北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

所谓学习，归类而已

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

做两道选择题

1、哪些形容词可以用来修饰机器学习到的规律/知识/模式/模型（　）

A、正确的　　B、有趣的

C、有用的　　D、令人新奇的

E、其它_____

2、就逻辑推理方式而言，机器学习 / 数据挖掘属于（　）

A、归纳　　B、演绎

C、都是　　D、都不是

归纳与演绎

归纳 科学推理 演绎

从特殊到一般：

从特殊事实或个别事例到一般结论的逻辑推理方法

每次走过隔壁老王的小卖部，他家的狗都会对我大叫，但是不咬我。因此，下次我再经过小卖部的时候，他的狗也不会咬我

从一般到特殊：

从普遍性结论或一般性事理推导出个别性结论

三段论推理示例

大前提：叫声大的狗都不咬人

小前提：隔壁老王家的狗叫声大

结论：隔壁老王家的狗不会咬人

机器学习主要的推理方式是归纳

- 《阿培丁·机器学习导论》：几乎所有的科学领域都在用模型拟合数据。科学家们设计实验、进行观测并收集数据。然后，通过找寻能解释所观测数据的简单模型，尝试抽取知识。**该过程称为归纳**，是从一组特别的示例中提取通用规则的过程
- 《周志华·机器学习》：归纳与演绎是科学推理的两大基本手段……而“从样例中学习”显然是一个**归纳**的过程

归纳与演绎

- 《道德经》：万物并作，吾以观复
- 《阅微草堂笔记》：无往不复，天之道也
- 由于在历史上一次次出现，我们也就有理由相信它在未来也会延续这种模式
- 归纳法由于没有（或无法）穷举考察对象的全体，因此它的结论带有猜想的性质，属于**似真推理，或然性推理**
- 要确保归纳出的知识是正确的，前提是自然是齐一的
- 归纳法得出的结论并没有办法直接证明，即便我们一次次证实了它

归纳与演绎

欧洲人观察了成百上千年，发现天鹅都是白的。因此，他们得到一个结论，天鹅都是白的。后来发现了澳洲，一上岸，发现居然天鹅是黑的……

归纳与演绎

一只老母鸡，被养了将近三年。它归纳总结出了1000天的经验模式：主人对我真好，每次伸手过来都是喂我好吃的。但是，大年三十那天，主人伸手过来没有给他喂食，而是抓住它的脖子，把它炖成了母鸡汤。于是，老母鸡通过归纳得出的结论被无情的推翻了

回到前述的选择题

1、哪些形容词可以用来修饰机器学习到的规律/知识/模式/模型（　）

A、正确的　　B、有趣的

C、有用的　　D、令人新奇的

E、其它_____

2、就逻辑推理方式而言，机器学习 / 数据挖掘属于（　）

A、归纳　　B、演绎

C、都是　　D、都不是

看看别人怎么形容规律/知识/模式/模型

- 《韩家炜等·数据挖掘：概念与技术》：数据挖掘是从大量数据中挖掘有趣模式和知识的过程
- 《Pang-Ning Tan等·数据挖掘导论（完整版）》：数据挖掘是在大型数据存储库中，自动地发现有用信息的过程。数据挖掘技术用来探查大型数据库，发现先前未知的有用模式
- (Fayyad et al., 1996): 识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程

Fayyad, U., Piatetskyshapiro, G., & Smyth, P. (1996). Knowledge discovery and data minin: towards a unifying framework. Shapiro, 82--88.

看看别人怎么形容规律/知识/模式/模型

有效的	Valid	Discovered patterns should be true on new data with some degree of certainty. Generalize to the future (other data).
新颖的	Novel	Patterns must be novel (should not be previously known).
潜在有用的	Useful	Actionable; patterns should potentially lead to some useful actions.
最终可理解	Understandable	The process should lead to human insight. Patterns must be made understandable in order to facilitate a better understanding of the underlying data.

Fayyad, U., Piatetskyshapiro, G., & Smyth, P. (1996). Knowledge discovery and data minin: towards a unifying framework. Shapiro, 82--88.

机器学习/数据挖掘是从大量的数据中归纳出（先前未知的）有用或有趣关系结构（模式、模型、知识、规律、…）的过程

谢谢聆听

Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics