

本课件仅用于教学使用。未经许可,任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等),也不得上传至可公开访问的网络环境

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

Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页:

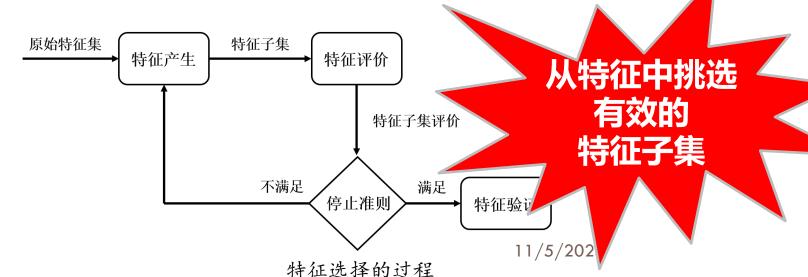
http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html

11/5/2025



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,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





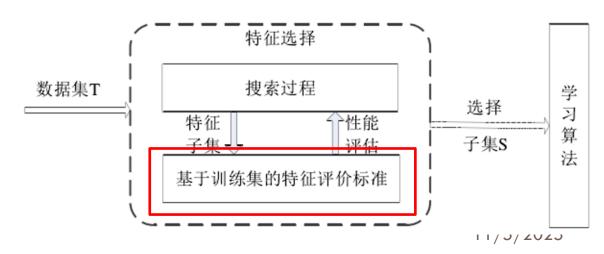
特征子集评价

26

□ 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

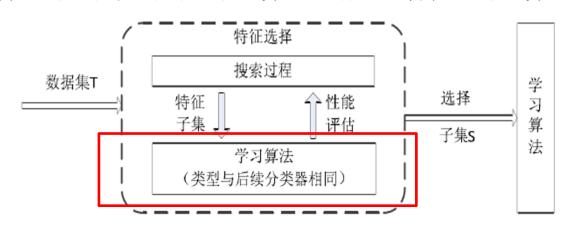
27

□ 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<mark>结合后续的学习算法</mark>,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

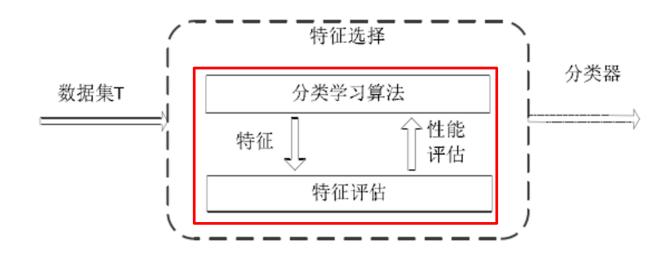
相对于Filter 选择方法,Wrapper 方法所选择的特征子集的规模要小得多,有利于关键特征的辨识,模型的分类性能更好。但Wrapper 方法泛化能力较差,当改变学习算法时,需要针对该学习算法重新进行特征选择,算法的计算**复杂度高**.





