Module 7:Data Representation

Lecture 36: V-optimal Histograms

Prev topic Next topic Prev page
Next page

The Lecture Contains:

V-optimal histograms

- A histogram of data with n bins
- Reduce n to b where $b \ll n$

V-optimal histograms

- A histogram of data with n bins
- Reduce n to b where $b \ll n$
- Formally, assume a set V of n (sorted) values $v_1, v_2, \ldots, v_n$ having frequencies $f_1, f_2, \ldots, f_n$ respectively
- Problem is to output another histogram H having b bins, i.e., b non-overlapping intervals on V
- Interval is of the form $[l_i, r_i]$ and has a value h_i
- If value $v_j \in I_i$, estimate $e(v_j)$ of f_j is h_i
- Error in estimation is distance $d(f, e)$

Module 7:Data Representation

Lecture 36: V-optimal Histograms

Details

- Histogram value h_i is average of values in l_i , i.e.,

$$h_i = ([l_i, r_i]) = \left(\sum_{k=l_i}^{r_i} f_k \right) / (r_i - l_i + 1)$$

- Error function is sum squared error (SSE) (or L_2 error)

$$SSE([l, r]) = \sum_{k=l}^{r} \left(f_k - \text{avg } ([l, r]) \right)^2$$

Algorithm

- Assume optimal partitioning for the first i values with at most k bins is $SSE^*(i,k)$
- Consider placement of the last bin
- Choice is any of the i gaps
- For each such placement at gap j , at most $k-1$ bins have been placed optimally for the first j values
- This leads to the recursion

$$SSE^*(i,k) = \min_{1 \le j \le i} \{SSE^*(j,k-1) + SSE([l_{j+1}, r_i])\}$$

Algorithm

- Assume optimal partitioning for the first i values with at most k bins is $SSE^*(i,k)$
- Consider placement of the last bin
- Choice is any of the i gaps
- For each such placement at gap j , at most $k-1$ bins have been placed optimally for the first j values
- This leads to the recursion

$$SSE^*(i,k) = \min_{1 \le j \le i} \{SSE^*(j,k-1) + SSE([l_{j+1}, r_i])\}$$

- Dynamic programming (DP) solution
- Table of size $n \times b$
- Start with cell $(1,1)$ and proceed in a column-scan order
- Computation for cell (i,k) requires values at cells $(j,k-1),\ \forall 1 \le j \le i$

Algorithm

- Assume optimal partitioning for the first i values with at most k bins is $SSE^*(i,k)$
- Consider placement of the last bin
- Choice is any of the i gaps
- For each such placement at gap j , at most $k-1$ bins have been placed optimally for the first j values
- This leads to the recursion

$$SSE^*(i,k) = \min_{1 \le j \le i} \{SSE^*(j,k-1) + SSE([l_{j+1}, r_i])\}$$

- Dynamic programming (DP) solution
- Table of size $n \times b$
- Start with cell $(1,1)$ and proceed in a column-scan order
- Computation for cell (i,k) requires values at cells $(j,k-1),\ \forall 1 \le j \le i$
- Running time: $O(n^2 b)$