数据挖掘与机器学习

潘斌

panbin@nankai.edu.cn

范孙楼227

课程安排

- 讲授+实验
- 1-18周，每周讲授3课时
- 第5周起安排上机课程，共5次实验+1次课程大作业
- 成绩=闭卷考试（60）+实验（20）+大作业（20）

实验安排

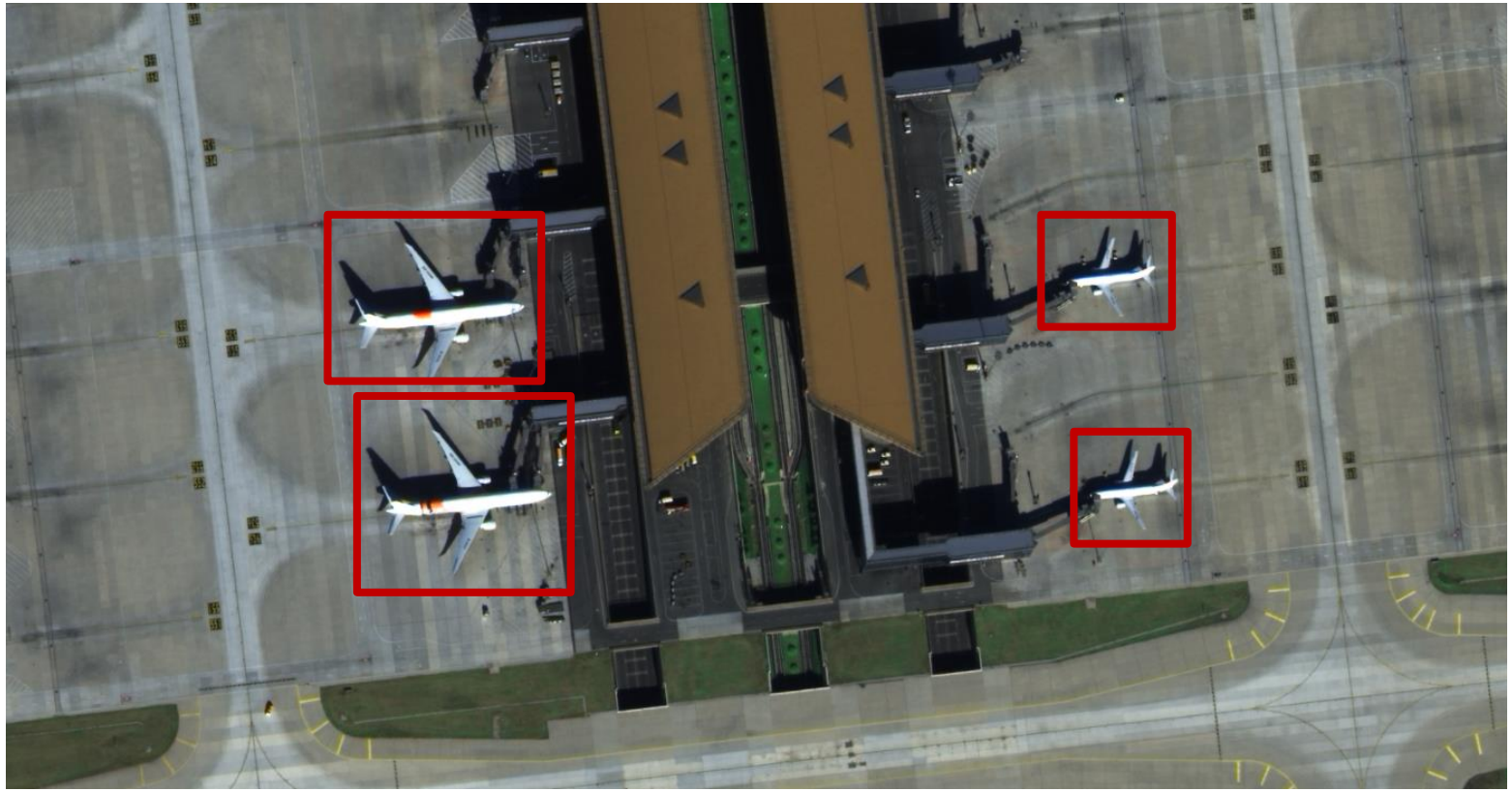
- 利用本课程学习的算法编程完成5项数据挖掘/机器学习任务（待续），并提交实验报告1份（11.27/12.4/12.11）
- 9-13周安排上机课，助教答疑
- 14、15、16周实验课随堂考核，针对报告提问

课程大作业

- 实现一项全链路机器学习任务（如目标检测、文字识别等）
- 做课堂PPT展示，汇报研究思路、算法设计、结果，20-25分钟，暂定15-18周
- 分组完成，5-7人一组，汇报时说明每个人的贡献

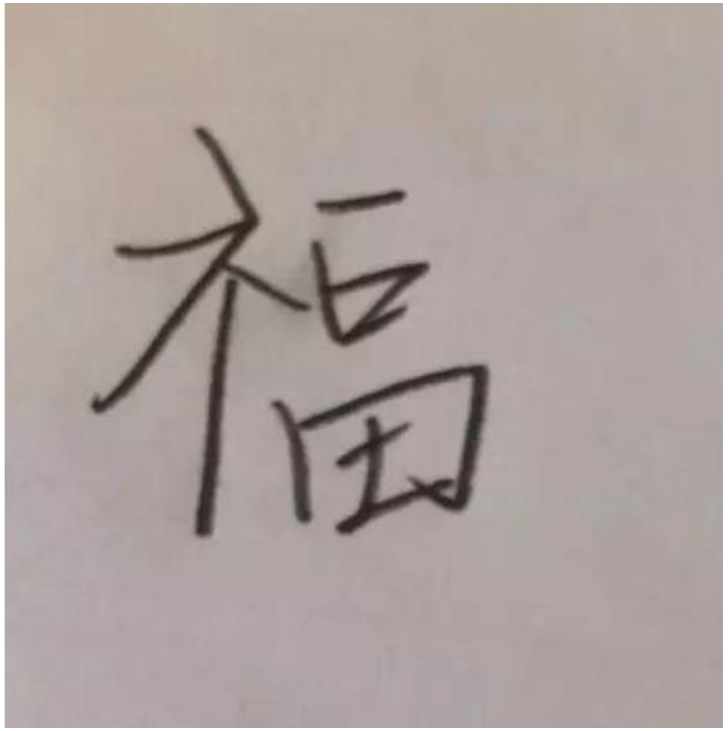
课程大作业（四选一，去年的）

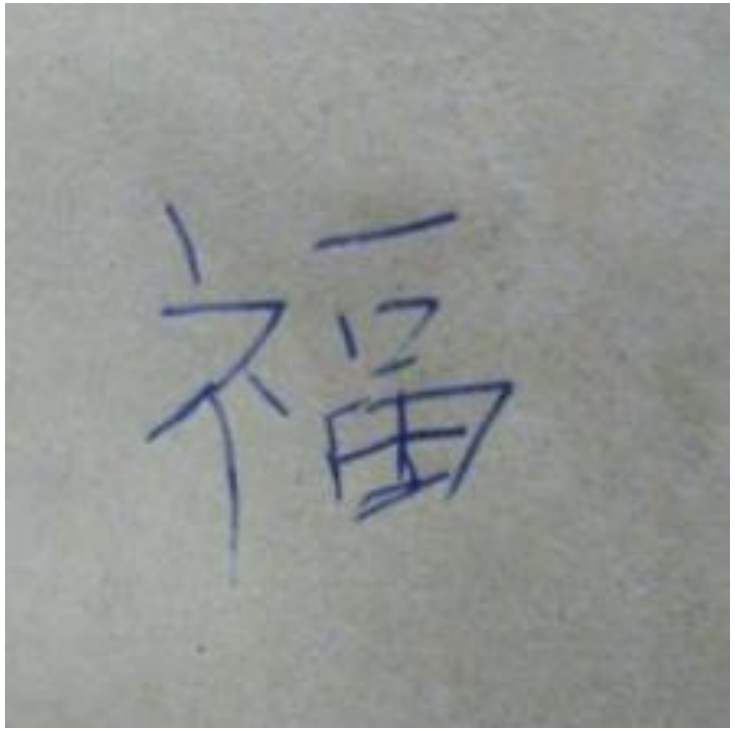
- 任务1：遥感图像飞机检测



课程大作业（四选一）

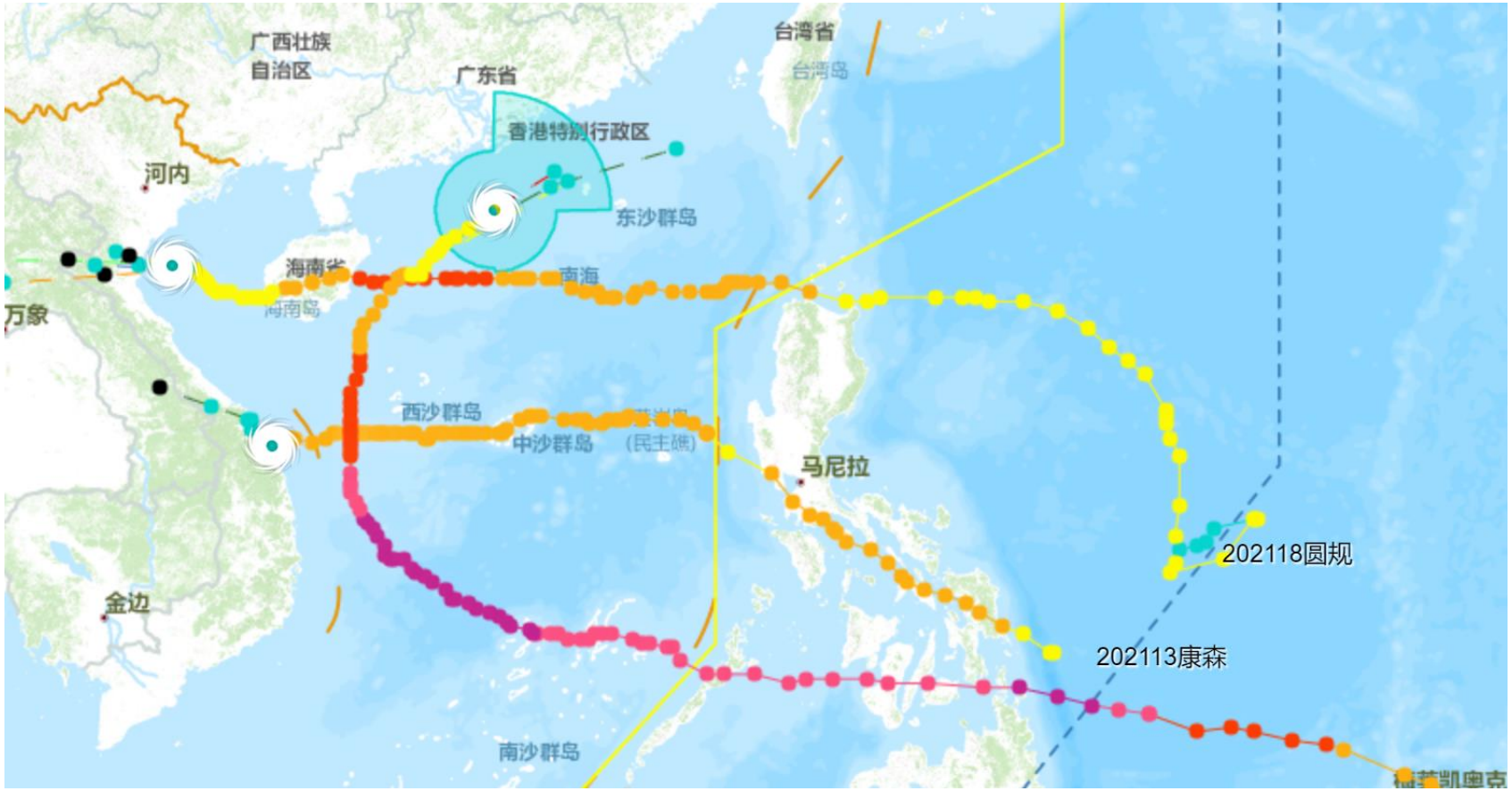
- 任务2：“福”字识别-解决类别不平衡问题





课程大作业（四选一）

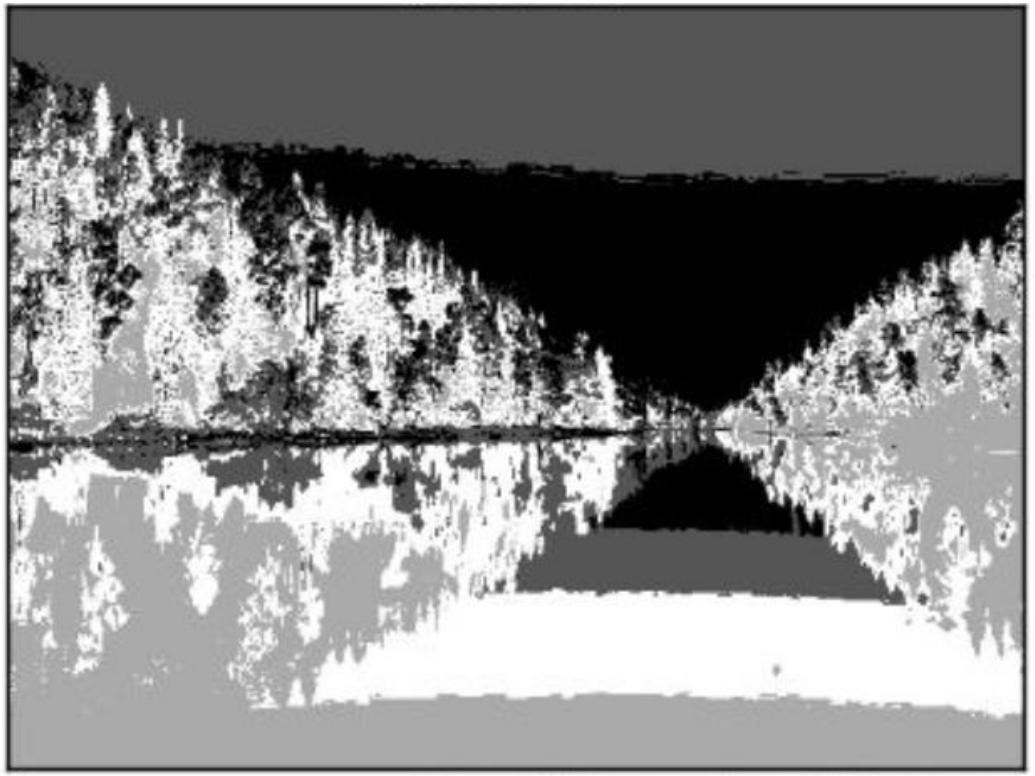
- 任务3：台风预报



66666 1815 21 0018 1815 0 6 LEEPI
2018081012 1 174 1448 1006 13
2018081018 1 179 1444 1004 15
2018081100 1 185 1441 1004 15
2018081106 1 192 1438 1004 15
2018081112 2 200 1435 1000 18
2018081118 2 208 1432 995 20
2018081200 2 217 1427 990 23
2018081206 2 227 1420 990 23
2018081212 3 235 1415 982 28
2018081218 3 245 1407 982 28
2018081300 3 253 1398 982 28

课程大作业（四选一）

- 任务4：图像区域分割提取-如何保持空间相关性?



