数据科学导论

Introduction to Data Science

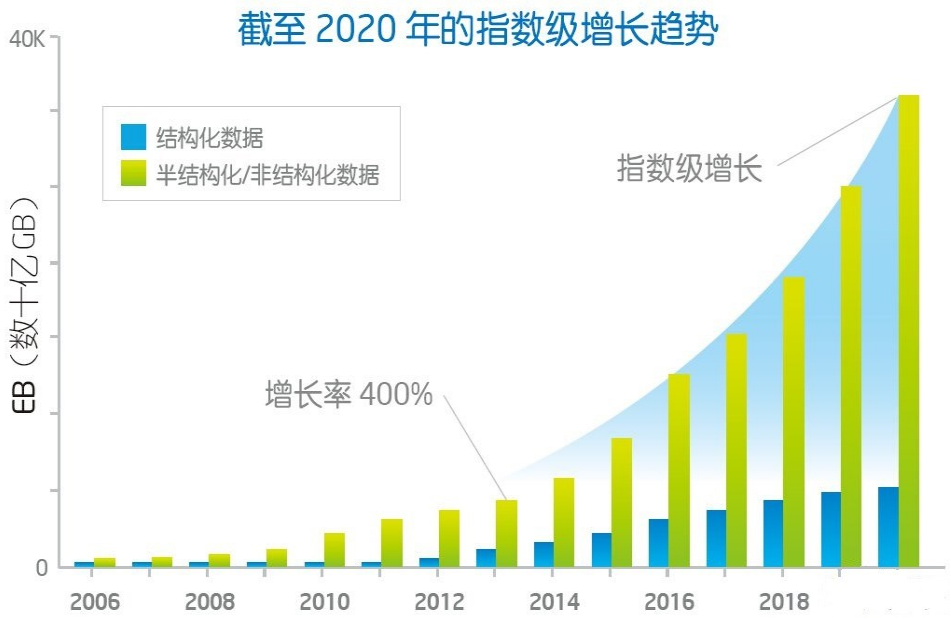
第二章 数据分析基础

黄振亚，陈恩红，刘淇
Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html

回顾：数据分析基础

- 数据采集 Data Collection
- 数据存储 Data Storage
- 数据预处理 Data Preprocessing
- 特征工程 Feature Engineering



回顾：数据预处理

- 大数据环境下的数据特征
- 为什么需要进行预处理
- 预处理的基本方法
  - 数据清理
  - **数据集成**
  - **数据变换**
  - 数据规约

回顾：数据集成

数据的相关性分析

- **无序数据：每个数据样本的不同维度是没有顺序关系的**
  - 余弦相似度、相关度、欧几里得距离、Jaccard
- **有序数据：对应的不同维度(如特征)是有顺序(rank)要求的**
  - 在信息检索中，如何判断不同检索方法返回的页面序列的优劣
  - 在推荐系统中，如何判断不同推荐序列的好坏
    - **Spearman Rank(斯皮尔曼等级)相关系数**
    - **标准化的折损累计增益(NDCG)**
    - 肯德尔相关性系数
      - kendall correlation coefficient
- 课外阅读：PageRank算法

i	相关度
1	3
2	3
3	2
4	0
5	1
6	2

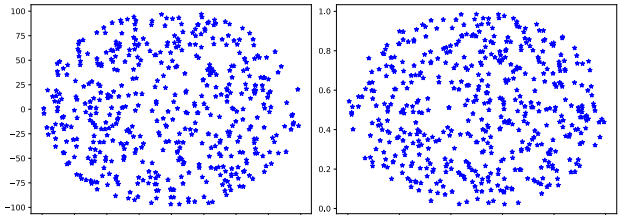
方法返回结果

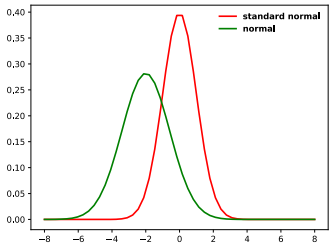
i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

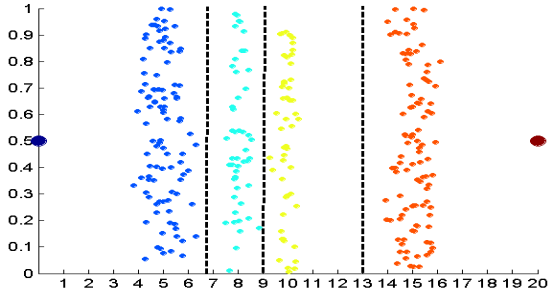
真实结果

回顾：数据变换

- 数据变换的目的是将数据转换成适合分析建模的形式
  - 前提条件：尽量不改变原始数据的规律
  - 数据规范化
    - 最小-最大规范化
    - z-score规范化
    - 小数定标规范化
  - 数据离散化
    - 非监督离散化
    - 监督离散化







回顾：数据变换-规范化

- 小结

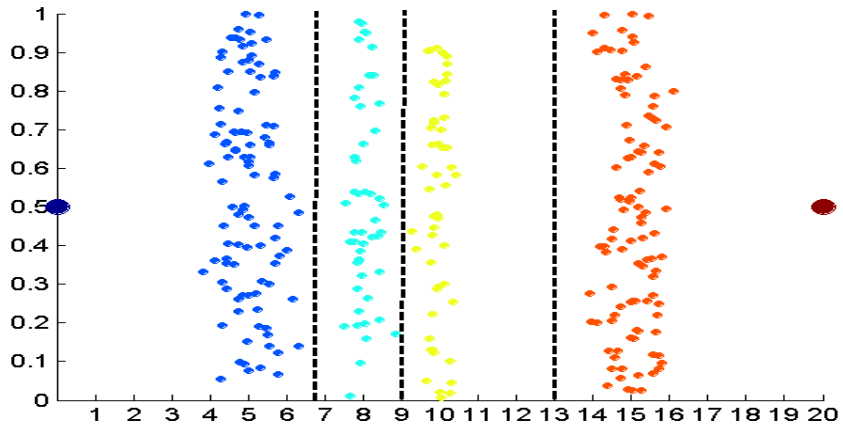
	优点	缺点	适用场景
最小-最大规范化	保留了原始数据中存在的关系，是消除量纲和数据取值范围影响的最简单方法	对最大最小值敏感，新数据加入时，可能改变最大最小值，需重新计算	适用于原始数据不存在很大/很小的一部分数据的时候
z-score规范化	算法简单方便，结果方便比较，应用于数值型的数据，且不受数据量级的影响	总体平均值和方差不一定可知，在一定程度上要求数据分布，结果没有具体意义，只用于比较	适用于最大最小值未知，或者离群点影响较大的时候
小数定标规范化	算法实现简单	不适用于不同含义数据的比较，无实际意义	使用含义相同的数据，且最大最小相差较大

回顾：数据变换-离散化

- 熵—计算不确定性以及不纯性
  - 假设数据已经离散，计算离散后的某个区间 ***t*** 中的熵:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

  - 其中，$p(j \mid t)$ 表示 第 j 类在区间 t 中的概率；一般对数log以2为底



数据变换-离散化

- **如何确定分隔点？**--计算分隔后的信息增益
  - 信息增益（Information Gain）：

$$GAIN_{\text{split}} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

    - 信息增益：表示在某个条件下，信息不确定性减少的程度。
    - 父节点 P 被分隔为 K 个区间
    - n 表示总记录数，n_i表示区间 i 中的记录数
  - 确定分隔点 j ：
    - 选择信息增益最大的分隔点，即

$$j = max(GAIN_{\text{split}})$$

数据预处理

- 大数据环境下的数据特征
- 为什么需要进行预处理
- **预处理的基本方法**
  - 数据清理
  - 数据集成
  - 数据变换
  - **数据规约**

数据预处理：数据规约

- 为什么需要进行数据规约？
  - **数据清理、数据集成之后，获得多源且质量完好的数据集**
  - 但数据规模很大：大规模数据样本，使得在整个数据集上进行复杂的数据分析与建模需要**很多计算资源和很长的时间**

- 例：电子商务中的商品类别

  CVPR 2021 AliProducts Challenge: Large-scale Product Recognition
  - 背景：电商企业面临的大规模、细粒度商品图像识别问题
  - 数据量：**300万张图片**，涵盖了**5万个SKU级商品类别**

数据预处理：数据规约

- **数据归约**
  - **目标：缩小数据挖掘所需的数据集规模**
  - 得到数据集的归约表示，它小得多，但可以产生相同的（或几乎相同的）分析结果
  - 用于数据归约的时间不应当超过或“抵消”在归约后的数据上挖掘节省的时间
- 常用的数据规约方法
  - 维度归约
    - 减少所考虑的随机变量或属性的个数
  - 数值归约
    - 用较小的数据表示形式替换原始数据
    - e.g. 使用模型来表示数据

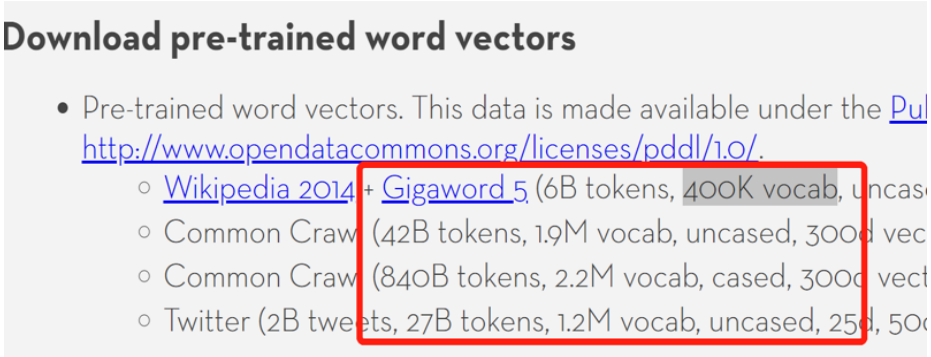
数据规约-维度规约

- **维度规约**
  - 大数据应用包含几万至几百万的属性，其中大部分属性与挖掘任务不相关，是冗余的，
  - **数据降维：**删除不相关的属性，并保证信息的损失最小
- **维度归约方法**
  - 主成分分析
  - 特征子集选择

个人借贷数据

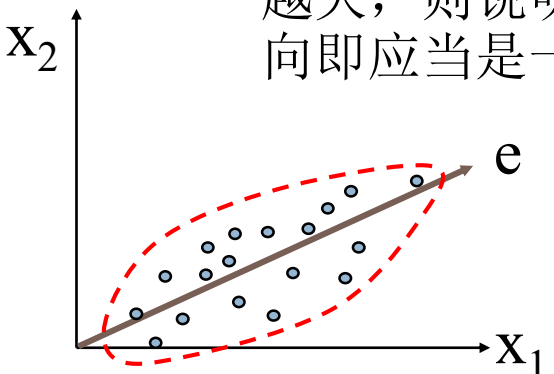
loan_id	119262	贷款记录唯一标识
user_id	0	用户唯一标识
total_loan	12000.0	贷款金额
year_of_loan	5	贷款期限（year）
interest	11.53	贷款利率
…	…	…