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

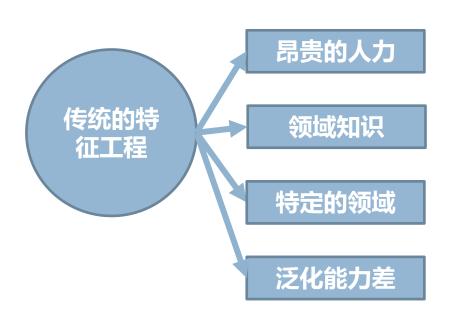




传统特征工程的缺点

29

□传统特征工程的缺点



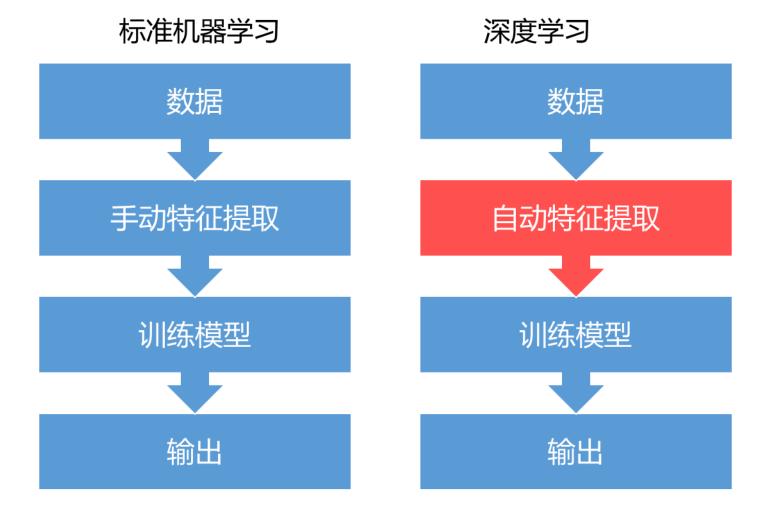




传统特征工程的缺点

30

□传统特征工程的缺点

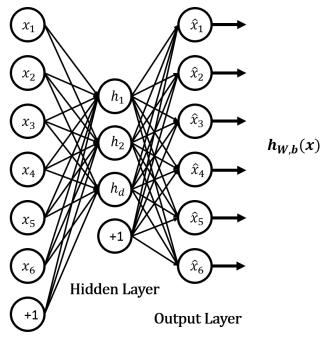




□特征学习

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)



Input Layer

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')

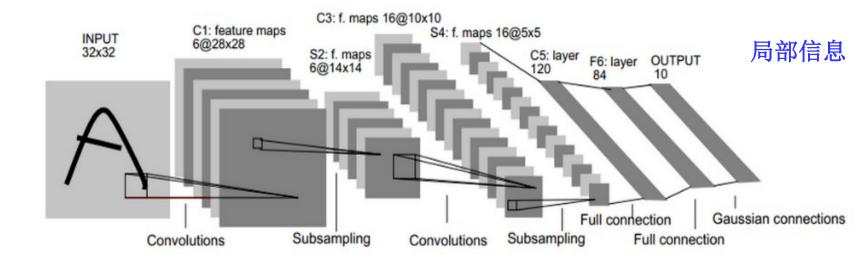
将中间的隐含层H的输出作为学习到的数据特征。

11/5/2025



32

□ 卷积神经网络(CNN): 常用于图像特征提取



LeNet

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

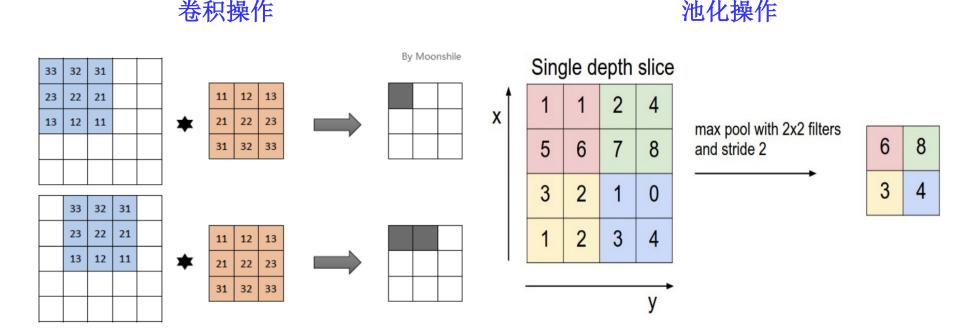
池化层: 计算某个区域的特征,提高模型的泛化能力**全连接层**: 通过多层的神经网络,抽取更高阶的特征。

最终全连接层的输出即为该图像的特征向量表示。



33

□ 卷积神经网络(CNN): 常用于图像特征提取

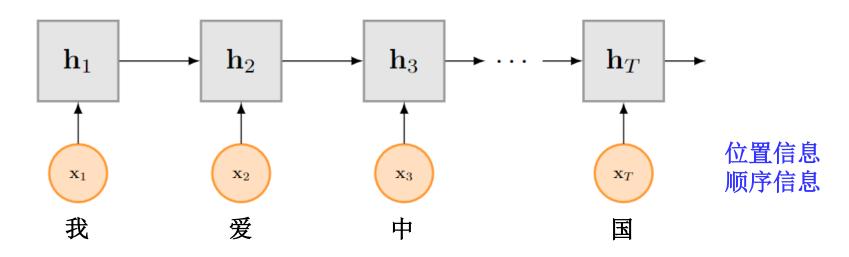


思考: 卷积神经网络能否用于文本处理? 如何做?