课程大作业

- 无需实验报告，只PPT展示即可
- 上述任务随机4选1，优先遵照意愿
- 第11周前完成分组（11.6），可在飞书群里自行组队

- 助教：潘奕安

参考文献

- 李航，统计学习方法，清华大学出版社
- 周志华，机器学习，清华大学出版社
- 张学工，模式识别，清华大学出版社
- 其他文献……
- 网络资源……

绪　论

Overview of Data Mining & Machine Learning

- Intel公司创始人之一戈登·摩尔（Gordon Moore）发现芯片的容量每18~24个月增加一倍。他据此推断，按此趋势发展下去，在较短时间内计算能力将呈指数增长。——“摩尔定律”（Moore‘s Law）。
- John Roth在联合国世界电信论坛上又提出了一个关于网络科技的观点：互联网宽带每9个月会增加一倍的容量，但成本降低一半，比芯片的变革速度还快。——“光纤定律”（Optical Law）。

- 1KB（Kibibyte，千字节）=1024B，
- 1MB（Mebibyte，兆字节，简称“兆”）=1024KB，
- 1GB（Gigabyte，吉字节，又称“千兆”）=1024MB，
- 1TB（Terabyte，万亿字节，太字节）=1024GB，
- 1PB（Petabyte，千万亿字节，拍字节）=1024TB，
- 1EB（Exabyte，百亿亿字节，艾字节）=1024PB，
- 1ZB（Zettabyte，十万亿亿字节，泽字节）= 1024EB，
- 1YB（Yottabyte，一亿亿亿字节，尧字节）= 1024ZB，
- 1BB（Brontobyte，一千亿亿亿字节）= 1024YB

天文学

- 2000年斯隆数字巡天（Sloan Digital Sky Survey）项目启动的时候，位于新墨西哥州的望远镜在短短几周内收集到的数据，已经比天文学历史上总共收集的数据还要多。到了2010年，信息档案已经高达1.4×2^{42}字节（200PB）。

沃尔玛

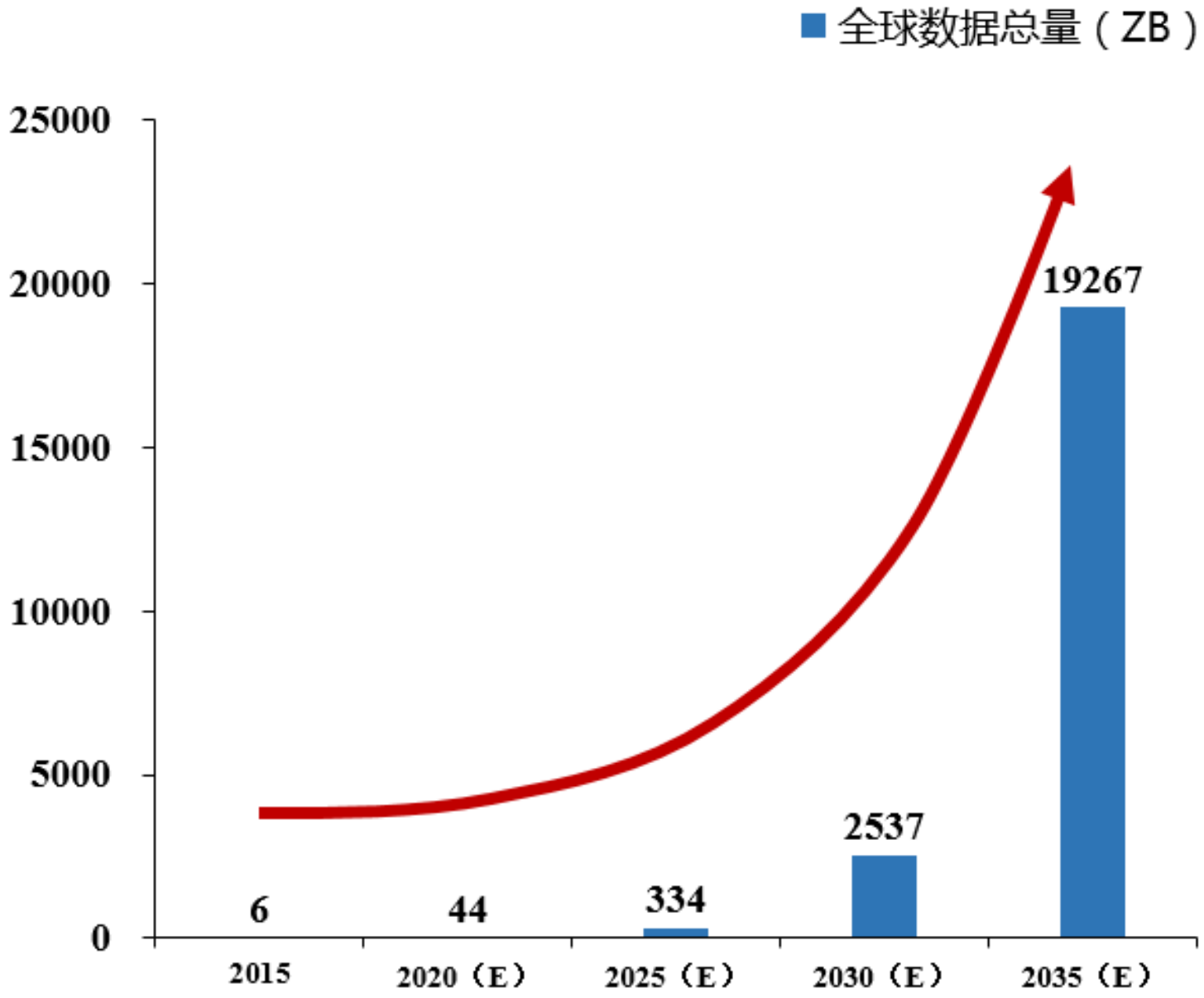
- 沃尔玛每隔1小时就要处理超过100万客户的交易，信息录入量超过2,5PB;

互联网

- 根据Visual Capitalist统计，2019年平均每分钟Facebook有100万账户登录，Youtube有450万次视频播放，谷歌有380万次搜索，以及100万美元的线上消费！
- 短视频、直播、自媒体等新一代互联网数据近年来迎来了大爆发。抖音2016年初创，2018年月活1亿，2019年5亿，2020年8亿，接近微信。