NLP中的Glove词表
https://nlp.stanford.edu/projects/glove/



数据规约-维度规约

- 主成分分析(principal component analysis, PCA)
  - 目的：数据降维、数据去噪、数据压缩（去除冗余属性）
  - 思想：将原高维（如维度为N）数据向一个较低维度（如维度为K）的空间投影，同时使得数据之间的区分度变大。这K维空间的每一个维度的基向量（坐标）就是一个主成分
  - **问题**：如何找到这K个主成分
    - **消除原始数据不同属性间的相关性**，要求：K个维度间相互独立
    - **最大化保留K维度上的数据多样性**，要求：最大化每个维度内的样本方差。使用方差信息，若在一个方向上发现数据分布的方差越大，则说明该投影方向越能体现数据中的主要信息。该投影方向即应当是一个主成分

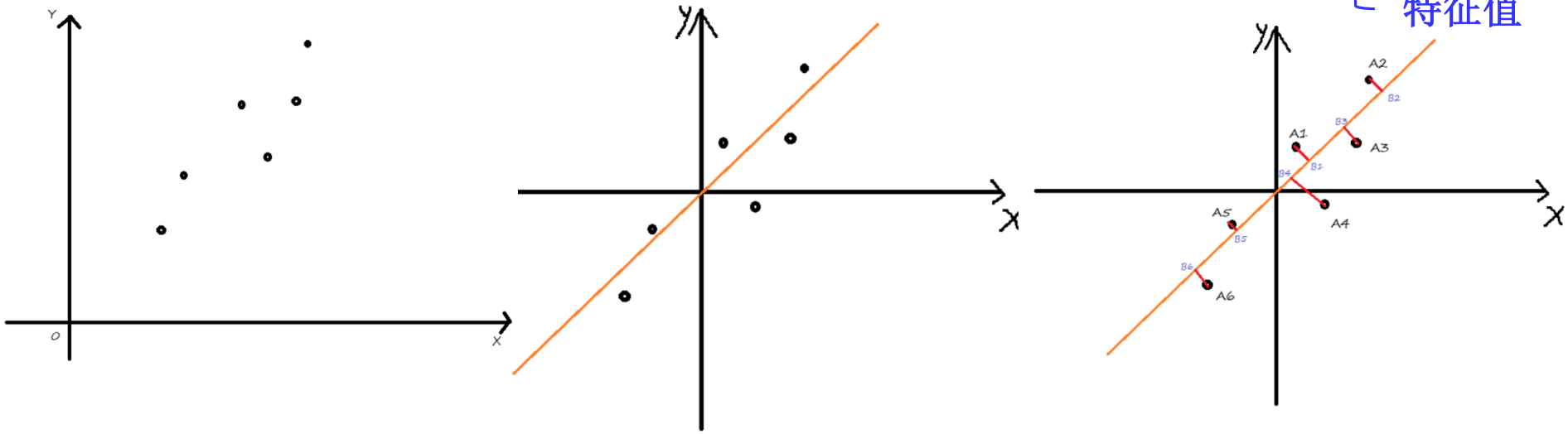


$$D=\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{bmatrix} \Rightarrow D'=\begin{bmatrix} x_1' & x_2' & x_3' & x_4' \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

数据规约-维度规约

- 主成分分析(principal component analysis, PCA)
  - 例：如何把二维数据降维至一维
    - 思想：方差越大，数据间的**差异**越大。让新数据的**方差**尽可能地大，使得新数据尽可能地不丢失原有数据的信息

原始二维数据　　去中心化　　协方差矩阵 → 特征向量、特征值



课后思考：一元线性回归与PCA的关系？

数据规约-维度规约

- 主成分分析(principal component analysis, PCA)
  - 原数据为X，降维后新数据为Y
    - **不同维度相互独立：**要求数据Y的协方差矩阵为对角阵 $B = \frac{1}{m}Y^\top Y \in \mathrm{R}^{k\times k}$

$$B = \frac{1}{m}Y^\top Y = \frac{1}{m}(XP)^\top(XP) = \frac{1}{m}P^\top X^\top XP = P^\top CP \qquad C = \frac{1}{m}X^\top X \in \mathrm{R}^{n\times n}$$

    - **最大化每个维度内的样本方差：**Y的协方差矩阵中的对角元素越大越好
      - 对 C 进行特征值分解，将求解得到的特征值从大到小依次作为协方差矩阵的对角线元素，特征向量组成变换矩阵 (P)
    - 保留最大的 k 个特征值：第K个主成分就是第K大的特征值对应的特征向量
  - 存储空间：
    - 对于原始的N*M（N维M个样本）的数据，原始存储空间是N*M
    - PCA以后为：K*M（M个K维样本）+N*K(K个特征向量)

数据规约-维度规约

- 主成分分析(principal component analysis, PCA)

Algorithm 5 PCA 算法

Input: 原始的样本矩阵 $X \in \mathrm{R}^{m \times n}$

Output: 压缩后的样本矩阵 $Y \in \mathrm{R}^{m \times k}$

1: 对样本矩阵进行去均值化 $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i, \forall i \in \{1, 2, \cdots, m\}$

2: 计算协方差矩阵 $C = \frac{1}{m} X^\top X$

3: 通过特征值分解求解 C 的特征值和特征向量

4: 将特征值从大到小排序，取最大的 k 个特征值对应的特征向量作为列向量构成变换矩阵 $P \in \mathrm{R}^{n \times k}$

5: 将原始数据转换到新的空间中 $Y = XP$

- 不足之处
  - 当原始数据的维度n特别大的时候，计算协方差时的 XᵀX 已经具有相当大的计算量
  - 针对协方差矩阵C的特征值求解过程计算效率不高

数据规约-维度规约

样本矩阵(10个城市样本，8个属性)的转置X^T

省份	GDP X_1	居民消费水平 X_2	固定资产投资 X_3	职工平均工资 X_4	货物周转量 X_5	居民消费价格指数 X_6	商品零售价格指数 X_7	工业总产值 X_8
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64

数据归一化并得到协方差矩阵

三个主成分（8*3）

第一特征向量 a_1	第二特征向量 a_2	第三特征向量 a_3
0.470641	0.107995	0.19241
0.456708	0.258512	0.109819
0.424712	0.287536	0.19241
−0.31944	0.400931	0.397525
0.312729	−0.40431	0.24505
0.250802	0.498801	−0.24777
0.240481	−0.48868	0.332179
−0.26267	0.167392	0.723351

特征值分解

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	.267	.951	.191	.617	−.274	−.264	.874
X_2	.267	1.000	.426	.718	−.151	−.234	−.593	.363
X_3	.951	.426	1.000	.400	431	−.282	−.359	.792
X_4	.191	.718	.400	1.000	−.356	−.134	−.539	.104
X_5	.617	−.151	.431	−.356	1.000	−.255	.022	.659
X_6	−.274	−.234	−.282	−.134	−.255	1.000	.760	−.126
X_7	−.264	−.593	−.359	−.539	.022	.760	1.000	−.192
X_8	.874	.363	.792	−.104	.659	−.126	−.192	1.000

数据规约-维度规约

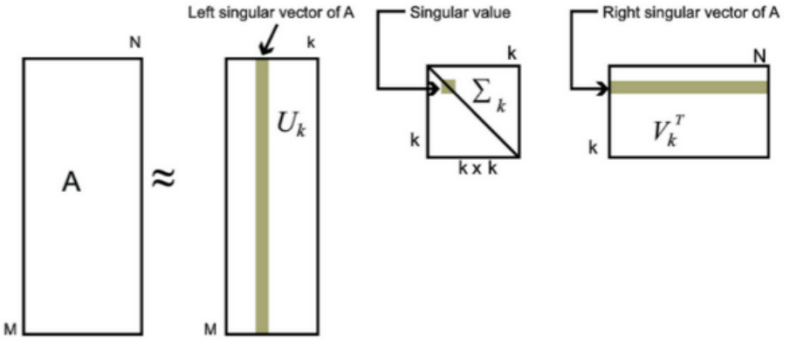
先把大图像分成16*16（256）的小图像块，把小图像块当成一个256维的向量，所有256维向量拼接成新的数据矩阵，对其进行归一化和PCA压缩（取前四个特征值，取前八个，前16个，一直到前256个特征值），压缩完以后需要重构图像就会得到以上的效果

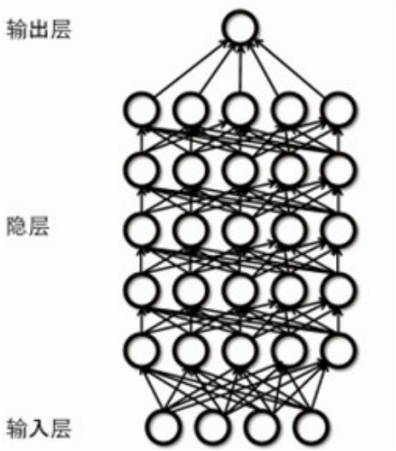
数据规约-维度规约

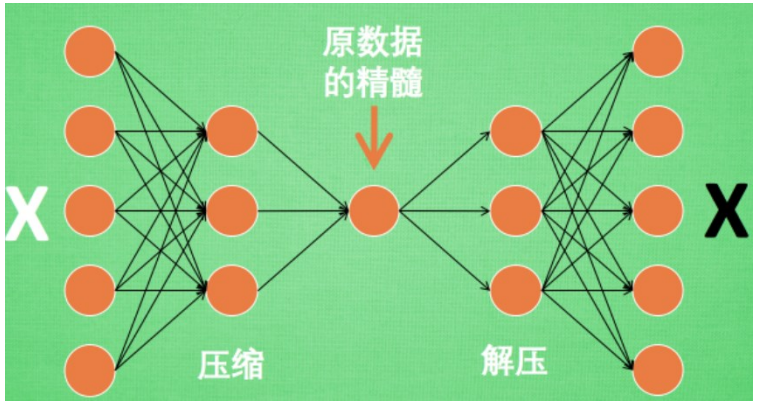
- 维度规约
  - 属性子集选择（特征子集）
    - 做法：删除不相关或冗余的属性来减少维度与数据量
    - 目标：找到最小属性集，使得数据的概率分布尽可能接近使用所有属性得到的原分布
    - 理解：从全部属性中选取一个特征属性子集，使构造出来的模型更好
  - 启发式步骤
    - 建立子集集合
    - 构造评价函数
    - 构建停止准则
    - 验证有效性
  - 例如：决策树

数据规约-维度规约

- 维度规约
  - 奇异值分解(SVD)
  - 矩阵分解(PMF)
  - 深度学习(Deep Learning)
    - DNN，AutoEncoder, etc







数据规约-数值规约

- **数值规约**
  - 通过选择替代的、较小的数据表示形式来减少数据量
- 参数化方法
  - 使用一个参数模型估计数据，最后只要存储参数即可，不用存储数据（除了可能的离群点）
  - 常用方法
    - 线性回归方法；多元回归；对数线性模型；
- 非参数化方法
  - 不使用模型的方法存储数据
  - 常用方法：直方图，聚类，抽样

资料推荐

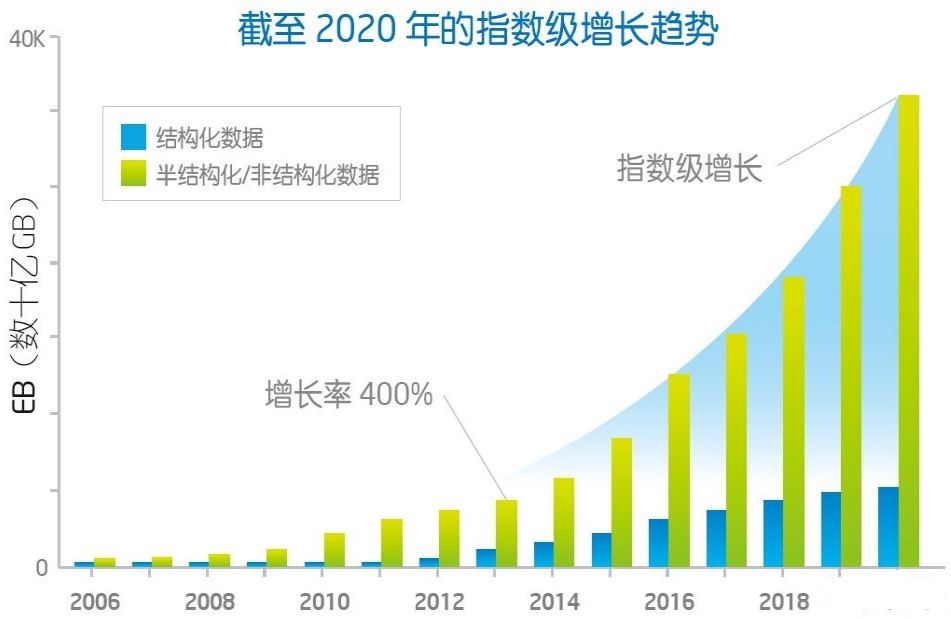
- 数据挖掘导论 第2章：数据，人民邮电出版社
- 数据挖掘原理与算法 第2章， 清华大学出版社
- T.C. Redman Data Quality: The Field Guide. January 2001
- I.T.Jolliffe. Principal Component Analysis. Springer Verlag, 2nd edition, October 2002.
- Feature selection algorithms: A survey and experimental evaluation, ICDM 2003
- Zhenya Huang, Qi Liu, Enhong Chen, Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students, ACM TOIS
- Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang,, Neural Cognitive Diagnosis for Intelligent Education Systems, AAAI'2020

回顾：数据预处理

- 大数据环境下的数据特征
- 为什么需要进行预处理
- 预处理的基本方法
  - 数据清理
  - 数据集成
  - 数据变换
  - 数据规约

数据分析基础

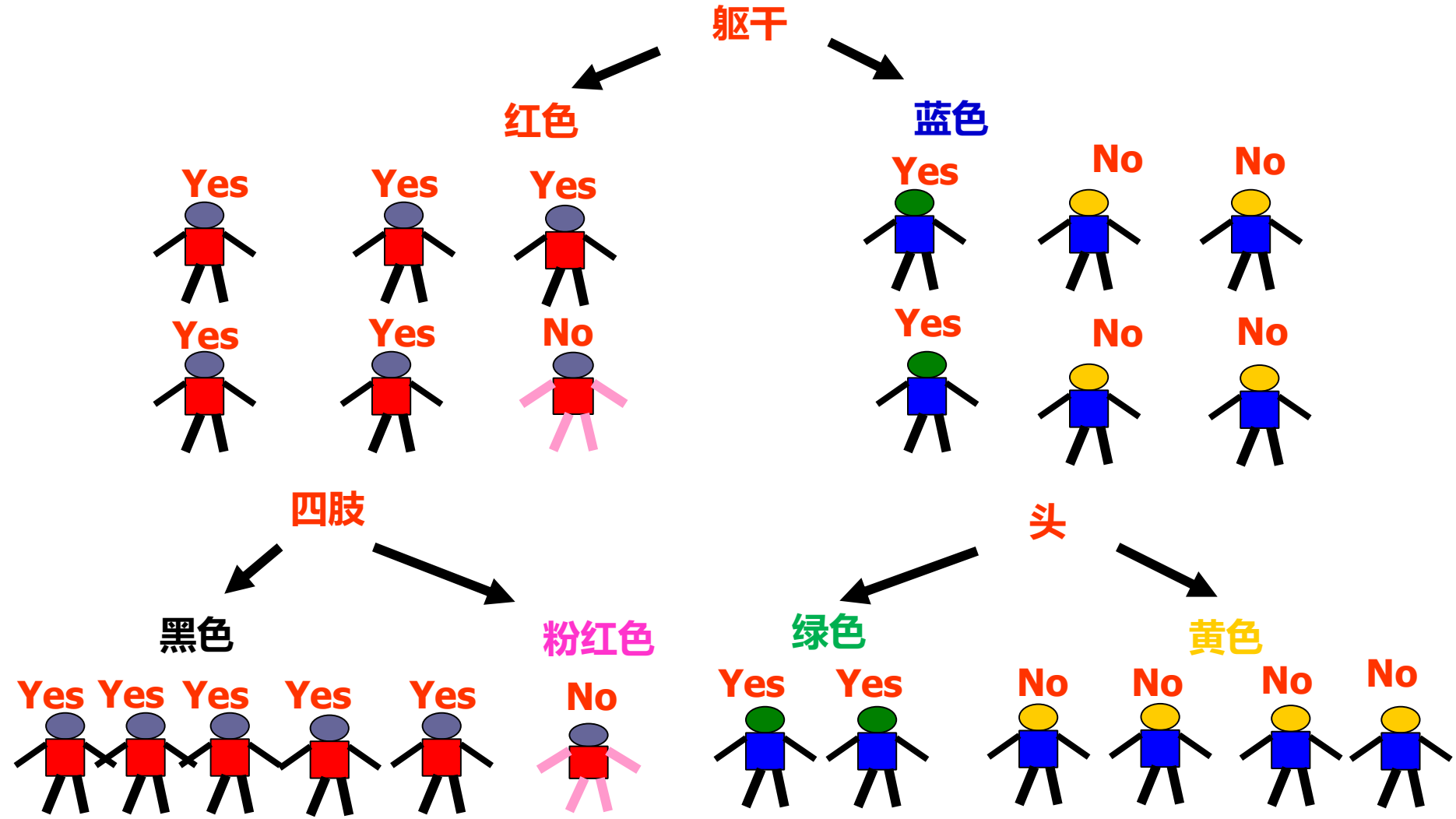
- 数据采集 Data Collection
- 数据存储 Data Storage
- 数据预处理 Data Preprocessing
- **特征工程 Feature Engineering**



特征工程

- 特征工程的定义
- 特征工程的流程
- 特征学习
- 案例学习
- 参考文献

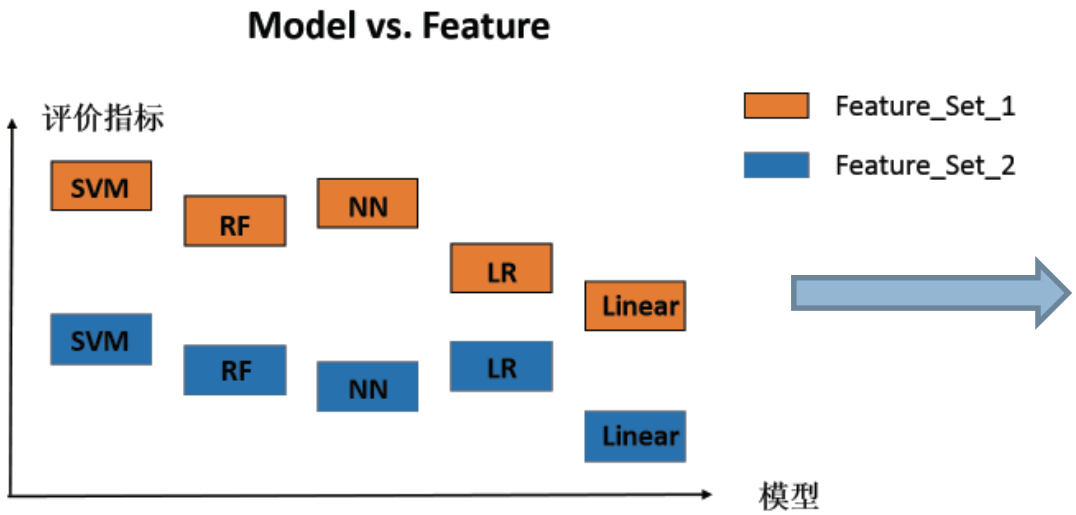
特征工程


躯干
红色
蓝色
Yes
Yes
Yes
Yes
Yes
No
Yes
No
No
Yes
No
No
四肢
头
黑色
粉红色
绿色
黄色
Yes
Yes
Yes
Yes
Yes
No
Yes
Yes
No
No
No
No

特征工程

- 特征工程是什么？ (Feature Engineering)

在数据预处理以后（或者数据预处理过程中），如何从数据中提取有效的特征，使这些特征能够尽可能的表达原始数据中的信息，使得后续建立的数据模型能达到更好的效果，就是特征工程所要做的工作。



特征工程

特征工程的意义

著名数据科学家Andrew Ng 对特征工程这样描述的：“虽然提取数据特征是非常困难、耗时并且需要相关领域的专家知识，但是机器学习应用的基础就是特征工程”

特征越好，灵活性越强

好的特征能使一般的模型也能获得很好的性能，在不复杂的模型上运行速度很快，并且容易理解和维护。

特征越好，构建的模型越简单

好的特征不需要花太多的时间去寻找最优参数，降低了模型的复杂度，使模型趋于简单。

特征越好，模型的性能越出色

好的特征能够使模型表现越出色是毫无疑问的，[illegible]就是提升模型的性能。

特征工程的流程

- 特征工程（**重复迭代**）的流程

1. 对特征进行头脑风暴

深入分析问题，观察数据的基本统计信息，结合问题的相关领域知识和参考其他问题的相关特征工程的方法并应用到自身的问题中来。

2. 特征的设计—基础且重要的步骤

人工设计特征、自动提取特征，或两者结合，得到模型使用的特征。

3. 特征的选择

使用不同的特征重要性评分方法或者特征选择方法，对特征的有效性进行分析，选出有效的特征。

4. 评估模型

利用所选择的特征对测试数据进行预测，评估模型的性能。

5. 上线测试

通过在线测试的效果判断特征是否有效，若不能达到要求，则重复2-5 步骤，直到模型的性能达到要求。

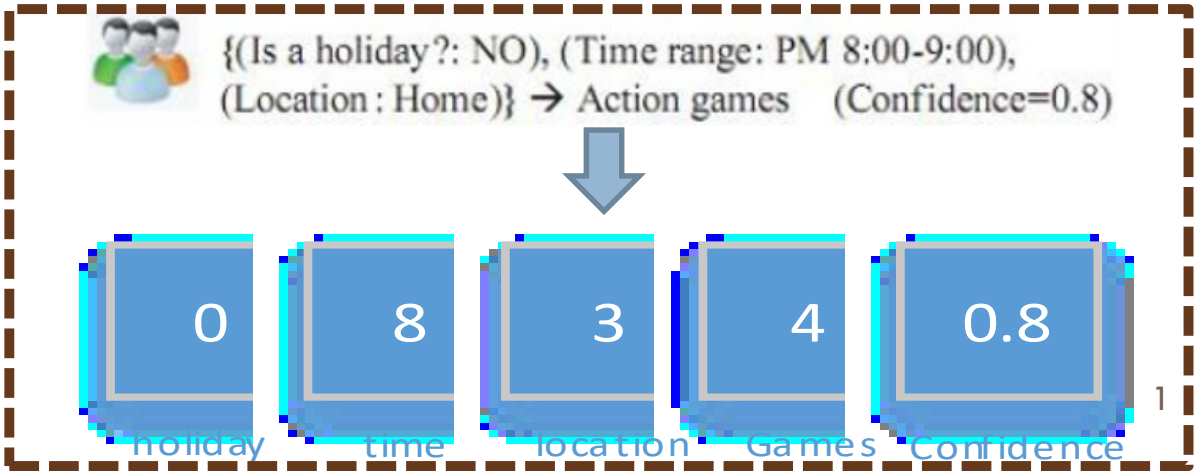
特征的设计

- 从原始数据中如何设计特征？
  - **基本特征的提取**

    基本特征的提取过程就是对原始数据进行预处理，将其转化成可以使用的数值特征。常见的方法有：数据的**归一化**、**离散化**、**缺失值补全**和**数据变换**等。

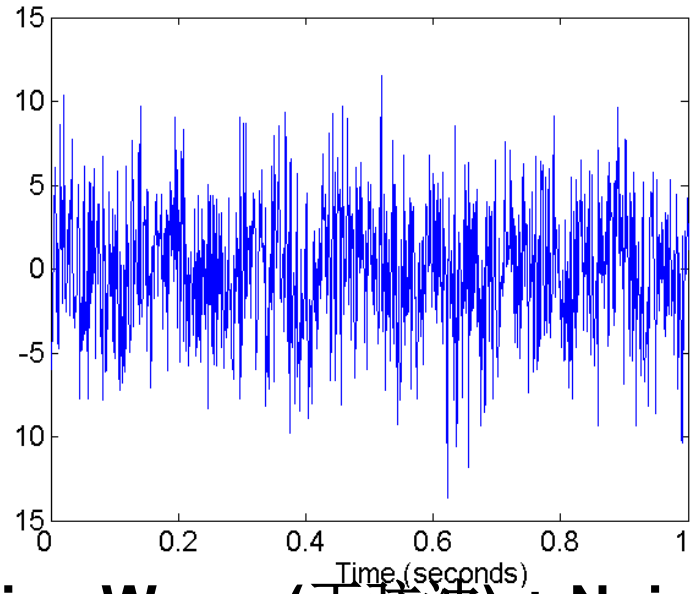
  - **创建新的特征**

    根据对应的领域知识，在基本特征的基础上进行特征之间的比值和交叉变化来构建新的特征。

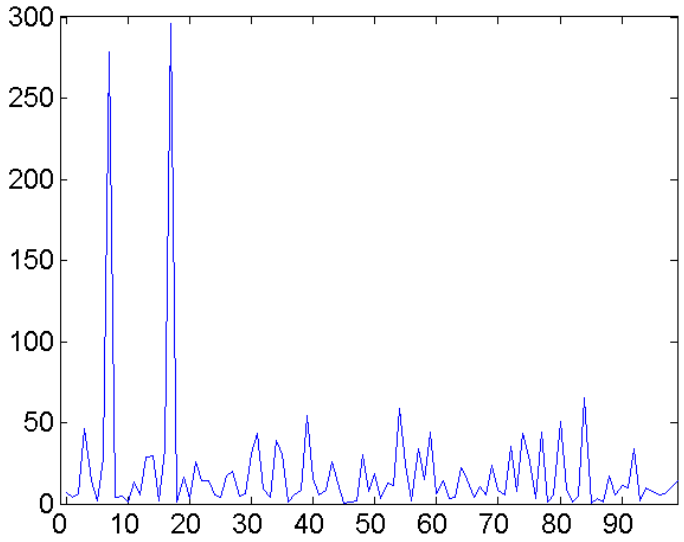


特征的设计

- 从原始数据中如何设计特征？
  - 函数变换特征
    - 左图是根据两个Sin函数（分别是每秒7个和17个周期），以及一些噪声数据得到的序列图；
    - 右图是由**傅立叶变换**得到了频率图，可以看出变换后成功得到了两个概率最大的频率7和17（其中纵坐标是振幅，即概率值）



Two Sine Waves(正弦波) + Noise



Frequency

特征的设计

推荐系统中的“用户”和“商品”

- 从原始数据中如何设计特征？
  - **独热特征表示 One-hot Representation**
  - 将每个属性表示成一个很长的向量（每维代表一个属性值，如词语）
    - **函数：**[0, 0, 1, 0, 0, …, 0, 0, 0, 0]
    - **图像：**[0, 0, 0, 0, 0, …, 0, 0, 0, 1]
  - **优点：**直观，简洁
  - **缺陷：**
    - **“维度灾难” 问题：**尤其是我们所构建的语料库包含的词语数据非常多的时候，独热表征在空间和时间上的开销都是十分巨大的
    - **“语义鸿沟” 现象：**任意两个词之间都是完全孤立的，是无法刻画句子中词语的语序信息的（之前提到的词袋模型也是如此）。例如，我们是无法通过独热表征来判断“函数”与“偶函数”之间的联系的（但实际上这两个词语是非常相关的）。

特征的设计

- 从原始数据中如何设计特征？
  - 独热特征表示 One-hot Representation
    - “维度灾难” 问题
    - “语义鸿沟” 现象

Download pre-trained word vectors

- Pre-trained word vectors. This data is made available under the Pu
  http://www.opendatacommons.org/licenses/pddl/1.0/.
  - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncas
  - Common Craw (42B tokens, 1.9M vocab, uncased, 300d vec
  - Common Craw (840B tokens, 2.2M vocab, cased, 300d vect
  - Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50

“dog”	“canine”
3	399,999
$\begin{pmatrix}0\\0\\1\\0\\\vdots\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\0\\0\\\vdots\\1\\0\end{pmatrix}$

特征的设计

- 从原始数据中如何设计特征？
  - **数据的统计特征，**如：文档中的词频统计

John likes to watch movies. Mary likes too.

John also likes to watch football games.

  - 字典

{"John": 1, "likes": 2, "to": 3, "watch": 4, "movies": 5, "also": 6, "football": 7, "games": 8, "Mary": 9, "too": 10}

  - 文档词频特征

[1, 2, 1, 1, 1, 0, 0, 0, 1, 1]

[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

特征的设计

- 从原始数据中如何设计特征？
  - **TF-IDF（词频-逆文档率）**
    - 算法简单高效，工业界用于最开始的**数据预处理**
  - 主要思想：找到能代表该文档中的**“关键词”**
  - 词频 （TF, Term Frequency）
    - TF = 某个词(特征值)在句子(数据)中出现的频率 $TF_{w,D_i} = \frac{count(w)}{|D_i|}$
  - 逆文档率 （IDF, Inverse Document Frequency）
    - IDF = log(语料库(数据库)的句子(数据)总数 / 包含该词(特征值)的句子(数据)总数) $IDF_w = \log \frac{N}{1+\sum_{i=1}^{N} I(w, D_i)}$
  - 每个特征值（词）的重要性
    - w_{ij}= tf*idf = TF_{ij}*log(N/DF_i)　　可能会有微小变型

特征的设计

$$TF_{w,D_i} = \frac{count(w)}{|D_i|} \qquad IDF_w = \log \frac{N}{1 + \sum_{i=1}^{N} I(w, D_i)}$$

- 从原始数据中如何设计特征？—计算 TF-IDF
  - d_1 (A, B, C, C, S, D, A, B, T, S, S, S, T, W, W)
  - d_2 (C, S, S, T, W, W, A, B, S, B)
  - d_3(不含ABCDSTW)

文档中词总数=25

TF

	d_1	d_2
A	0.08	0.04
B	0.08	0.08
C	0.08	0.04
D	0.04	0.00
S	0.16	0.12
T	0.08	0.04
W	0.08	0.08

IDF

	IDF
A	0.4
B	0.4
C	0.4
D	1.1
S	0.4
T	0.4
W	0.4

TF-IDF

	d_1	d_2
A	0.032	0.016
B	0.032	0.032
C	0.032	0.016
D	0.044	0.000
S	0.064	0.048
T	0.032	0.016
W	0.032	0.032

特征的设计

- 从原始数据中如何设计特征？
  - **TF-IDF（词频-逆文档率）** w_{ij}= tf*idf = TF_{ij}*log(N/DF_i)
- 优点
  - 简单快速的词（特征）重要性表示方法，结果比较符合实际情况
  - 应用广泛：不仅限于文本数据
- 缺点
  - 单纯以“词频”衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多
  - 无法体现词的**位置信息、顺序信息，**出现位置靠前的词与出现位置靠后的词，都被视为重要性相同
  - 无法发现词（特征）的隐含联系，如同义词等

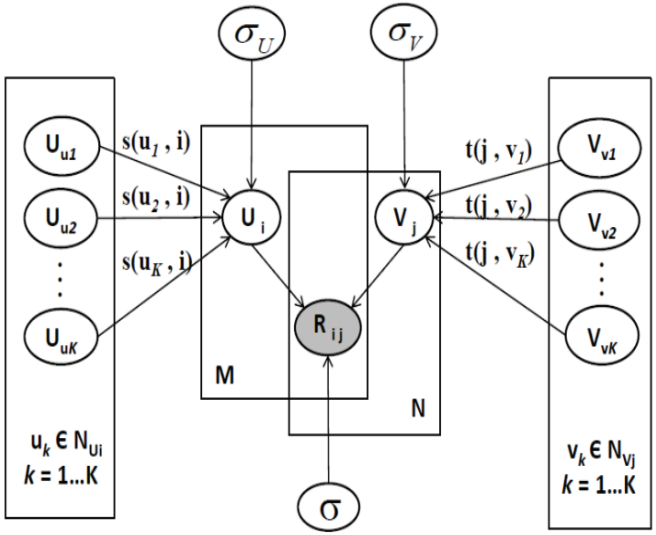
特征的设计

- 从原始数据中如何设计特征？
  - **TF-IDF（词频-逆文档率）—应用**
    - 搜索引擎；关键词提取；文本相似性；文本摘要
  - 推荐系统
    - 可以计算“用户-标签-商品”的特征
    - 用户-标签的TF-IDF

$$P_{il} = tf(i,l) \times \ln(\frac{M}{df(l)})$$

- 用户：i。标签：l。用户总数：M。

$$s(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_F * \|\vec{j}\|_F}$$



Le Wu, Enhong Chen, Qi Liu, Leveraging Tagging for Neighborhood-aware Probabilistic Matrix Factorization. CIKM'2012

