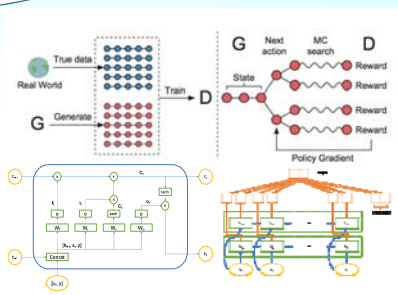
很可能是多系统萎缩

不可能是多系统萎缩

健康医疗数据的机器学习应用探索举例


关嘉麒


现病史：3天前无明显原因出现咳嗽，有痰不能自行咳出，伴喘憋，不能平卧，无发热，无腹泻，无便血及黑便，无尿频、尿急及排尿困难，症状进行性加重，今日患者家属发现患者不能自行进食，来我院门诊查血常规白细胞6.4×10⁹/L，中性粒细胞89.6%，血红蛋白103g/l，血小板123×10⁹/L，胸片双肺间质样改变，合并双下肺渗出病变，心电图为窦性心律，异常Q波，ST段改变。为进一步诊治住院治疗。起病后神志清，精神弱，饮食差，睡眠差，二便失禁。
既往史：20余年前患2型糖尿病，目前服用糖适平1片 tid，拜糖平1片 tid，3年前患高血压病，血压最高为180/90mmHg，目前服用硝苯地平，1年前患者不能与人正常交流：1月前出现双下肢肿胀；曾因子宫肌瘤行手术治疗。右眼因白内障手术治疗，左眼失明。否认有肝炎、结核病史，否认有外伤史，否认有食物及药物过敏史。



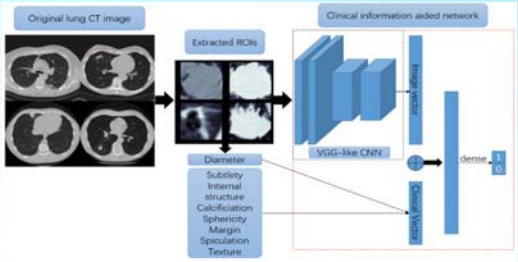
仿真肺炎病例

1. 患者于 1 周前无明显诱因出现咳嗽，咳白色粘痰，伴活动后加重，休息后可缓解，间断服用镇咳药物等治疗，未行正规诊治。
2. 患者 9 年前受凉后出现咳嗽、咳痰、气喘，活动后加重，喘憋明显，给予抗感染等对症治疗后好转，后进食后症状可改善。
3. 患者 1 年前无明显诱因出现咳嗽、咳痰伴气喘，多于受冷或天气转凉时发作，持续时间在 2 周至半年不等，经抗感染等治疗后好转。无明显诱因，冬季多发。

仿真肺癌病例

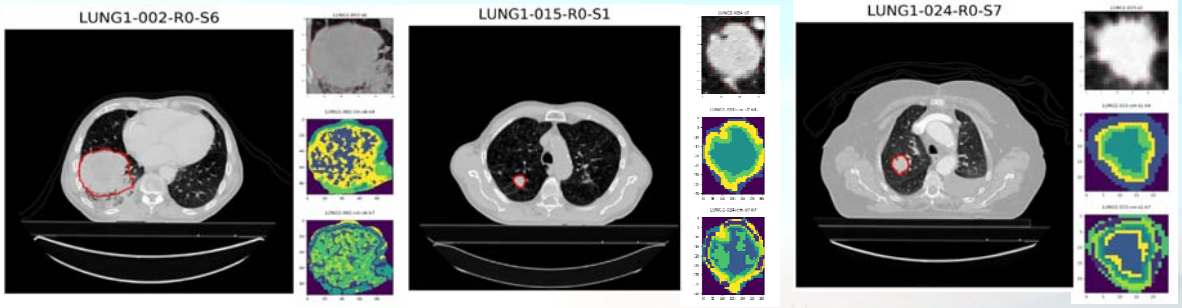
1. 患者反复出现饮食呛咳，无咯血、胸痛、咯血，就诊于北京友谊医院查胸部 CT 示右肺中叶占位性病变，因病变较小定期，于急诊后间断咳嗽、咳少量白痰。
2. 患者 10 多年前开始出现咳嗽、咳痰，痰中带血，当地医院查胸部 CT 示纵隔肿大淋巴结，右肺下叶结节，后于北京人民医院行气管镜时治疗收入院。
3. 患者间断用中药治疗，偶有咯血。4 年前体检症状右侧酸痛，无胸闷、胸痛，无心悸、头晕。2015-9-2 出现发热。

健康医疗数据的机器学习应用探索举例



	Accuracy	AUC
6-Layer CIAnet	89.10	0.9412
12-Layer CIAnet	93.32	0.9709
14-Layer CIAnet	95.04	0.9892

李想之、方翔



本章知识点

- 模式识别和机器学习的基本概念
- 基本术语：样本、特征、类、分类面、…


单选题 1分
设置
休息4分钟，回到座位后请答题
A
已回座位
B
还没有
Xuegong Zhang
提交
61