2018 This Is What Happens In An Internet Minute
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
375,000 Apps Downloaded
174,000 Scrolling Instagram
481,000 Tweets Sent
1.1 Million Swipes
tinder
187 Million Emails Sent
936,073 Views
twitch
67 Voice-First Devices Shipped
amazon echo
38 Million Messages
25,000 GIFs Sent via Messenger
2.4 Million Snaps Created
$862,823 Spent Online
NETFLIX
266,000 Hours Watched
Google
3.7 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd


2019 This Is What Happens In An Internet Minute
facebook
1 Million Logging In
18.1 Million Texts Sent
YouTube
4.5 Million Videos Viewed
390,030 Apps Downloaded
347,222 Scrolling Instagram
87,500 People Tweeting
1.4 Million Swipes
tinder
188 Million Emails Sent
1 Million Views
twitch
41 Music Streaming Subscriptions
180 Smart Speakers Shipped
amazon echo
4.8 Million Gifs Served
GIPHY
41.6 Million Messages Sent
2.1 Million Snaps Created
$996,956 Spent Online
NETFLIX
694,444 Hours Watched
Google
3.8 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd



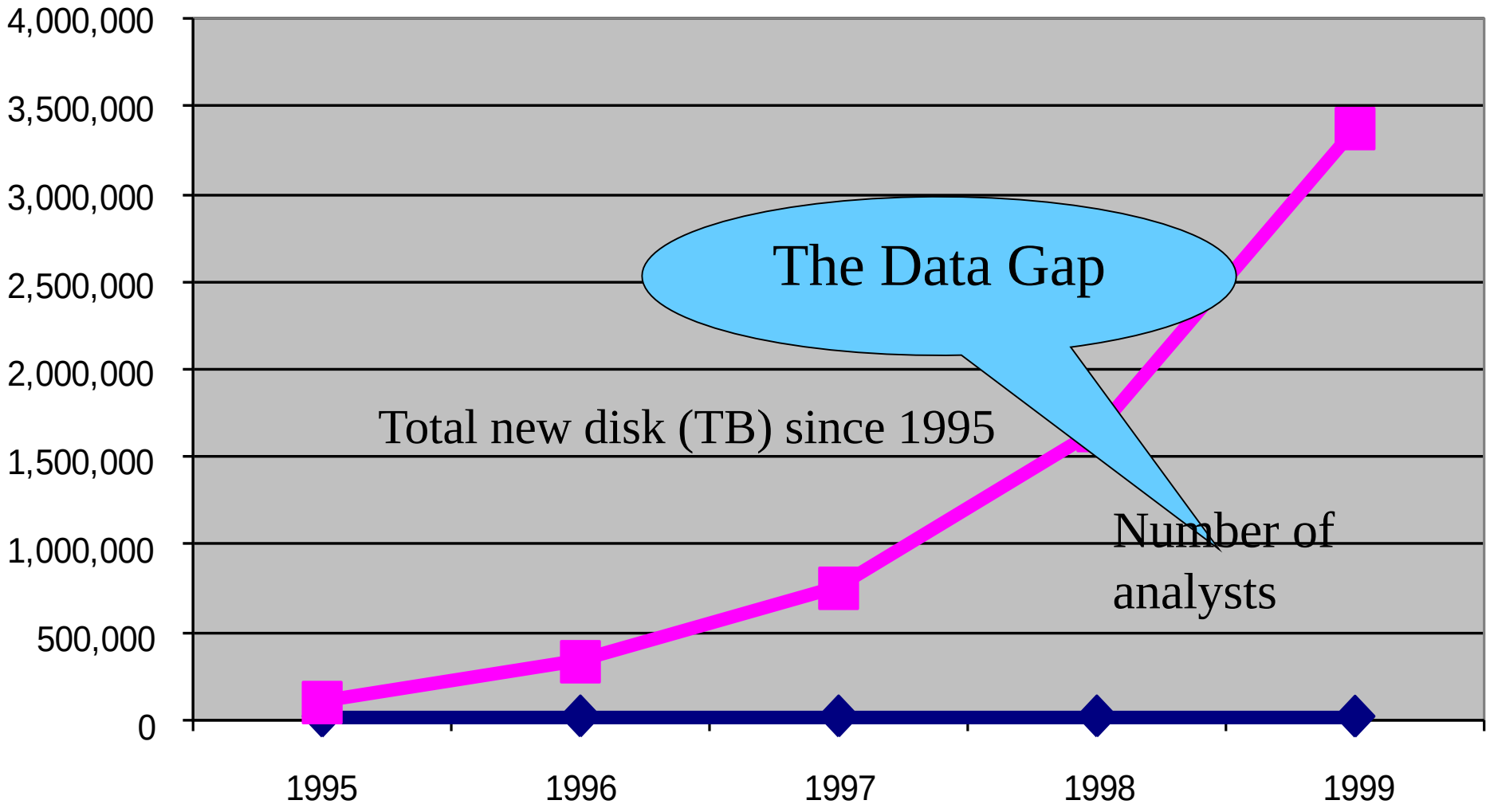
数据来源：IDC，中国电子学会整理

http://www.sohu.com/a/252632116_179557

- 我们已经进入了“Big Data”时代！

大数据的特点（4V）

- 数据量太大Volume
- 数据本身多样性强Variety
- 大数据的产生非常高速Velocity
- 价值低Value——历史数据变成“数据坟墓”


4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
1995
1996
1997
1998
1999
The Data Gap
Total new disk (TB) since 1995
Number of analysts

“We’re drowning in information, but starving for knowledge.”

- John Naisbett

新数据的特点及对分析方法的要求（1）

- 海量数据集越来越普遍
  - 如，处理的数据可能不能放入内存
  - 算法必须是可伸缩的（scalable）
- 常遇到高维数据集
  - 如，生物学领域已产生涉及数千特征的基因表达数据
  - 算法需要具有高维性（稀疏）

新数据的特点及对分析方法的要求（2）

- 很多数据带有异构属性
  - 含有半结构化文本和超链接的Web页面集；描述目标属性是所使用的.doc文件和.mp4文件；包含地球表面不同位置上的时间序列测量值（温度、气压等）的气象数据
  - 算法要能处理这些非传统数据集

新数据的特点及对分析方法的要求（3）

- 需要分析的数据并非存放在一个站点或归属一个机构，而是地理上分布在属于多个机构的资源中（贵州大数据）
  - 需要开发分布式数据分析技术
    - 降低执行分布式计算所需的信息量
    - 有效地统一从多个资源得到的数据分析结果
    - 处理数据安全性问题

新数据的特点及对分析方法的要求（4）

- 待分析的数据集通常不是精心设计的实验的结果
  - 通常代表数据的时机性样本（opportunistic sample）而非随机性样本（random sample）
- 数据集常涉及非传统的数据类型和数据分布
- 数据分析的任务需要产生和评估数千种假设

- 数据挖掘（Data Mining）是在大型数据存储库中，自动的发现有用信息的过程。
  - 发现先前未知的有用模式（描述任务）
  - 预测未来观测结果（预测任务）

- 数据挖掘：目的
- 机器学习：手段
- 大数据：原料
- 大数据技术?

大数据 vs 数据挖掘

- 数据：大量的、不完全的、有噪声的、模糊的、随机的
- 发现：隐含在其中的、人们事先不知道的、但有潜在的有用信息和知识

数据挖掘过程

- 数据采集
- 数据预处理
- 数据挖掘（机器学习）
- 评价和呈现

描述任务

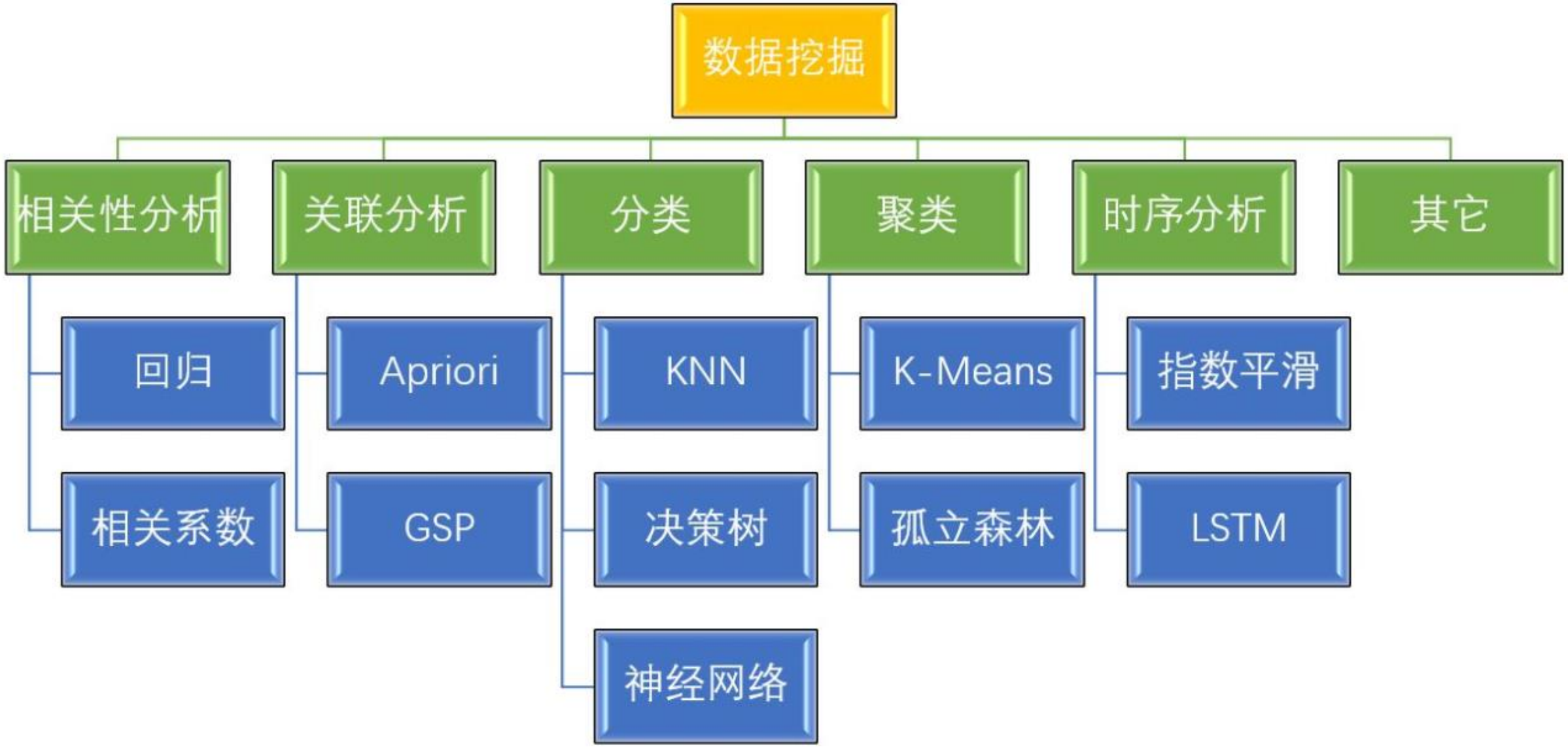
- 导出概括数据中潜在联系的模式（相关、趋势、聚集、轨迹、异常等）
- 本质上是探查性的
- 常需要后处理技术验证和解释结果
- ChatGPT

预测任务

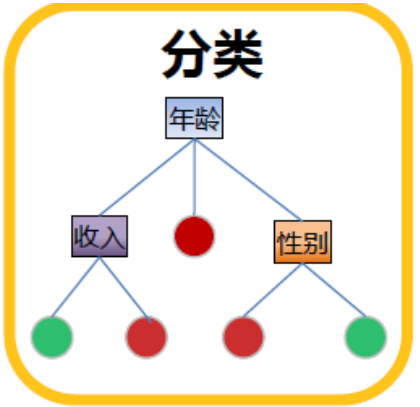
- 根据其他属性的值，预测特定属性的值
- 分类
  - 用于预测离散的目标变量
- 回归
  - 用于预测连续的目标变量

主要数据挖掘任务

- 分类（Classification） [Predictive]
- 聚类（Clustering） [Descriptive]
- 关联规则（Association Rule Discovery） [Descriptive]
- 回归（Regression） [Predictive]
- 序列模式识别（Sequential Pattern Discovery） [Descriptive]
- 异常检测（Anomaly Detection） [Predictive]


数据挖掘
相关性分析
关联分析
分类
聚类
时序分析
其它
回归
相关系数
Apriori
GSP
KNN
决策树
神经网络
K-Means
孤立森林
指数平滑
LSTM

分类



- 给定一个记录集合
  - 每条记录（样本）由一些属性（特征+标签）组成，其中一个属性为类别。
  - $(x_1, x_2, \ldots, x_n, c)$
  - 找到一个将类别属性表示为其他属性的函数的模型（如$c = f(x)$）。
  - 目标：未见过的记录尽可能准确地被分类。

分类步骤

- 给定的数据集被分成独立的训练集和验证集，训练集用于建立模型，而验证集用于检验该模型。
- 模型创建：对一个类别已经确定的数据创建模型。
- 模型检验与应用：对创建的模型进行评价，用其预测未来或者类别未知的记录。

- [例] 某鱼类加工厂生产线混合生产鲑鱼、鲈鱼两种鱼类罐头，需要对原材料鲑鱼、鲈鱼进行人工分拣（分类识别）。
- 若用一机器学习系统来进行自动分拣，采用什么样的识别方法和识别系统?

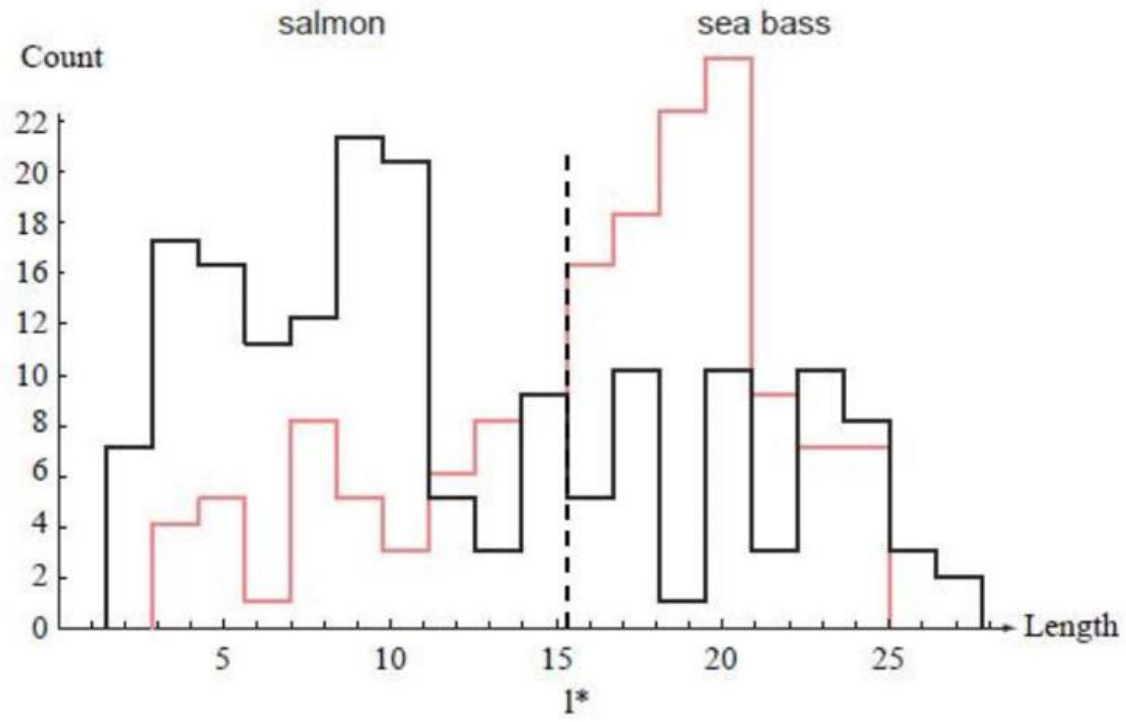
问题和要求

机器学习系统概述

- 水产工人目视识别的参考依据
  - 长度差异（鲑鱼相对鲈鱼短一些）
  - 宽度差异（鲑鱼相对鲈鱼窄一些）
  - 亮度差异（鲑鱼相对鲈鱼暗一些）
  - 鳞片形状
  - 嘴的形状
  - 等等