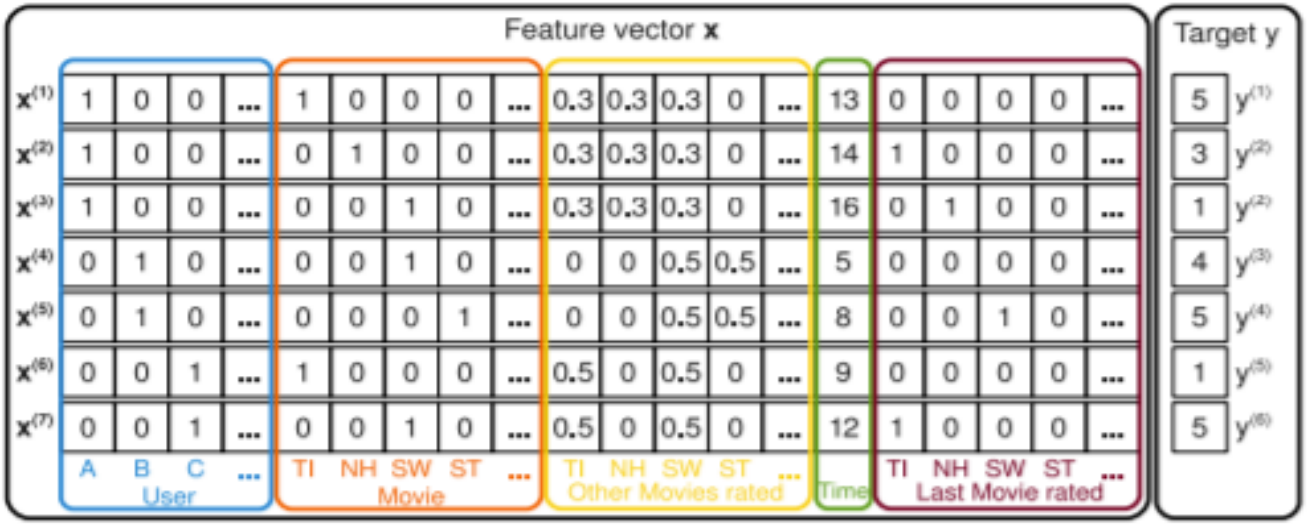
特征的设计

*ratings.csv - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
userId,movieId,rating,timestamp
1,1,4.0,964982703
1,3,4.0,964981247
1,6,4.0,964982224
1,47,5.0,964983815
1,50,5.0,964982931
1,70,3.0,964982400
1,101,5.0,964980868
```

- 从原始数据中如何设计特征？
  - **特征组合：构造高阶特征**
  - 上述所有构造的特征均可以：两两、三三、… 进行组合
    - Factorization Machine (2012)



$$S = \{(\text{A}, \text{TI}, \text{2010-1}, 5), (\text{A}, \text{NH}, \text{2010-2}, 3), (\text{A}, \text{SW}, \text{2010-4}, 1), (\text{B}, \text{SW}, \text{2009-5}, 4), (\text{B}, \text{ST}, \text{2009-8}, 5), (\text{C}, \text{TI}, \text{2009-9}, 1), (\text{C}, \text{SW}, \text{2009-12}, 5)\}$$

$$\tilde{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} x_i x_j$$

特征的设计

- 举例：第二届“中国高校计算机大赛-大数据挑战赛”
- 赛题描述/数据：http://bdc.saikr.com/vse/bdc/2017
- 简单的说，该赛题的求解目标是利用数据分析将人工的鼠标轨迹和代码生成的鼠标轨迹区分开来。这里的鼠标轨迹是指一种完成一种验证手段——拖动滑块到指定区域时鼠标的轨迹。

  - 原始数据格式：一系列连续点的坐标及其对应时间，目标点的坐标

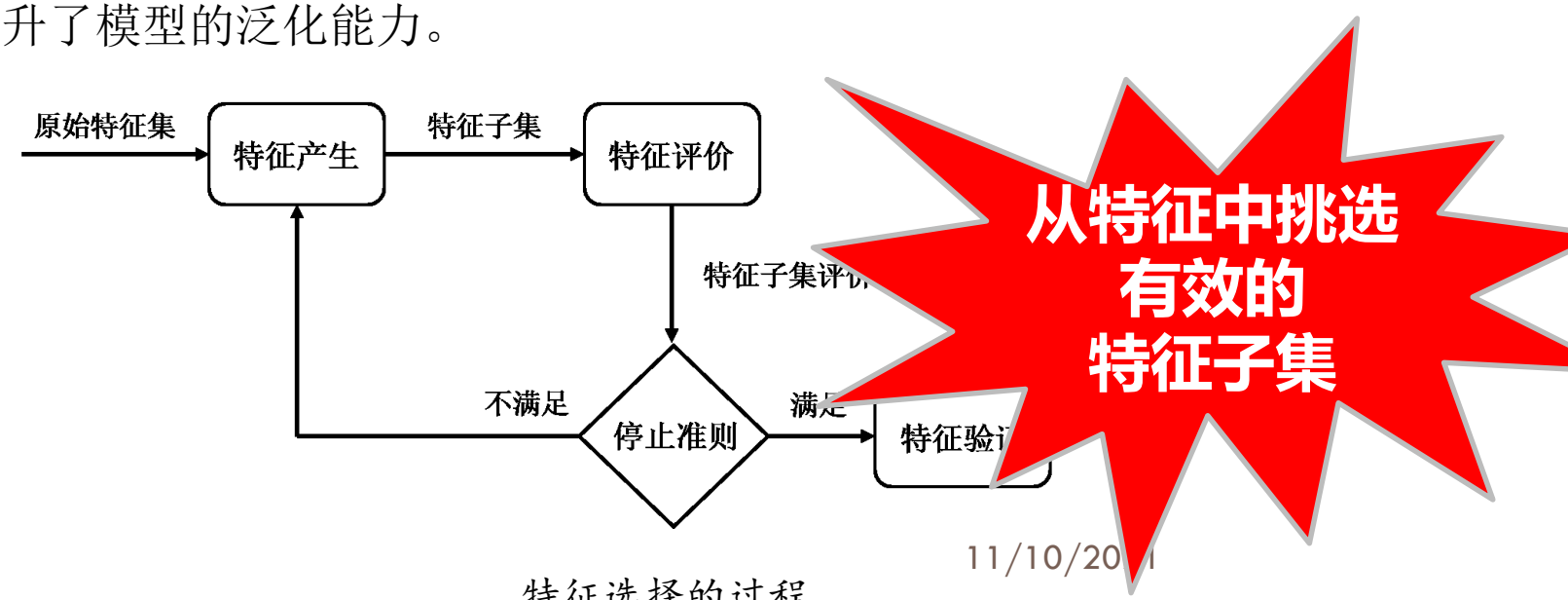
例如：(2,3,4),(2,5,6)(4,3,7) (4,3)，该轨迹中含有三个点的坐标，以(x,y, time)的时间表示，终点坐标为(4,3)

特征的设计

- 从原始数据中如何设计特征？
  - **基本特征的提取**
    - 轨迹运动数据的统计值：运动速度/加速度/角加速度/角速度的均值/极值/最值/中位数 等
    - 轨迹的描述：运动在x轴方向是否为单向，曲线平滑程度， 等
  - **创建新的特征**
    - 基本特征的简单二元运算， 加/减/乘/除/平方和/和平方/倒数和
    - 运动数据在某一维上的偏导
    - 领域专家知识

特征的选择

- 如何挑选有效的特征（**Subset Selection问题**）？
  - 在实际应用中，特征的数量往往比较多，其中可能会存在不相关的特征。
  - 特征数量越多，分析特征、训练模型所需要的时间就越长，同时容易引起**“维度灾难”**，使得模型更加复杂。
  - **特征选择**通过剔除不相关的特征或冗余的特征来**减少特征数量**，从而简化了模型并且提升了模型的泛化能力。



特征选择的过程

特征子集产生过程

- 如何生成特征子集？

特征选择的本质上是一个组合优化的问题，求解组合优化问题最直接的方法就是搜索.根据不用的搜索策略，可以将搜索的算法分为**完全搜索**(Complete), **启发式搜索**(Heuristic) 和**随机搜索**(Random) 三大类。

1. 采用**全局最优搜索策略**的过程产生方法

全局最优搜索策略可以分为穷举搜索与非穷举搜索两类。穷举搜索策略有遍历所有特征和以广度优先搜索的策略，这两种搜索策略都枚举了所有的特征组合，复杂度为2^n。

2. 采用**启发式搜索策略**的过程产生方法

启发式搜索的基本思想是增加关于要解决问题的解某些特征，以便指导搜索向最有希望的方向发展。启发式搜索是搜索是在搜索的最优性和计算量之间做一个折中的搜索策略。

3. 采用**随机算法搜索策略**的产生方法

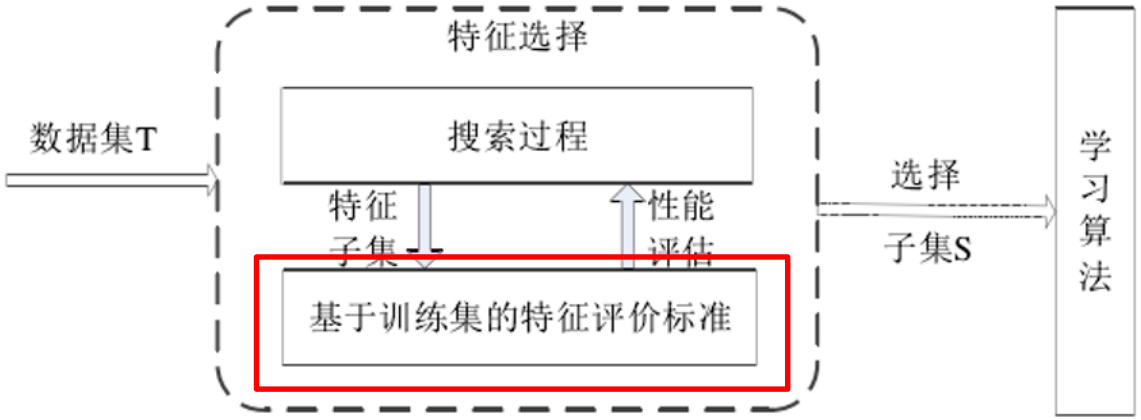
特征选择本质上是一个组合优化问题，求解这类问题可采用非全局最优目标搜索方法，其实现是依靠带一定智能的随机搜索策略。(如模拟退火，遗传算法等)

特征子集评价

- 如何评价特征子集？

不同的特征选择算法不仅对特征子集评价标准不同，有的还需要结合后续的学习算法模型。因此根据特征选择中子集评价标准和后续算法的结合方式主要分为**过滤式(Filter)、封装式(Wrapper)** 和**嵌入式(Embedded)**三种

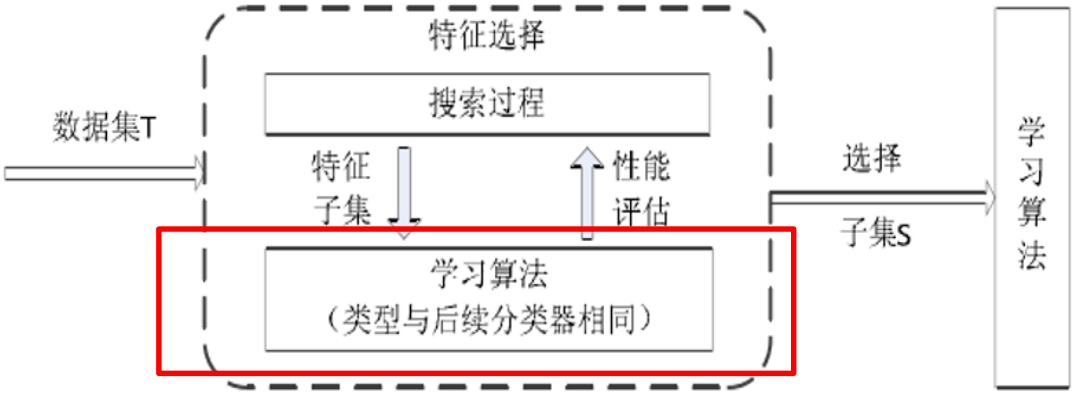
- **1. 过滤式(Filter)评价策略方法**
  - 独立于后续的学习算法模型来分析数据集的固有的属性
  - 采用一些基于信息统计的启发式准则来评价特征子集
  - 启发式的评价函数：距离度量、信息度量、依赖性度量、一致性度量



特征子集评价

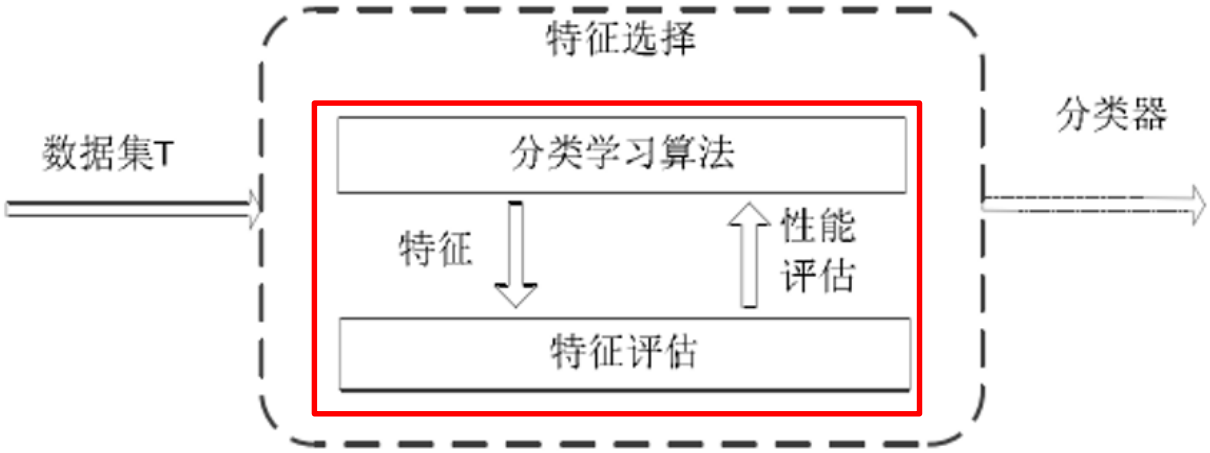
- 如何评价特征子集？
  - **2. 封装式(Wrapper)评价策略方法**
    - 将特征选择作为学习算法一个组成部分，需要结合后续的学习算法，并直接将学习算法的分类性能作为特征重要性的评价标准
    - 直接使用**分类器的性能作为评价的标准**，选出来的特征子集对分类一定有最好的性能

相对于Filter 选择方法，Wrapper 方法所选择的特征子集的规模要小得多，有利于关键特征的辨识，模型的分类性能更好。但Wrapper 方法泛化能力较差，当改变学习算法时，需要针对该学习算法重新进行特征选择，算法的计算**复杂度高**.



特征子集评价

- 如何评价特征子集？
  - **3. 嵌入式(Embedded)评价策略方法**
    - 特征选择算法嵌入到学习和分类算法中，**特征选择是算法模型中的一部分**
    - 算法模型训练和特征选择同时进行，互相结合。即，算法具有自动进行特征选择的功能



特征子集评价

- 如何评价特征子集？

3. 嵌入式(Embedded)评价策略方法

1). 带惩罚项的特征选择方法

- 在模型损失函数上加上一个惩罚项，模型训练时通过惩罚项来**对特征的系数进行惩罚处理**，而在特征选择方法中常使用的是L1 正则化(regularization)项

线性回归的损失函数：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$

岭回归的损失函数：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\lambda\sum_{j=1}^{n}\theta_j^2$

Lasso回归的损失函数：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\lambda\sum_{j=1}^{n}|\theta_j|$

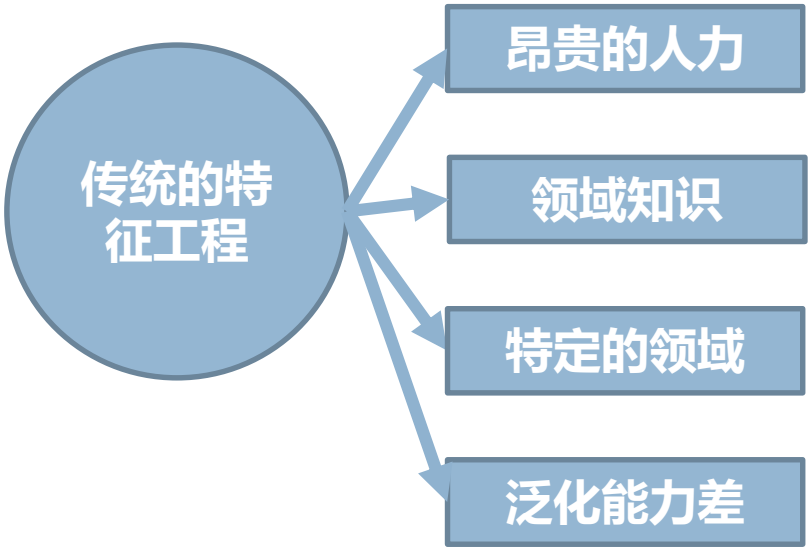
- **2范数，让系数的取值变得平均。对于关联特征，能够获得更相近的对应系数**
- **1范数，系数w的l1范数作为惩罚项加到损失函数上，由于非零，迫使那些弱的特征所对应的系数变成0，达到特征选择**

2). 基于树模型的特征选择方法

- 每一步都必须选择一个特征，将样本集划分为纯度更高的子集，而每次选择出的都是使划分效果最佳的特征。**决策树、迭代决策树(GBDT)、随机森林**
- **第四章：数据挖掘算法**

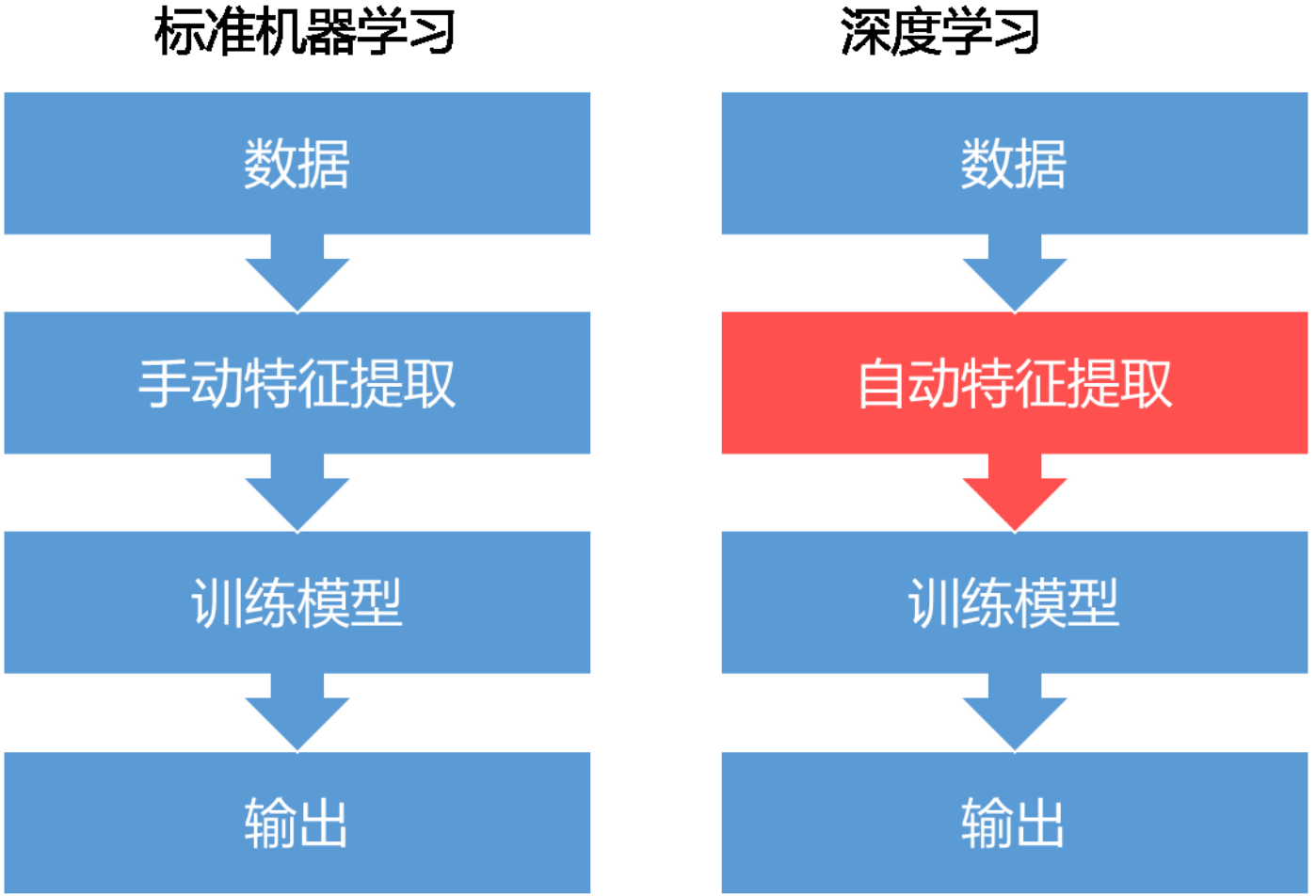
传统特征工程的缺点

- 传统特征工程的缺点



传统特征工程的缺点

- 传统特征工程的缺点

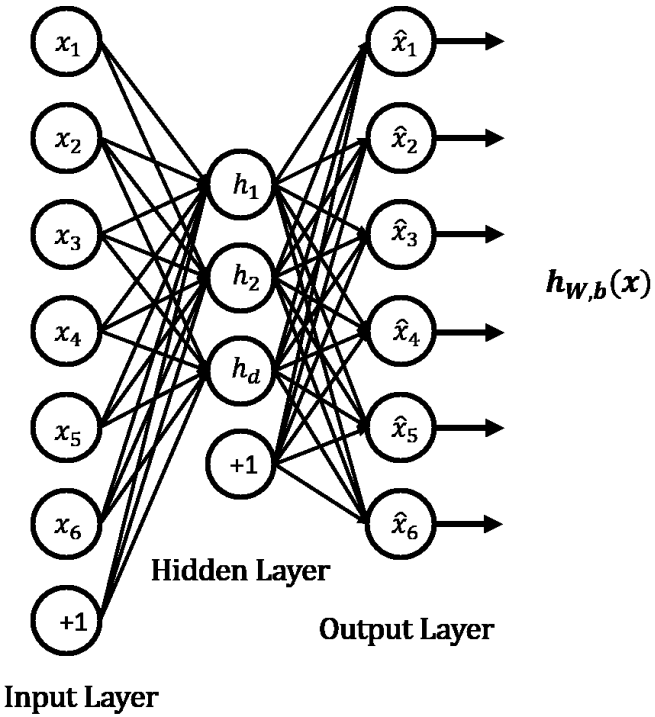


特征学习

- 特征学习

如何从数据中能够自主的学习特征，在这里我们主要介绍在深度学习中常用的三种网络结构。

- 自编码结构(Auto-Encoder)



$h_{W,b}(x)$

将数据的特征X作为Input Layer输入
同样将原始数据特征X'作为Output Layer的输出来重构出原数据。

Encoder: $H = f(A * X + b)$

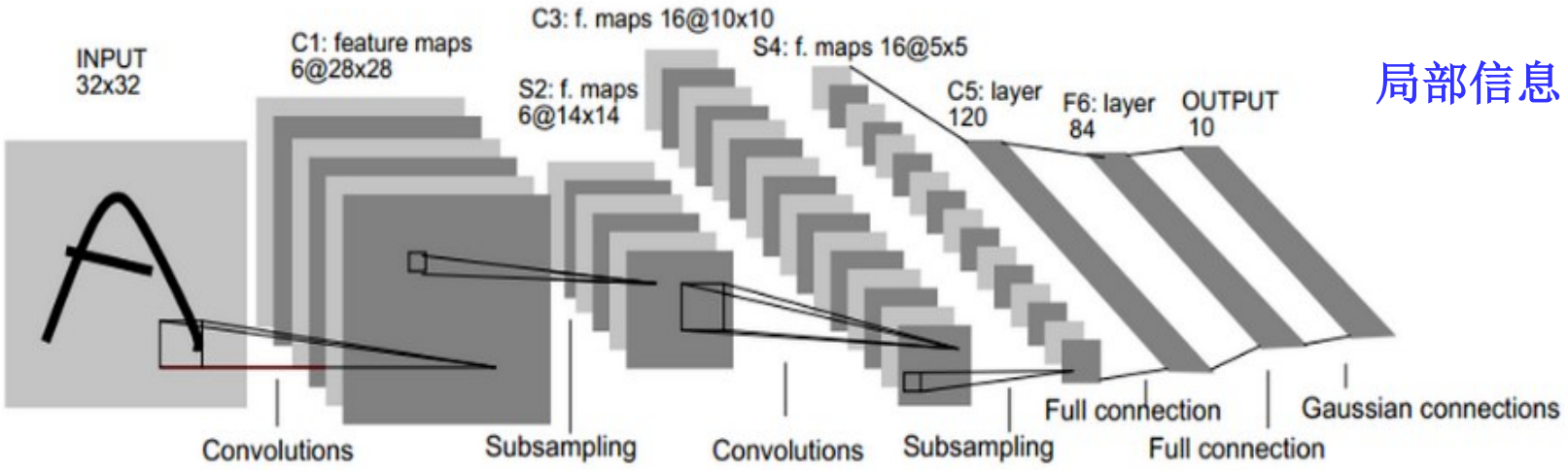
Decoder: $X' = f(A' * H + b')$