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



□ 循环神经网络(RNN): 常用于序列数据的特征提取



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



- □ 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 e
 - 常见应用: 文本分类, 文本推理, 翻译等

训练好的特 征即可直接 作为其它模 型的输入来 使用 Efficient estimation of word representations in vector space

<u>T Mikolov, K Chen, G Corrado, J Dean</u> - 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 ...

☆ 55 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

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

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

特征工程:可解释性

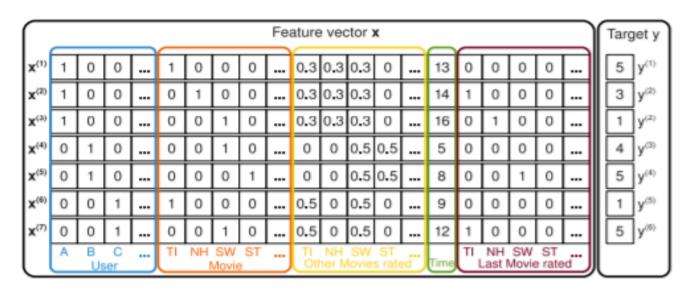
Factorization machines

11/5/2025

S Rendle - 2010 IEEE International confere In this paper, we introduce Factorization Macombines the advantages of Support Vector SVMs, FMs are a general predictor working

☆ 切 被引用次数: 1862 相关文章

- □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



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

特征工程:可解释性

"dog" "canine"

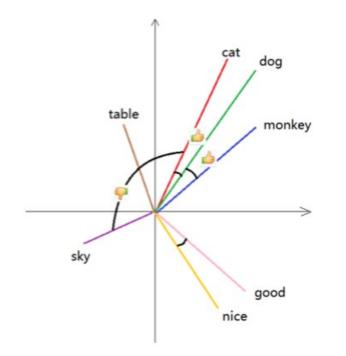
3 399,999

(0)
0
1
0
:
0
:
1

38

- □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

<u>T Mikolov</u>, <u>K Chen</u>, <u>G Corrado</u>, <u>J Dean</u> - 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 ...

☆ 57 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 are

☆ 切 被引用次数: 13612 相关文章 所有 18 个版本

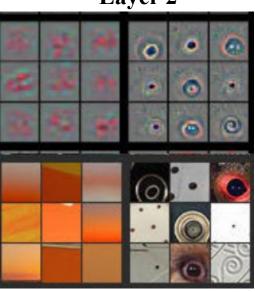
- □ 特征的含义—可解释性
 - □ CNN Feature map——特征图可视化
 - □ 学习图像的特点:图形图像的"颜色,边角,轮廓,图形"等

Layer 1





Layer 2

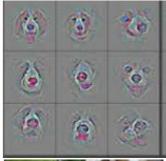


Layer 3





Layer 4

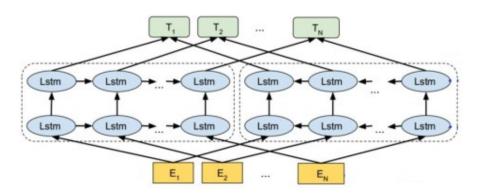




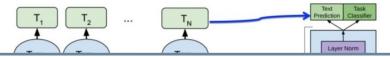
Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