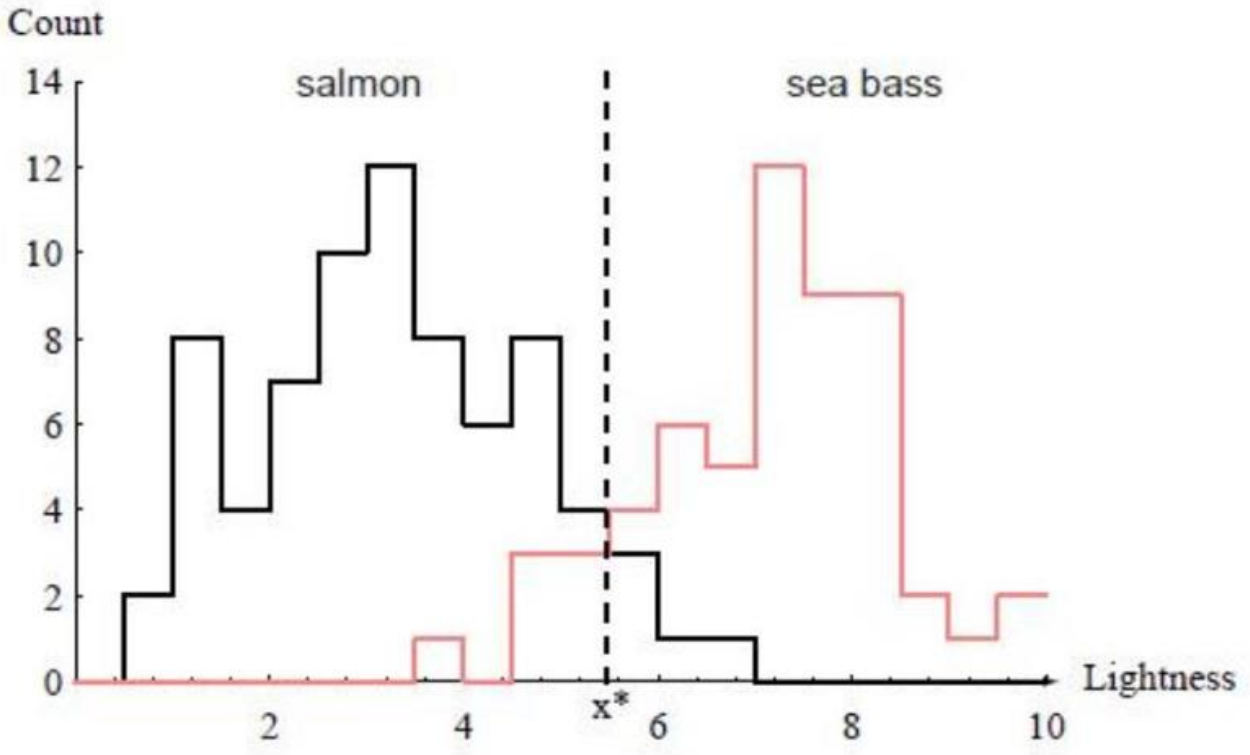
机器学习系统概述

- 长度差异（样本统计图）



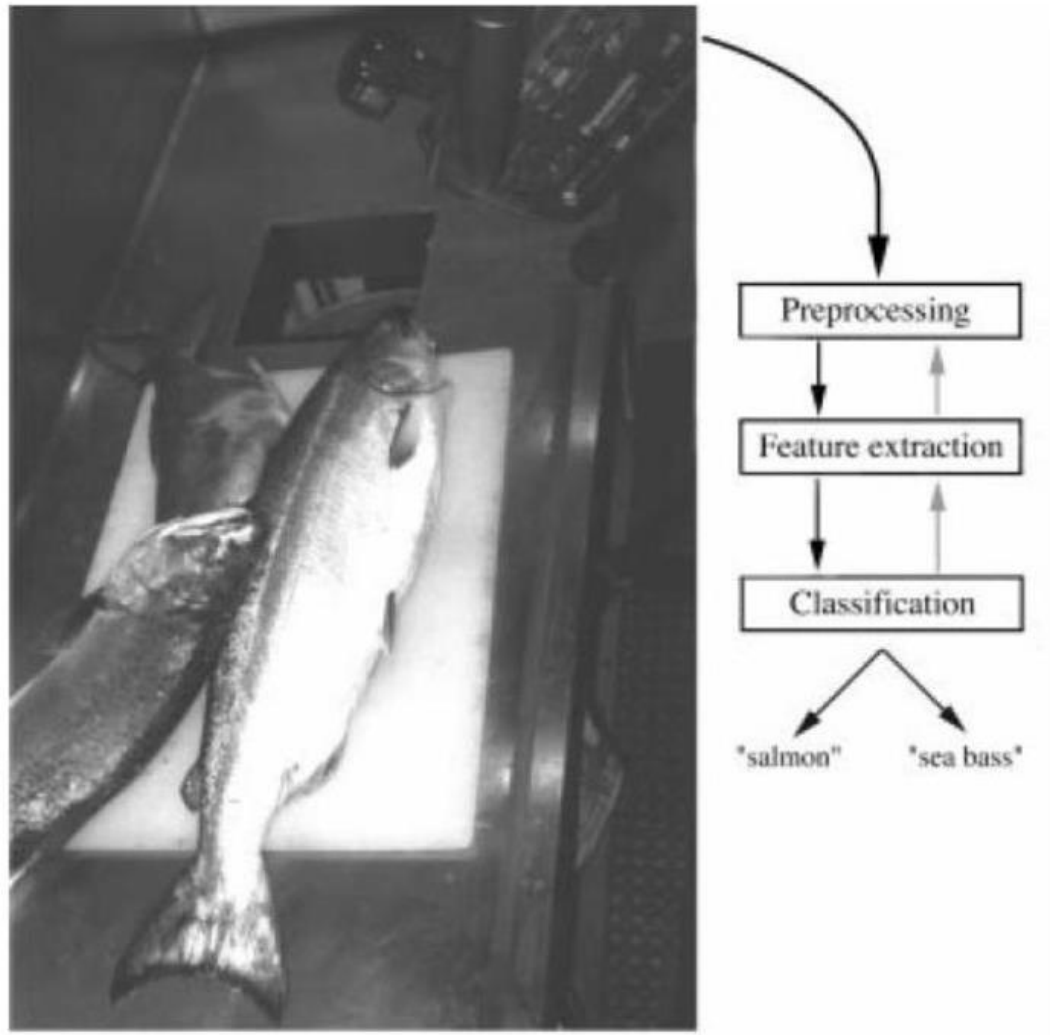
机器学习系统概述

- 亮度差异（样本统计图）



机器学习系统概述

- 自动识别系统的主要组成部分
  - 数据源：通过相机拍摄获取数字图像
  - 预处理：采用图像处理算法提取鱼的区域
  - 特征计算：采用图像处理算法计算鱼的长度、宽度等
  - 分类：利用已知的两种鱼的长度、亮度等差异进行判别

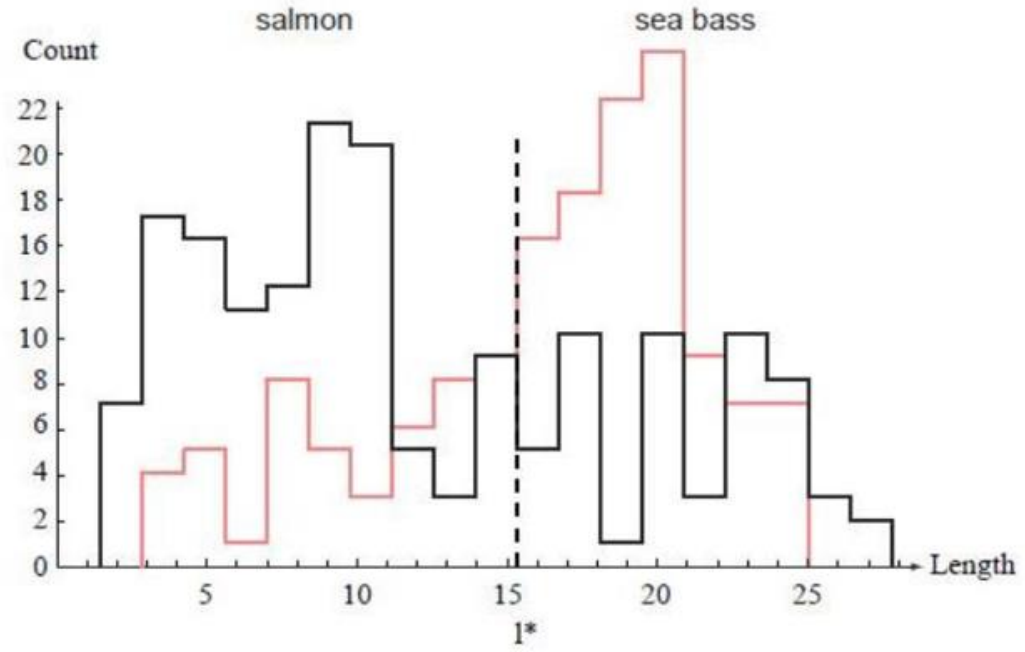

Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

应用问题抽象化

- 类别抽象化
  - 鲑鱼——ω_1类
  - 鲈鱼——ω_2类
- 分类依据抽象化
  - 若采用长度来区分，则定义长度为特征，记为x

应用问题抽象化

- 分类依据抽象化
  - 已知目视识别经验
    - 鲑鱼相对较短——ω_1类x值偏小
    - 鲈鱼相对较长——ω_2类x值偏大



应用问题抽象化

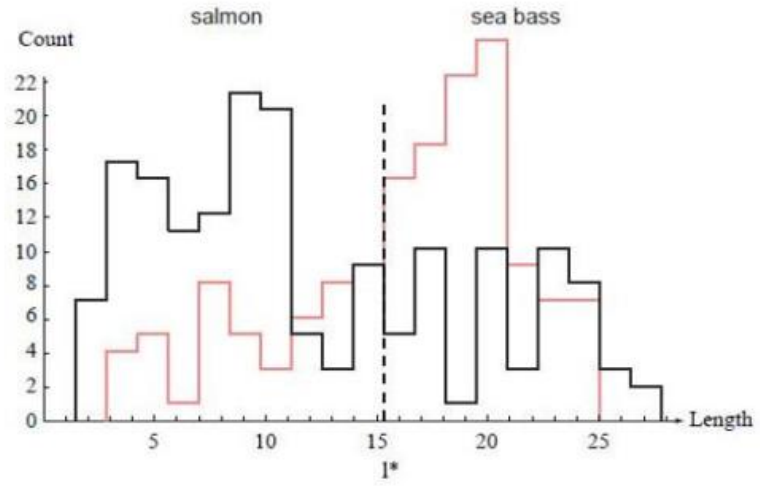
- 分类方法抽象化
  - 定义长度特征x，统计对象的亮度分布，求出临界点x_0（阈值）
  - 从而确定分类的过程
    - 如果$x < x_0$，则$x \in \omega_1$类（鲑鱼）
    - 如果$x > x_0$，则$x \in \omega_2$类（鲈鱼）
  - 上述过程称为决策规则

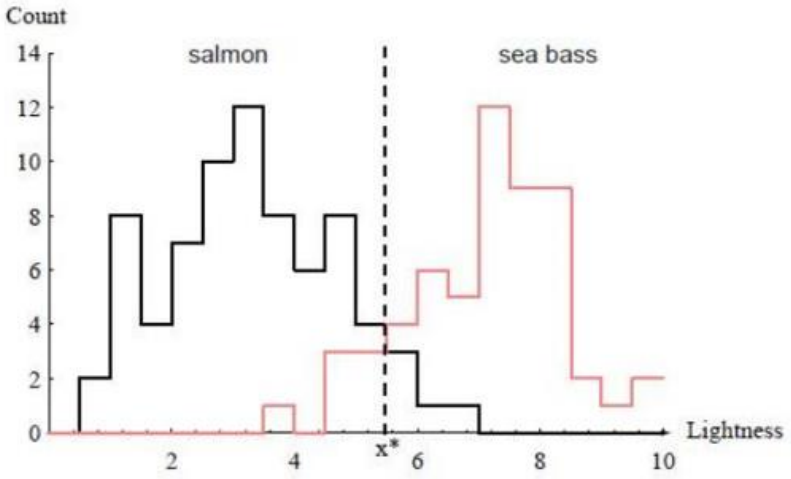
应用问题抽象化

- 特征需要优选
  - 采用长度特征
  - 采用亮度特征

  - 尽量多的选用特征（通常可以提高识别率）
  - 尽量少的选用特征（减少设计和计算难度）

应用问题抽象化


salmon
sea bass
Count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
Length
l*


Count
14
12
10
8
6
4
2
0
salmon
sea bass
2
4
x*
6
8
10
Lightness

应用问题抽象化

- 优选结果案例
  - 若采用亮度特征x_1和宽度特征x_2
  - 从而确定分类的过程
    - 如果……，则$x \in \omega_1$类
    - 如果……，则$x \in \omega_2$类

示例

- 待分析数据集

ID	X1	X2	Y
1	A	7	T
2	A	7	T
3	B	5	T
4	A	3	F
5	A	2	F
6	A	6	F
7	A	7	F
8	A	7	T
9	B	2	T
10	A	3	F

- 训练集

ID	X1	X2	Y
1	A	7	T
2	A	7	T
4	A	3	F
6	A	6	F
9	B	2	T
10	A	3	F

- 验证集

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	F
8	A	7	T

- 测试集

模型创建

ID	X1	X2	Y
1	A	7	T
2	A	7	T
4	A	3	F
6	A	6	F
9	B	2	T
10	A	3	F

- 模 型 检 验（参数、超参数）

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	F
8	A	7	T

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	T
8	A	7	T

- 模 型 应 用

- 不 能 在 测 试 集 调 参 !

