**3.4.5. Mô hình hồi quy tuyến tính và log-linear: Giảm thiểu dữ liệu tham số**

**-** Có thể được sử dụng để tính toán xấp xỉ các dữ liệu đã cho.

- Trong hồi quy tuyến tính, dữ liệu được mô hình hóa theo một đường thẳng. Biến y (biến phản hồi), có thể được mô hình hóa như một hàm tuyến tính của một biến khác biến ngẫu nhiên, x (được gọi là biến dự đoán), với phương trình

y = ax + b

( Phương sai của y được giả định là hằng số. x và y là các thuộc tính cơ sở dữ liệu số. Các hệ số w và b được gọi là hệ số hồi quy)

* Đa hồi quy tuyến tính là một phần mở rộng của hồi quy tuyến tính (cơ bản), cho phép một biến phản hồi, y được mô hình hóa như một hàm tuyến tính của hai hoặc nhiều biến dự đoán
* Các mô hình log-linear xấp xỉ các phân bố xác suất đa chiều rời rạc.
* Mô hình log-linear có thể được sử dụng để ước tính xác suất của mỗi điểm trong một không gian đa chiều cho một tập hợp các thuộc tính rời rạc, dựa trên một tập hợp con nhỏ hơn của các kết hợp chiều.
* Điều này cho phép một không gian dữ liệu chiều cao hơn được xây dựng từ các không gian chiều thấp hơn. Do đó, các mô hình log-linear cũng hữu ích cho việc giảm tính chiều hướng và làm mịn dữ liệu

*Kết luận:*

Cả hai mô hình hồi quy và log-linear đều có thể được sử dụng trên dữ liệu thưa thớt, mặc dù chúng ứng dụng có thể bị hạn chế. Trong khi cả hai phương pháp đều có thể xử lý dữ liệu bị lệch, hồi quy làm đặc biệt tốt.

Một số phần mềm tồn tại để giải quyết các vấn đề hồi quy. Ví dụ: *e SAS (www.sas.com), SPSS (www.spss.com), and S-Plus (www.insightful.com). Another useful resource is the book Numerical Recipes in C, by Press, Teukolsky, Vetterling, and Flannery [PTVF07].*

**3.4.6. Biểu đồ tần suất**

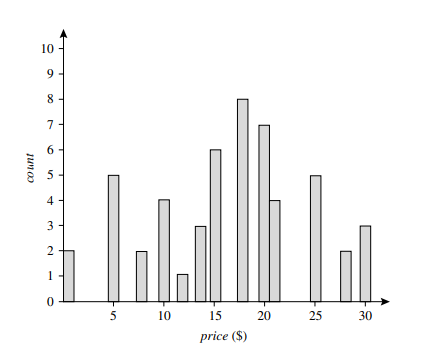
Biểu đồ tần suất sử dụng binning để phân phối dữ liệu gần đúng và là một dạng phổ biến giảm dữ liệu. Biểu đồ đã được giới thiệu trong Phần 2.2.3. Biểu đồ tần suất cho một thuộc tính, A, phân vùng phân phối dữ liệu của A thành các tập con rời rạc, được gọi là buckets hoặc bins.

* Nếu mỗi vùng lưu trữ chỉ đại diện cho một cặp giá trị/tần suất thuộc tính duy nhất, thì biểu tượng buckets được gọi là singleton buckets.

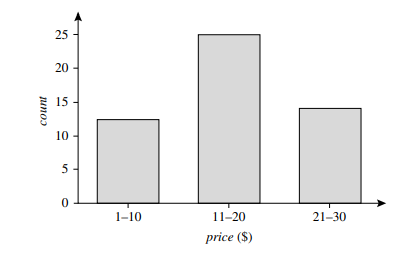
***Ví dụ*** 3.3: Biểu đồ tần suất. Dữ liệu sau đây là danh sách giá AllElectronics thường được bán các mặt hàng (làm tròn đến đô la gần nhất). Các số đã được sắp xếp: 1, 1, 5, 5, 5,5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,

18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,

30, 30.



3.7 Biểu đồ tần suất cho giá sử dụng các singleton buckets — mỗi vùng lưu trữ đại diện cho một giá–giá trị/ cặp tần số.



3.8 Biểu đồ tần suất có chiều rộng bằng nhau cho giá, trong đó các giá trị được tổng hợp để mỗi vùng lưu trữ có một chiều rộng đồng nhất là 10$.

*Câu hỏi được đặt ra :*

"Các vùng lưu trữ được xác định như thế nào và các giá trị thuộc tính được phân vùng như thế nào?"

*Trả lời*

Có một số quy tắc phân vùng, bao gồm những điều sau đây:

* Chiều rộng bằng nhau: Trong biểu đồ chiều rộng bằng nhau, chiều rộng của mỗi phạm vi vùng lưu trữ là đồng phục
* Tần số bằng nhau (hoặc độ sâu bằng nhau): Trong biểu đồ tần số bằng nhau, các vùng lưu trữ là được tạo ra sao cho, đại khái, tần suất của mỗi vùng lưu trữ là không đổi

*Tác dụng của biểu đồ :*

* Biểu đồ có hiệu quả cao trong việc xấp xỉ cả dữ liệu thưa thớt và dày đặc, như cũng như dữ liệu có độ lệch cao và thống nhất. Các biểu đồ được mô tả trước đây cho đĩa đơn thuộc tính có thể được mở rộng cho nhiều thuộc tính. Biểu đồ đa chiều có thể nắm bắt sự phụ thuộc giữa các thuộc tính. Những biểu đồ này đã được tìm thấy có hiệu quả trong xấp xỉ dữ liệu với tối đa năm thuộc tính. Cần có thêm các nghiên cứu liên quan đến hiệu quả của biểu đồ đa chiều cho tính chiều cao.
* Singleton buckets rất hữu ích để lưu trữ các ngoại lệ tần số cao.