# 平时作业 2

1160300314 朱明彦

November 9, 2019

## 1 作业要求

基于统计的异常探测问题中，有一种非参数方法称为直方图异常检测法。如图1所示，通过按照数据范围进行划分，将数据分配到相应的直方图中，可以统计出每个直方图中数据所占的比例。

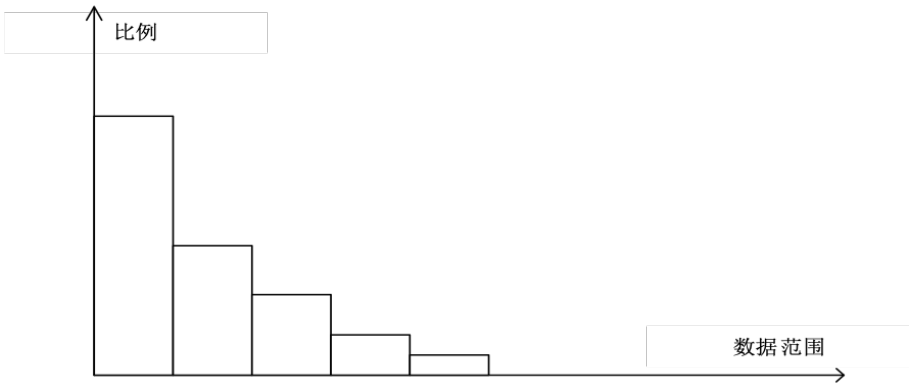


Figure 1: 直方图示意

请设计一种算法，利用直方图进行异常检测

- 给出构造直方图和检测异常点的主要步骤。
- 给出异常得分的计算公式。

## 2 作业解答

### 2.1 主要步骤

以下算法伪代码参考 [2] 给出。

算法 1 基于直方图的异常检测算法

输入: 原始数据集合 $S$, $N$ 为 $S$ 的大小, $d$ 为每个原始数据的维度, $k$ 每个维度桶数

输出: HBOS（Histogram-based Outlier Score）

**function** HBOS_CAL($S$, $N$, $d$, $k$, 样本 $p$)
  $result \leftarrow 0, i \leftarrow 0$
  **while** $i < d$ **do**
    $histogram_i$ = GETHISTOGRAM($S$, $N$, $i$, $k$)
    $result = result + \log\left(\frac{1}{histogram_i(p)}\right)$
  **end while**
  **return** $result$
**end function**
**function** GETHISTOGRAM($S$, $N$, $i$, $k$)
  **if** $i$ 维数据为离散数据 **then**
    对于每一个种类进行计数, 以频率进行作为高度
  **else**
    // $i$ 维为连续数据, 静态的桶划分方法
    将数据范围均匀的划分为 $k$ 个桶, 以落入每个桶的频率作为高度
  **end if**
  **return** $histogram_i$
**end function**

在上述算法中, 对于连续的数据采取了静态的桶划分方法, 这种方法比较快速, 但是在处理"Long Tailed" 型数据时表现不佳. 此时可以使用动态的桶划分方法: 首先将数据排序, 将数据均匀划分, 保证每个桶的数据大致为 $\frac{N}{k}$(对于整型数据, 保证相同值在同一个桶中) [2].

得到每个样本的 HBOS 值之后, **对于 HBOS 值较小的样本, 即可以认为是异常样本.**

## 2.2 异常得分公式

对于样本 $p$, 其异常值计算得分计算公式如下 [2]:

$$HBOS(p) = \sum_{i=1}^{d} \log\left(\frac{1}{hist_i(p)}\right)$$

推导过程如下, 假设 $p$ 的第 $i$ 个特征的概率密度为 $P_i$, 则得到 $p$ 的概率密度可以计算为

$$P(p) = P_1(p)P_2(p)\cdots P_d(p)$$

两边取对数则有

$$\begin{aligned}\log(P(p)) &= \log\left(P_1(p)P_2(p)\cdots P_d(p)\right)\\ &= \sum_{i=1}^{d}\log\left(P_i(p)\right)\end{aligned}$$

由于概率密度越大, 其异常评分应该越小, 所以取反

$$-\log(P(p)) = -1\sum_{i=1}^{d}\log\left(P_i(p)\right) = \sum_{i=1}^{d}\frac{1}{\log\left(P_i(p)\right)}$$

从而有

$$HBOS(p) = -\log(P(p)) = \sum_{i=1}^{d}\frac{1}{\log\left(P_i(p)\right)}$$

# References


[1] Matthew Gebski and Raymond K Wong. An efficient histogram method for outlier detection. In *International Conference on Database Systems for Advanced Applications*, pages 176–187. Springer, 2007.

[2] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. 2012.