将中间的隐含层H的输出作为学习到的数据特征。

特征学习

- 卷积神经网络(CNN): 常用于图像特征提取

LeNet



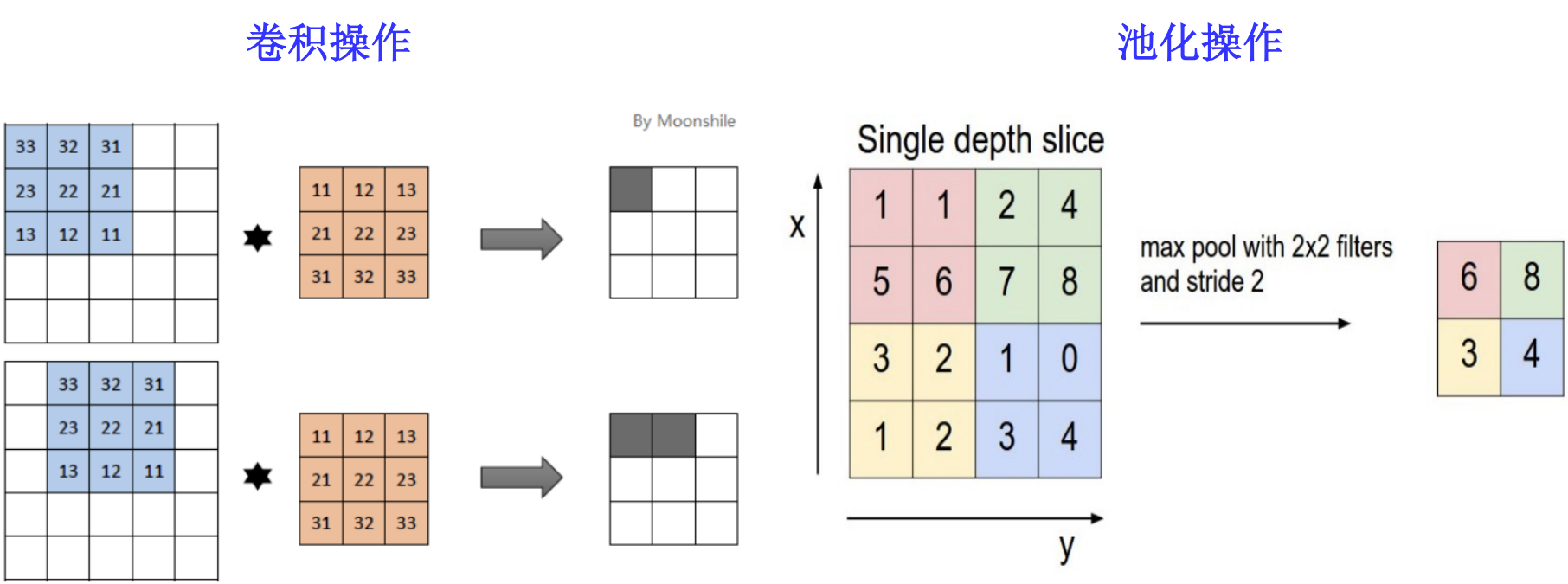
局部信息

卷积层：通过局部平移，利用不同的卷积核来提取图像中不同的特征
池化层：计算某个区域的特征，提高模型的泛化能力
全连接层：通过多层的神经网络，抽取更高阶的特征。

最终**全连接层的输出**即为该图像的特征向量表示。

特征学习

- 卷积神经网络(CNN): 常用于图像特征提取

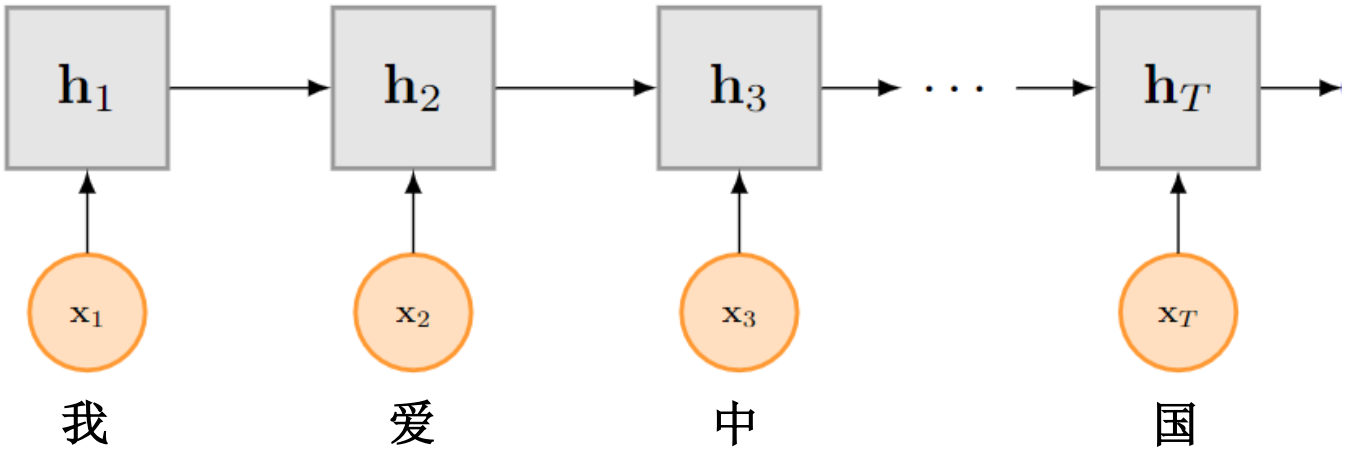


思考：卷积神经网络能否用于文本处理？如何做？

Zhenya Huang, Qi Liu, Enhong Chen, Question Difficulty Prediction for READING Problems in Standard Tests，AAAI'2017

特征学习

- 循环神经网络(RNN): 常用于序列数据的特征提取



位置信息
顺序信息

将序列中的每个数据依次作为RNN的输入，如上图中的文本数据‘我’、‘爱’、‘中’、‘国’，并将最后一层网络的输出h_T作为最终序列数据的特征向量

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, Guoping Hu, EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction, IEEE TKDE), 33(1): 100-115,2021

特征学习

- 利用标准数据集进行特征学习（特征预训练）
  - 作用：模型效果验证 & 应用问题中的模型预训练
  - 图像数据预训练：ImageNet
    - http://www.image-net.org/
    - 1400万张图片数据，2万类别，已标注
    - 常用模型：ResNet，AlexNet，VGG等
    - 常见应用：图像分类、目标检测、目标定位
  - 文本数据预训练：Twitter，Wiki
    - https://nlp.stanford.edu/projects/glove/
    - 2 Billon tweets, 27 Billion 词数，1.2M 词表
    - 常用模型：CBOW，Skip-gram，Glove等Word2
    - 常见应用：文本分类，文本推理，翻译等

特征学习

Efficient estimation of word representations in vector space

T Mikolov, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing …

☆ 99 被引用次数：24421 相关文章 所有 43 个版本 »

- Word2Vec—自然语言处理的预训练
  - 哪句话更像自然语句

S1: 语言模型的本质是对一段自然语言的文本进行预测概率的大小
S2: 语言模型的本质是对自然一段语言的文本进行预测概率的大小
S3: 语言模型的本质是对自然语言一段的文本进行预测概率的大小

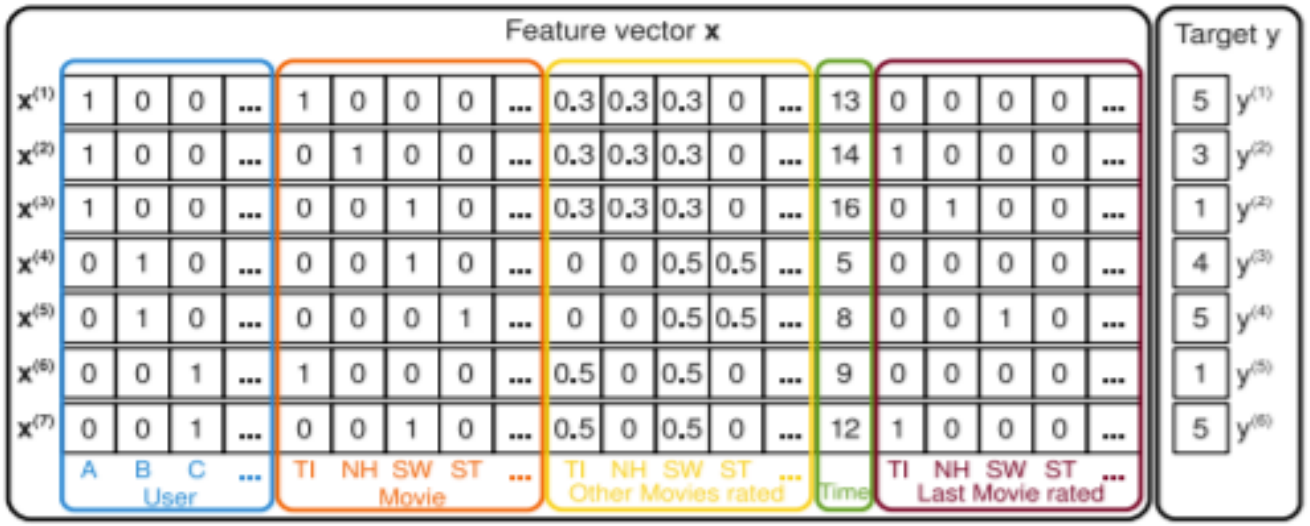
  - 计算词构成句子的概率—最大化

$$P(w_1, w_2, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1, \dots, w_{n-1})$$

$$L = \sum_{w \in C} \log P(w|context(w))$$

特征工程: 可解释性

- 特征的含义
  - 人工设计的特征
  - 特征的构造基于先验知识，天然具有可解释性

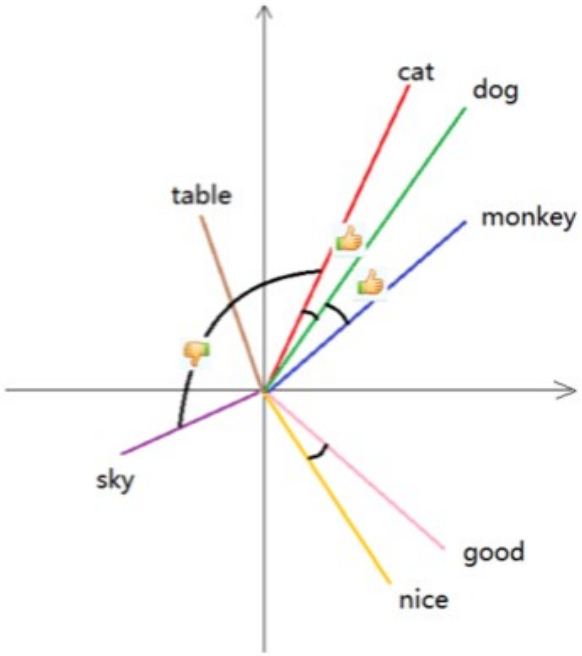


$$S = \{(\text{A}, \text{TI}, \text{2010-1}, 5), (\text{A}, \text{NH}, \text{2010-2}, 3), (\text{A}, \text{SW}, \text{2010-4}, 1),$$
$$(\text{B}, \text{SW}, \text{2009-5}, 4), (\text{B}, \text{ST}, \text{2009-8}, 5),$$
