Data Preprocessing

- 关于数据
- 为什么要预处理数据？
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据规约
- 小结

数据清理任务

- ✓ 填充遗漏值
- ✓ 标记离散点和平滑噪音数据
- ✓ 修正不连续数据
- ✓ 解决数据整合引起的冗余

数据遗漏可能是由于

- ✓ 设备故障
- ✓ 与其他记录的数据不一致因此删掉了
- ✓ 由于误解没有将数据输入
- ✓ 数据的历史记录或改变

丢失的数据可能需要推测

怎么处理丢失的数据?

- 忽略这个元组
- 人为填充遗失数据
- 用下面的方法自动填充
  - ✓ 一个全局变量：例如,“未知”, 一个新的类别?!
  - ✓ 属性的均值
  - ✓ 属于同一类别的所有样本的均值: 更灵活
  - ✓ 最大可能值：推理为基础的，如贝叶斯公式或决策树

有噪音数据

- 噪音:被测变量中的随机错误或者不一致
- 不正确的属性值可能是由于
  - ✓ 有缺陷的数据收集设备 **(Ex.1:** 摄像机**)**
  - ✓ 数据输入错误
  - ✓ 数据传输问题 **(Ex.2:** 监控电视内容**)**
  - ✓ 技术限制
  - ✓ 命名约定不一致（命名约定）
- 其他需要数据清理的数据问题
  - ✓ 重复的记录
  - ✓ 不完整的数据
  - ✓ 不连续的数据

怎么处理有噪音的数据?

Binning（分箱）

- ✓ 先对数据排序并（等频率）分割成箱
- ✓ 然后可以通过箱均值，箱中位数，箱边界值等平滑.

聚类

- ✓ 检测和移除离散点

结合计算机和人工检查

- ✓ 检测可疑值并人工核准 (例如, 处理可能的离散点)

回归

- ✓ 对数据进行回归函数拟合进行平滑

简单的离散化方法: 分箱

等距 分区:

- ✓ 分成**N**个大小相等的区间: 均匀网格
- ✓ 如果**A**和**B**是属性值的最小值和最大值，间隔的宽度就是：$W = (B - A)/N$.
- ✓ 最简单的，但是离散值可能主导呈现的形式
- ✓ 处理有偏的数据不是很好.

等深 (频率) 分区:

- ✓ 分成**N**个区间，每个区包含近似相同数量的样本
- ✓ 好的数据尺度
- ✓ 管理类别的属性可能会非常棘手.

平滑数据的分箱方法

- 按价格对数据排序(以美元计): **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

* 分成等频率的箱:

  - Bin 1: 4, 8, 9, 15

  - Bin 2: 21, 21, 24, 25

  - Bin 3: 26, 28, 29, 34

* 用箱均值平滑数据:

  - Bin 1: 9, 9, 9, 9

  - Bin 2: 23, 23, 23, 23

  - Bin 3: 29, 29, 29, 29

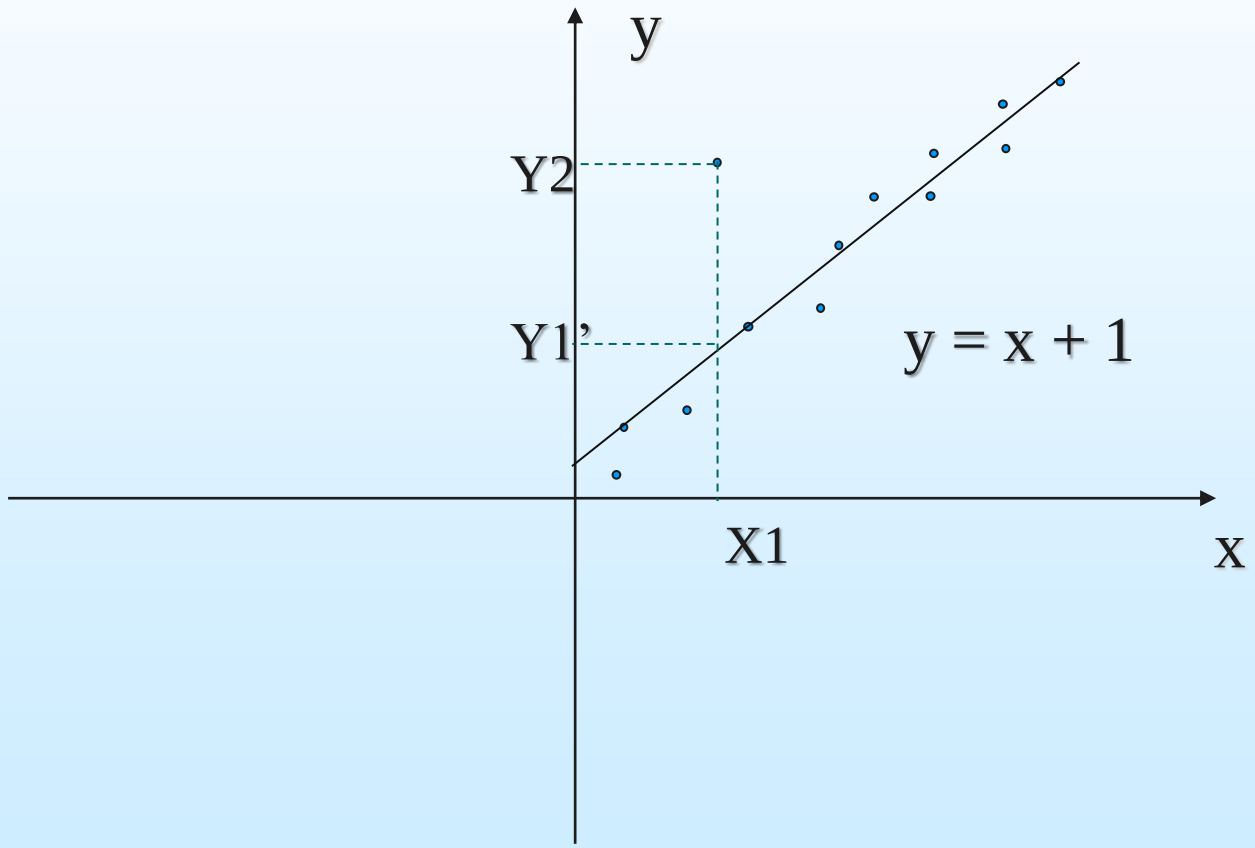
* 用边界值平滑数据:

  - Bin 1: 4, 4, 4, 15

  - Bin 2: 21, 21, 25, 25

  - Bin 3: 26, 26, 26, 34

回归



聚类分析