人工智能深度学习技术在医学考试试题难度预估中的应用研究

国家医学考试中心 清华大学 科大讯飞联合实验室

汇报人 吕萍
汇报时间 2019/10/18


背景
方法
结果
结论

背景

- **重要性**

  确定试题难度、固定合格分数线是行业准入考试的要求，也是考试权威性和公平性的体现

- **存在问题**

  传统的试题难度预估方法通过专家结合试题的知识点、自身专业背景以及对目标群体的掌握情况等因素对难度进行预估，存在预估结果不稳定、不同专家意见不统一、成本较高等缺点

- **技术尝试**

  人工智能技术的发展和大量数据的积累使得深度学习技术在试题难度预估的研究成为可能

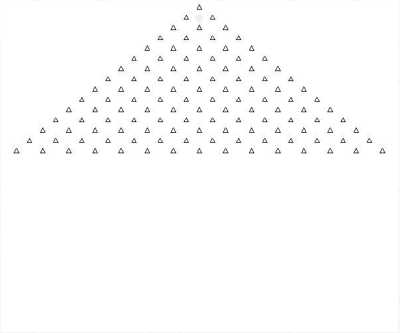
起源：回归分析

Francis Galton

在大数据分析中，回归分析是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。

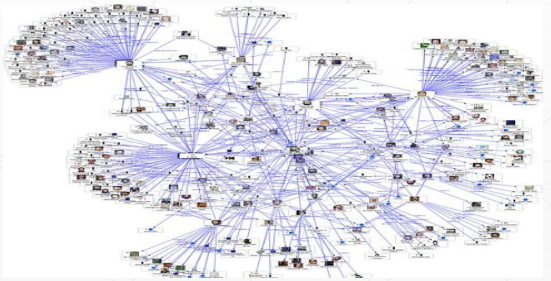
机场客流量预测

商品销售量预测

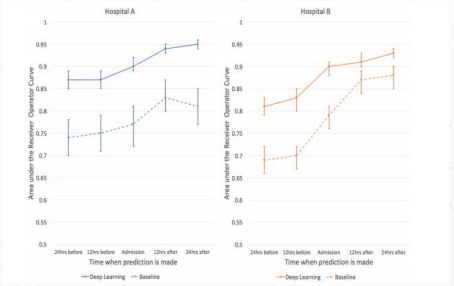


高尔顿板

《遗传的身高向平均数方向的回归》



微博转发量预测



出院、再入院预测

试题样例

男性，55岁。间歇性头晕、头痛1个月。近两年体检偶有血压高，未进一步诊治。有烟酒嗜好20余年。支气管哮喘、痛风病史10年，具体用药情况不详。查体：BP 155/95mmHg，心率55次/分，律齐。血尿酸550μmol/L。给予药物治疗后，患者反映有干咳，影响睡眠。

最可能引起此反应的药物是

A.普萘洛尔
B.氢氯噻嗪
C.卡托普利
D.哌唑嗪
E.氯沙坦

答案：C

难易度：中

认知层次：简单应用

大纲代码：110.12.4.3.3
（解释：临床医师-内科-心血管-高血压-治疗）

题型：A2

专业：心血管

类别：临床医师

研究方法

- **数据集**
  - 训练集：若干模拟题，若干考试真题，大量教科书等医学相关文本
  - 测试集：某年考试真题
- **难度预估方法**
  - 专家预估
  - 人工智能预估
- **研究假设（目的）**
  - 通过人工智能技术对试题进行难度预估
  - 人工智能预估、专家预估与实测难度进行比较
  - 分析各预估方法优劣及与实测难度的关系

专家预估难度

- **确定目标群体**
  - 教育测量学基本知识培训（试题评价理论、指标和评价方法）
  - 根据岗位胜任力标准确定临界组考生能力水平
- **往年实测数据学习**
  - 根据往年考试实测难度将试题排序，了解知识掌握程度
  - 分析讨论专家预估难度和实测难度之间差距的原因
- **多位专家新命试题难度预估**
  - 学科专家讨论新命试题的难易程度
  - 各学科专家单独预估，每题取平均值为最终预估难度

人工智能预估

- **属性模型**

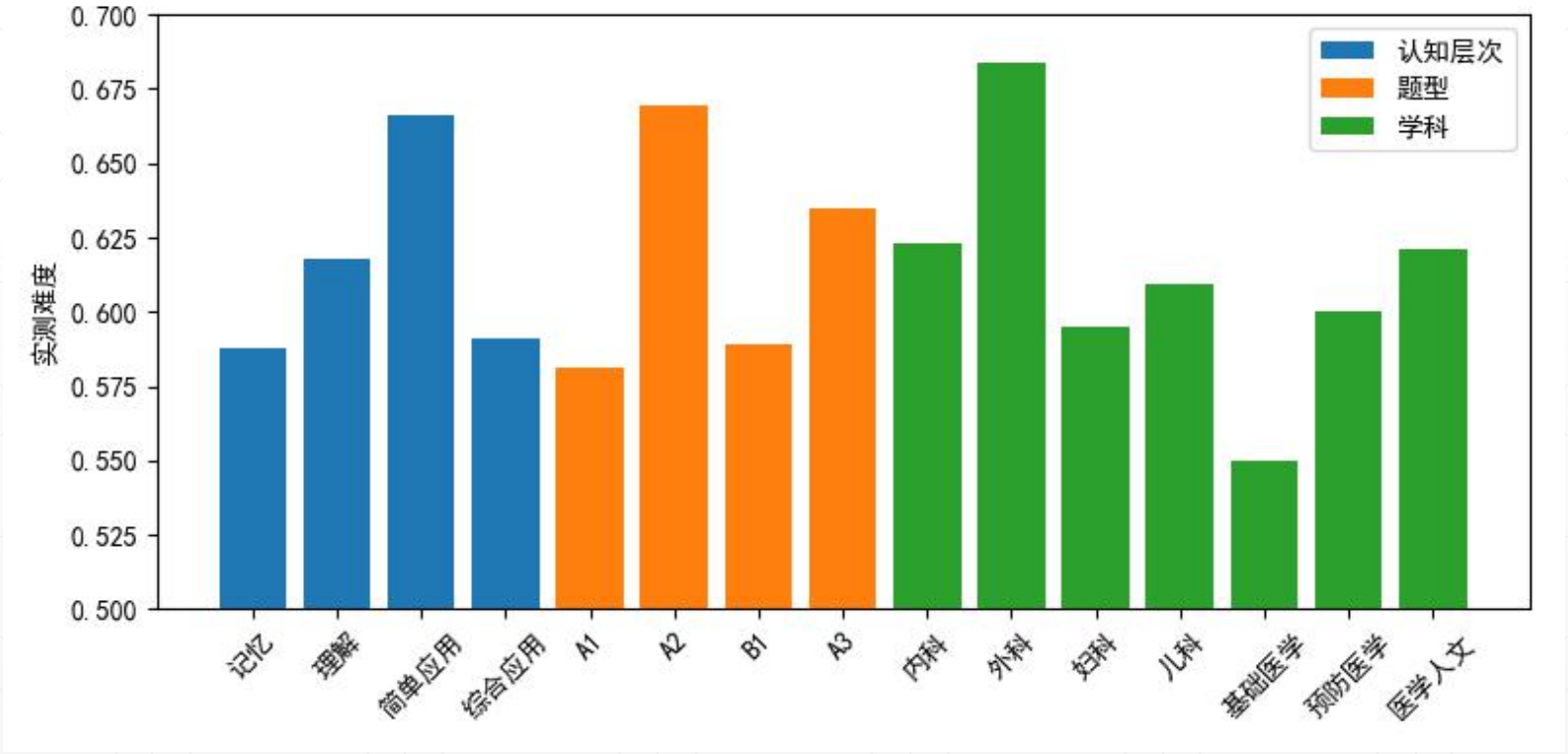
  从试题涵盖的知识点等属性因素考虑，符合专家出题的思路

- **语义模型**

  对试题进行语义分析和综合推理，符合考生做题的过程

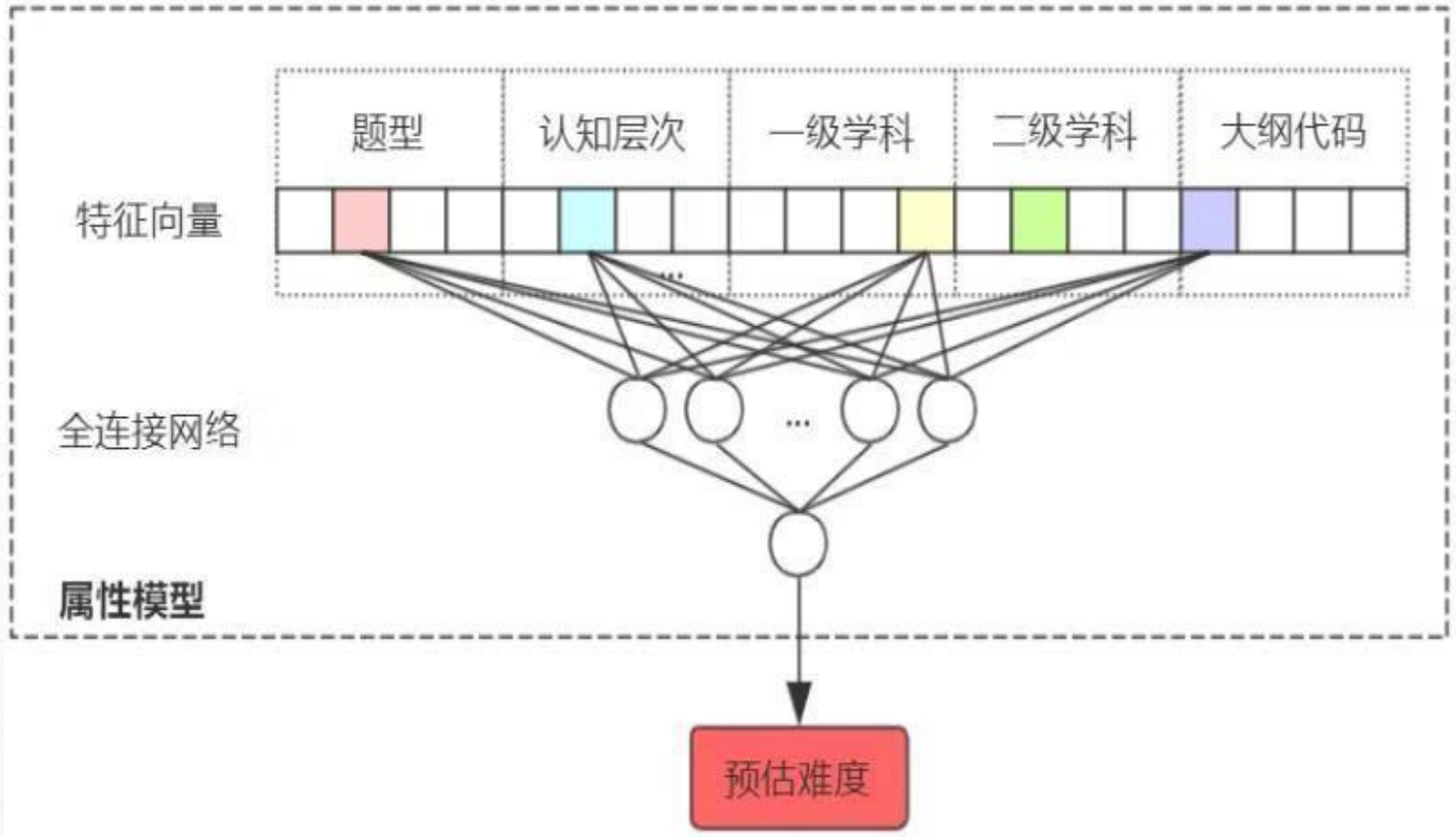
人工智能预估-属性模型

- 根据医学试题的命制规则，主要考虑试题题型、认知层次、学科、考试大纲等属性
- 统计试题在训练集上的实测难度如图



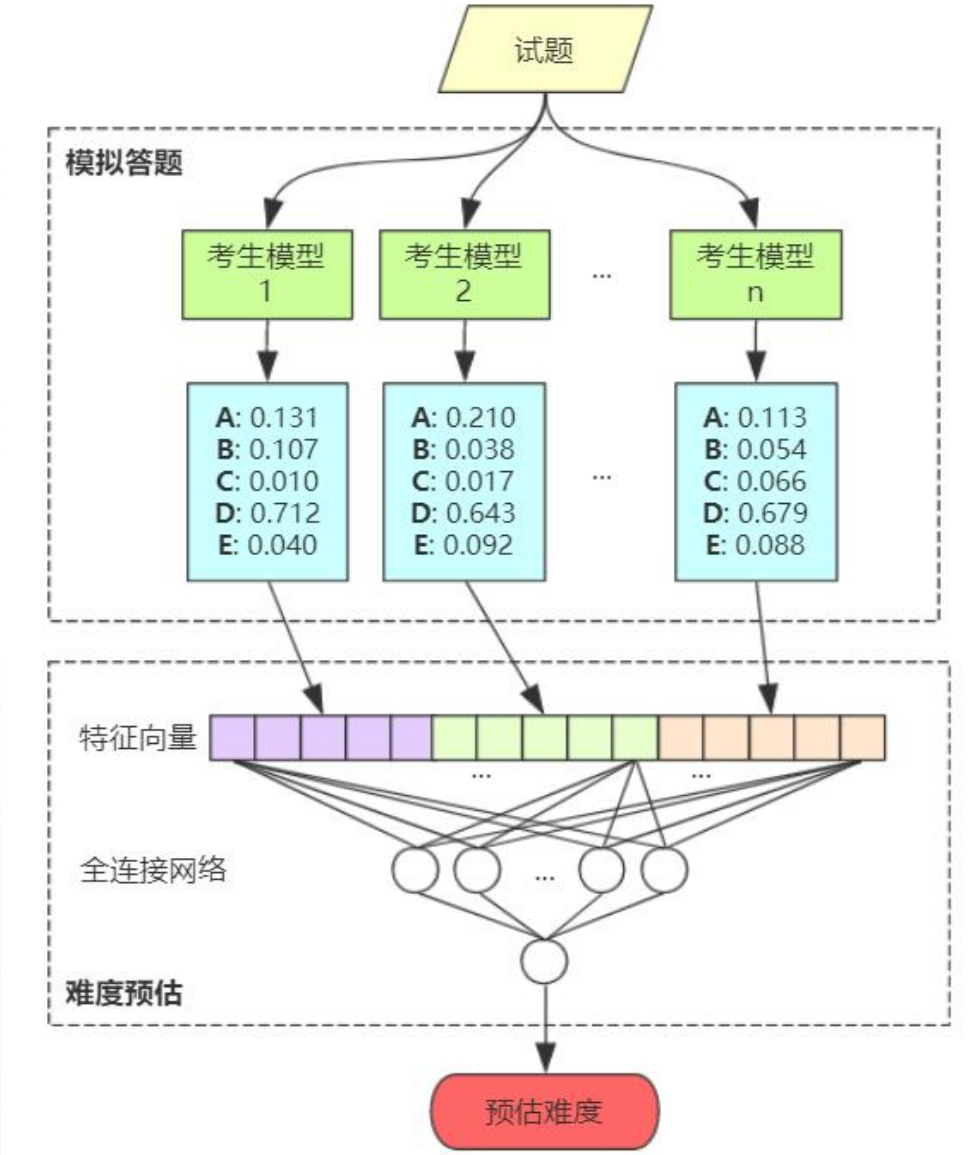
人工智能预估-属性模型

- 通过将试题属性进行特征提取和编码，经过多层神经网络模型进行试题难度的预估



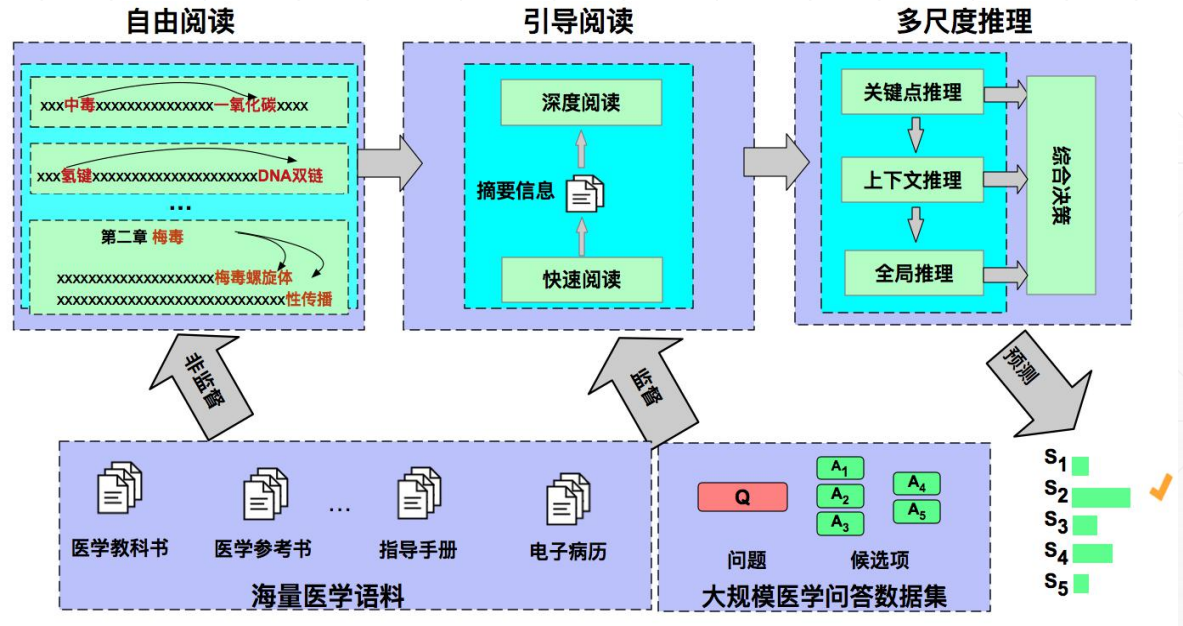
人工智能预估-语义模型

- 通过模拟不同的“考生群体”进行答题
- 从“考生群体”和“答题结果”中提取特征
- 经过多层神经网络模型进行试题难度的预估



人工智能预估-语义模型

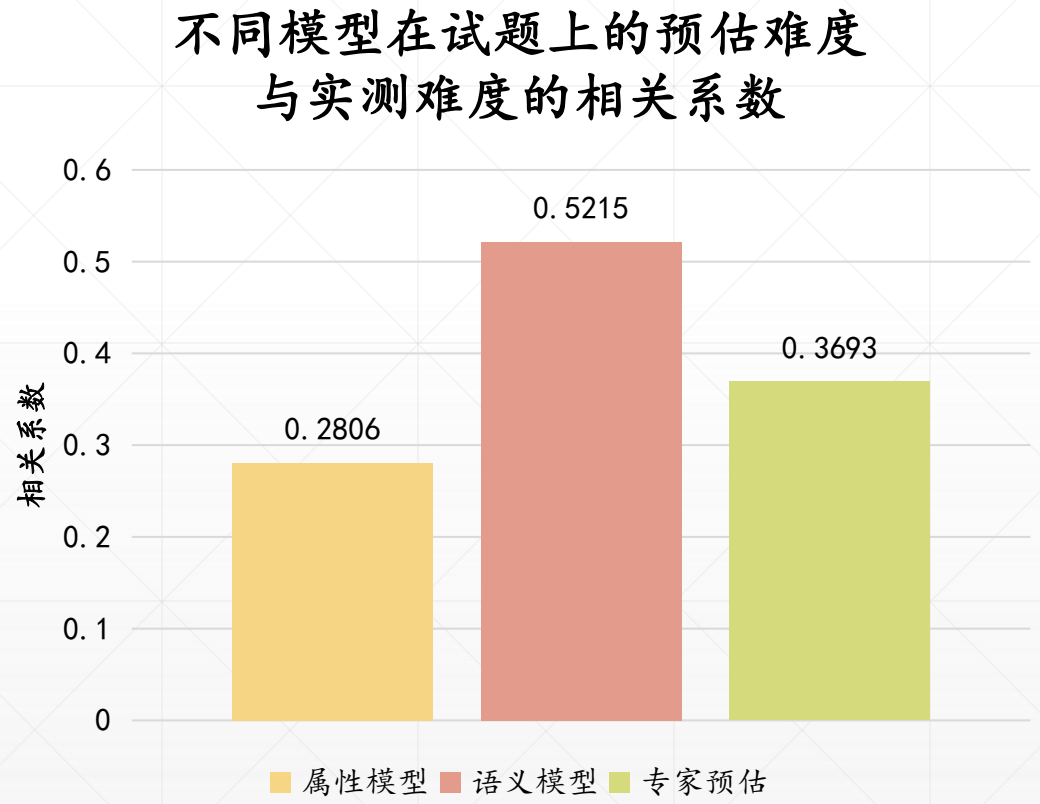
- 考生模型*
  1. 通过无监督模式泛读医学背景知识
  2. 通过有监督模式有针对性地学习试题的考点
  3. 构建多层语义推理模块进行试题的推理



* Wu Ji, Liu Xien, Zhang Xiao, He Zhiyang, Lv Ping. Master clinical medical knowledge at certificated-doctor-level with deep learning model.[J]. Nature communications,2018,9(1).

结果-相关分析

- 分别使用属性模型和语义模型在某年600道试题上进行难度预估测试，计算结果与实测难度的皮尔森相关系数如图



结果-方差分析

- 使用SPSS分析软件对各个模型的预估结果和实测结果进行单因素方差分析，其中事后检验的部分结果如下

多重比较						
因变量：难度						
LSD（最小显著差法）						
（I）方法	（J）方法	平均值差值（I-J）	标准误差	显著性	95% 置信区间	
					下限	上限
实测	专家	-.0322640*	.0090970	.000	-.050103	-.014425
	语义	.0153247	.0090970	.092	-.002514	.033164
	属性	.0342929*	.0090970	.000	.016454	.052132

结论

- 语义模型预估的试题难度比专家主观预估的难度更接近试题的实测难度
  - 可尝试将语义模型用于辅助专家进行试题难度的预估
  - 可尝试将语义模型的方法推广到其他考试的试题难度预估任务中
- 深度学习技术在试题难度预估任务上体现出可行性，相较传统方法有一定的优势