$$(\text{C}, \text{TI}, \text{2009-9}, 1), (\text{C}, \text{SW}, \text{2009-12}, 5)\}$$

特征工程: 可解释性

- 特征的含义
  - Word2Vec—自然语言处理的预训练
  - 学习语言的语义特性：解决“语义鸿沟”

词的相似性



词的类比性

King – Queen ~= Man – Woman

China – Beijing ~= UK – London ~= Capital

Efficient estimation of word representations in vector space

T Mikolov, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org

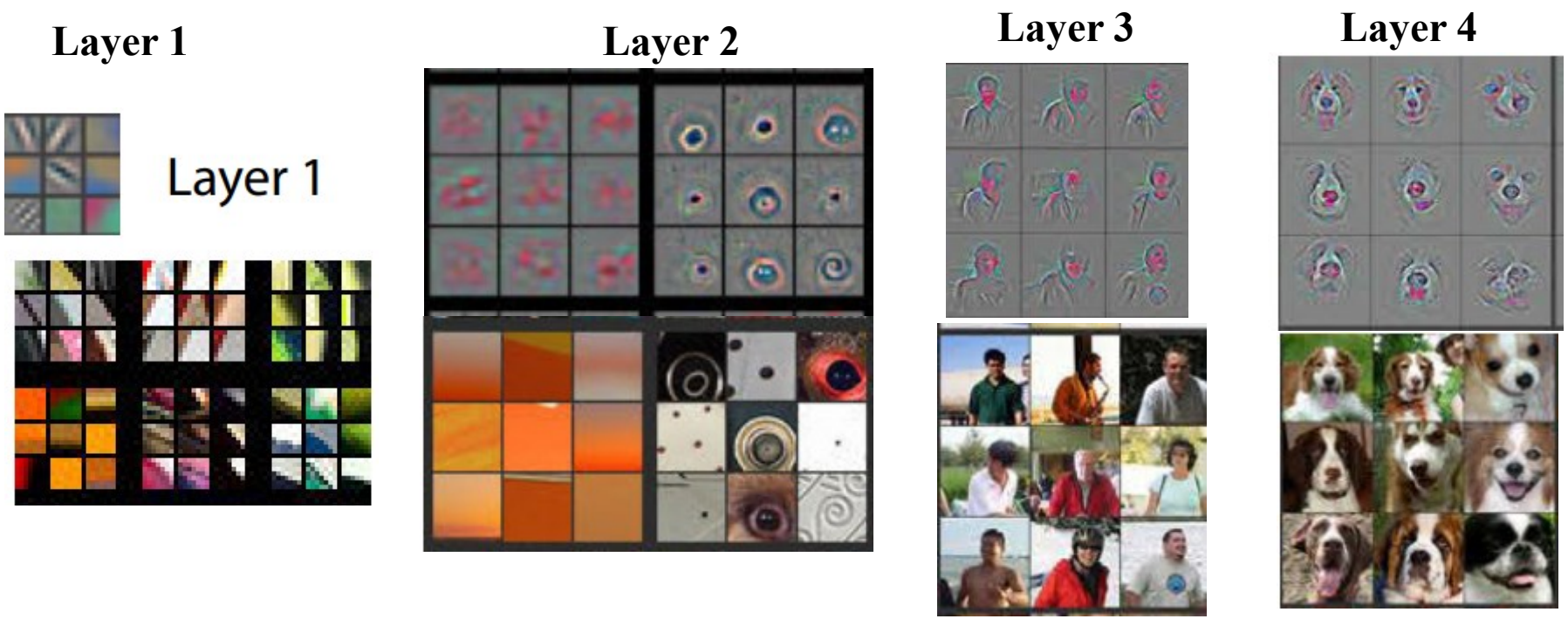
We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing …

☆ 99 被引用次数: 24421 相关文章 所有 43 个版本 »

特征工程: 可解释性

Visualizing and understanding convolutional networks
MD Zeiler, R Fergus - European conference on computer vision, 2014 -
Abstract Large Convolutional Network models have recently demonstra
classification performance on the ImageNet benchmark Krizhevsky et a
is no clear understanding of why they perform so well, or how they migh
paper we explore both issues. We introduce a novel visualization techn
insight into the function of intermediate feature layers and the operation
Used in a diagnostic role, these visualizations allow us to find model arc
☆ 99 被引用次数: 13612 相关文章 所有 18 个版本

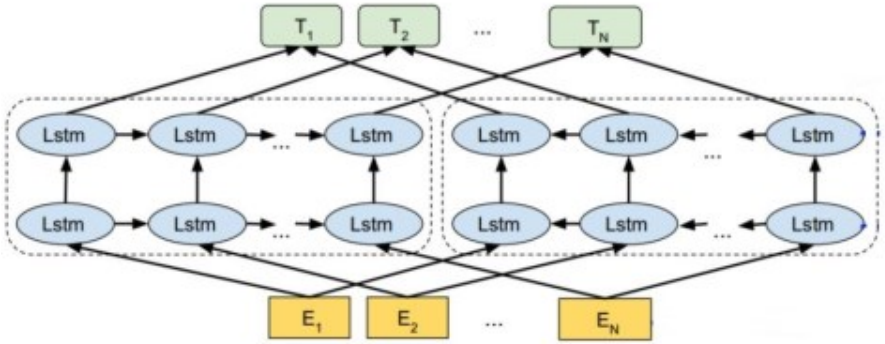
- 特征的含义—可解释性
  - CNN Feature map——特征图可视化
  - 学习图像的特点：图形图像的“颜色，边角，轮廓，图形”等

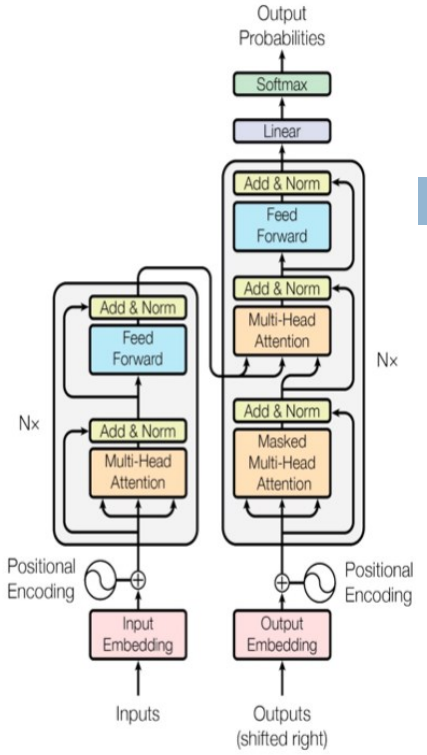


Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

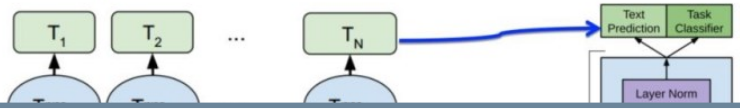
特征学习

- 自然语言处理的预训练模型
  - Elmo，GPT，Transformer, BERT

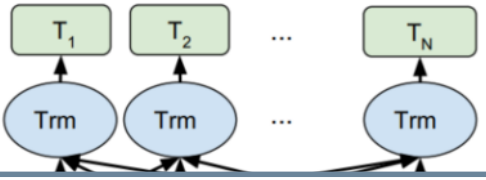




OpenAI GPT



BERT (Ours)



特征学习过程已经不局限于人工的思考、构造、统计等方法。它已经成为一个重要的研究方向，专门的特征学习模型已经在CV、NLP等领域取得重要的突破。

参考文献

- 书籍
  - 数据挖掘导论
  - 机器学习
- 论文
  - 《An Introduction to Variable and Feature Selection》
  - 《特征选择常用算法综述》
- 实战经验
  - Sklearn官方文档
  - Kaggle和天池比赛论坛

第二章数据分析基础小结

- 数据采集 Data Collection
  - 信息检索
  - 网络爬虫
- 数据存储 Data Storage
- 数据预处理 Data Preprocessing
  - 数据清洗
  - 数据集成
  - 数据变换
  - 数据规约
- 特征工程 Feature Engineering
  - 特征设计
  - 特征理解