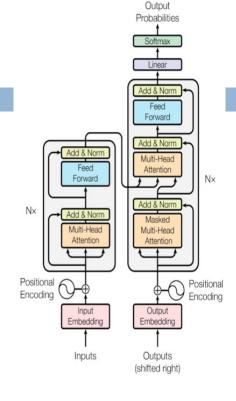
40

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

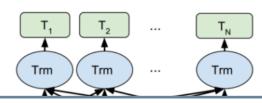


OpenAl GPT





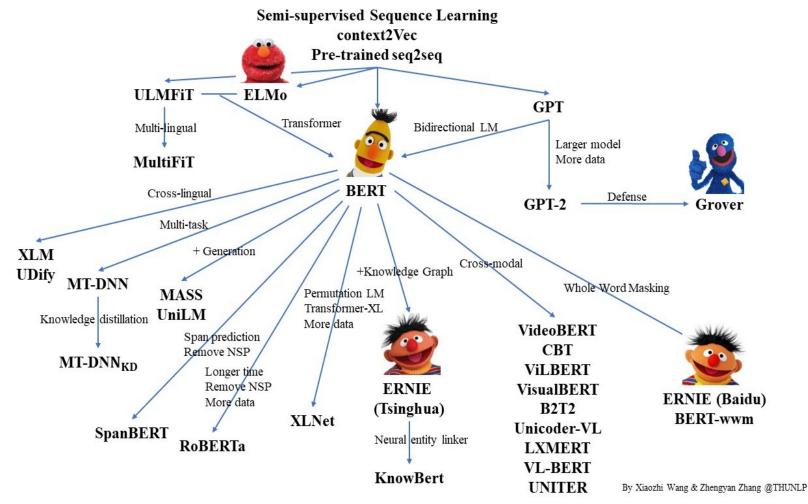
BERT (Ours)



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, graph等领域取得重要的突破。



□自然语言处理的预训练模型





42

□大语言模型的技术迭代

GPT

无监督预训练,有监督微调

5G文本数据

1.17亿模型参数

在9/12任务上最优,包括问答、语义相似度、文本分类

2018

GPT-2

多任务、零样本学习 (zero-shot)

40G文本数据

15亿模型参数

在7/8任务上最优,包括阅读理解、翻译、问答

2019

GPT-3

小样本学习 (few-shot)

45T文本数据

1750亿模型参数

在阅读理解任务上超越当时所有 zero-shot模型

2020

GPT-4o

多模态,可处理图像和文本输入

GPT-4的升级版模型,其中"O"是 Omni的缩写,意为"全能"。其在 响应速度、多模态能力、实时交互性 方面较GPT-4能力有极大的提升 GPT-4

多模态,可处理图像和文本输入

在大多数专业和学术考试中表现出人 类水平,且能通过律师资格考试,排 名考生中前10%,相较之下GPT-3.5排名低于后10% ChatGPT (3.5)

基于InstructGPT进行优化

能生成更翔实的回复: 标注数据质量

更高

更擅长连续对话: 源于标注人员标注

的多轮对话数据

捕获人类意图 进一步优化

大规模预训练

模型

2024.5 2023.3

2022.11

43

□ 前沿: 大语言模型

提出并应用了 GRPO, 增强了 数学推理能力

DeepSeek v1

2024.1 初步发现强化学 习能大幅度提高 模型推理能力; 提出了**R1使用的** MoE基本架构,

使用细粒度专家

和共享专家

2024.2

DeepSeekMath DeepSeek v2

> 2024.5 引入了多头潜在 注意力 (MLA) 和遵循

DeepSeekMoE 架构

提出MTP, 采用 FP8混合精度以及 DualPipe高效流 水线并行训练框架

DeepSeek v3

2024.12

DeepSeek R1

2025.1 推出DeepSeek-R1-Zero 和 DeepSeek-R1, 提出**基于规则奖** 励的纯强化学习, 以及多阶段SFT 和强化学习训练 流程



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

45

- □数据采集
 - □信息检索
 - □网络爬虫
- □数据预处理
 - □ 数据清洗
 - □ 数据集成
 - □数据变换
- □特征工程
 - □ 特征设计
 - □特征理解

Data Collection

Data Preprocessing

Feature Engineering