常用分类方法

- 决策树
- KNN
- SVM
- ANN
- Bayes分类
- ……

关联分析

某超市Pos机上记录如下的销售数据：

顾客	购买商品
1	面包，黄油，尿布，啤酒
2	咖啡，糖，小甜饼，鲑鱼，啤酒
3	面包，黄油，咖啡，尿布，啤酒，鸡蛋
4	面包，黄油，鲑鱼，鸡
5	鸡蛋，面包，黄油
6	鲑鱼，尿布，啤酒
7	面包，茶，糖鸡蛋
8	咖啡，糖，鸡，鸡蛋
9	面包，尿布，啤酒，盐
10	茶，鸡蛋，小甜饼，尿布，啤酒

从这个销售数据中可以得出什么结论？（因果推断）

关联分析

- 给定一系列记录，找到其中隐含的令人感兴趣的联系，用以根据记录中某些项的出现来预测其他项的产生。
- 在关联规则挖掘算法中，通常的目的是发现数据中强关联特征的模式。
- 所发现的模式通常用蕴含规则或特征子集的形式表示。
- 目标是以有效的方式提取最有趣的模式。

示例

- 购物篮分析

商场购物篮事务

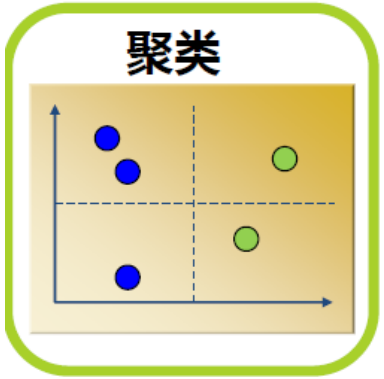
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

关联规则：

{Diaper} →{Beer}

{Bread} →{Milk}

聚类

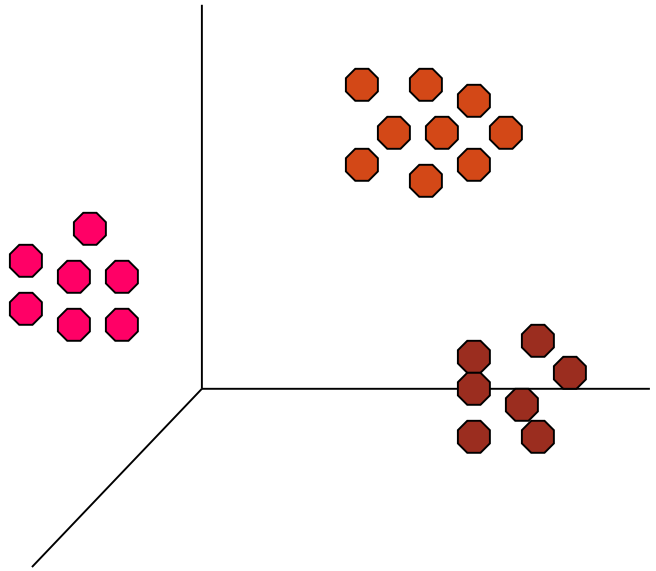


- 聚类是指一组彼此间非常“相似”的数据对象的集合。每一个类中的数据相近，不同类之间的数据相差较大。
- 通过聚类分析可以根据部分数据发现规律，找出对全体数据的描述，可用于异常点检测。

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



- 特征空间构建
- 距离度量方式

异常检测

- 识别特征显著不同于其他数据的观测值（异常点或离群点）。
- 应用于检测信用卡欺诈、网络攻击、疾病的不同模式、生态系统扰动等。

回归

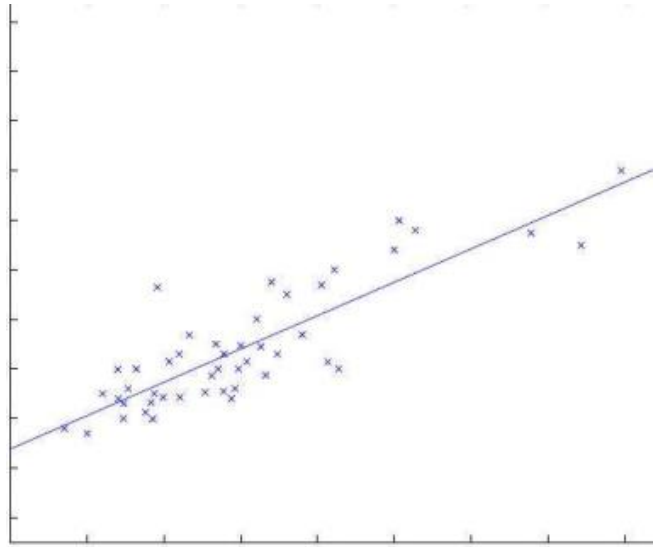
- 监督学习，给定输入x和输出y，任务是学习从输入到输出的映射y=f(x|θ)。

线性回归试图学得

$$f(x_i) = wx_i + b, \text{ 使得 } f(x_i) \simeq y_i \ .$$

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

$$= \arg\min_{(w,b)} \sum_{i=1}^{m} (y_i - wx_i - b)^2 \ .$$



序列模式发现

- 主要用于分析数据仓库中的某类与时间相关的数据，搜索类似的序列或子序列，并挖掘时序模式、周期性、趋势和偏离等。
- 例如，它可以导出类似“若A股票连续上涨两天且B股票不下跌，则第三天C股票上涨的可能性为75%”的数据关系。
- 序列模式可以看成是一种特定的关联模型，它在关联模型中增加了时间属性。

监督学习和无监督学习

- 监督学习（Supervised Learning）
  - 训练集带有类标签，新的数据基于训练集进行分类。
  - 如回归、分类等
- 无监督学习（Unsupervised Learning）
  - 数据集没有类标签，提供一组属性，然后寻找出数据集中存在的类别或者聚集
  - 如聚类、降维等

监督学习和无监督学习

Supervised vs Unsupervised

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, generative models, etc.

数据挖掘应用

- 数据新闻让英国撤军
  - 2010年10月23日《卫报》利用维基解密的数据做了一篇“数据新闻”。将伊拉克战争中所有的人员伤亡情况均标注于地图之上。地图上一个红点便代表一次死伤事件，鼠标点击红点后弹出的窗口则有详细的说明：伤亡人数、时间，造成伤亡的具体原因。密布的红点多达39万，显得格外触目惊心。一经刊出立即引起朝野震动，推动英国最终做出撤出驻伊拉克军队的决定。

THE END !