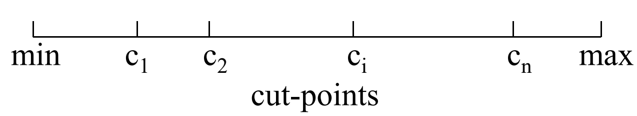
Data Preprocessing

- Why preprocess the data?
- Data integration
- Data cleaning
- transformation
- Data reduction
  - Feature Selection
  - Case Reduction
  - Value Reduction
- Discretization

49

What is discretization?

- A discretization algorithm
  - converts continuous attributes into discrete attributes by partitioning the range of a continuous attribute into intervals.
  - Interval labels can then be used to replace actual data values.



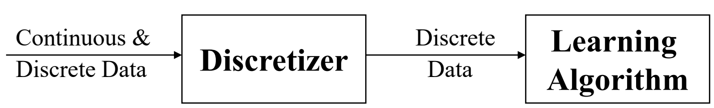
50

Why Need Discretization?

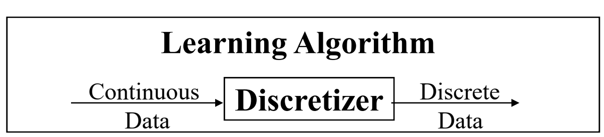
- Some learning algorithms are limited to discrete inputs.
- Efficiency: handling (lots of) continuous values tends to slow down learning considerably. (*Value reduction*)
- Accuracy: in the presence of noise good discretization can sometimes improve predictive accuracy. (*Smoothing out noise*)
- Intelligibility: discretization may lead to smaller sizes of induced trees or rule sets.

Two Architectures

- Discretization before learning starts (Static discretization)

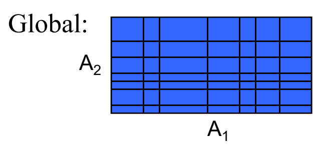


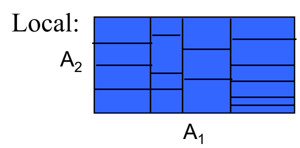
- Discretization during the learning process (Dynamic discretization)



Classification of Discretization Methods

- Supervised vs. unsupervised.
  - Supervised discretization uses class information.
  - Unsupervised does not use class labels.
- Bottom-up vs. top-down
  - Bottom-up: start from intervals with one value each and repeatedly merge intervals until some stopping criterion is satisfied.
  - Top-down: start from one interval with all values and repeatedly split intervals until some stopping criterion is satisfied.
- Global vs. local
  - Global: an attribute is partitioned over the entire continuous range, using global information and independent of other attributes.
  - Local: partition is applied to local regions of an attribute range.





Unsupervised Discretization

- Equal-width binning
  - Use discrete values, such as 1, 2, 3, …, to represent intervals instead of bin means or boundaries
- Equal-depth/frequency binning
  - Use discrete values, such as 1, 2, 3, …, to represent intervals instead of bin means or boundaries
- k-means clustering
  - Given k bins, distribute the values in the bins to minimize the average distance of a value from its bin mean.

K-mean Clustering

- Input: (1) a set of values for an attribute
  (2) k = number of bins
- Sort the input values and keep the unique values
- Create k bins using equal-depth binning
- Compute bin means ($mean_1$, $mean_2$, ……, $mean_k$)
- Compute global distance: $D_{new} = \sum_i \sum_j (v_{ij} - mean_i)^2$
  where $mean_i$ is the mean in bin_i and v_{ij} is the jth value in bin_i.
- Repeat
  - $D_{old} = D_{new}$
  - for each bin_i
    - for each v_{ij} in bin_i
      - If (v_{ij}- $mean_{i-1}$) < (v_{ij}-$mean_i$), move v_{ij} to bin_{i-1}.
      - If (v_{ij}- $mean_{i+1}$) < (v_{ij}-$mean_i$), move v_{ij} to bin_{i+1}.
  - Compute new bin means and D_{new}
- Until D_{new} is not less than D_{old}.

55

Supervised Discretization

- ChiMerge
  - Based on chi-square test
- Entropy-based discretization method
  - Based on an entropy minimization heuristic

56

ChiMerge: a Bottom-up Supervised Method

- ChiMerge is based on the statistical χ^2 test
- The purpose of a χ^2 test is to determine whether two variables are related.
  - E.g., we want know if there is any relationship between the gender of undergraduate students in a university and their footwear preferences.
- Observations about the two variables in a sample are usually expressed in a contingency table:

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

57

Chi Square Significance Test

- The null hypothesis is that the two variables are unrelated (that is, only randomly related).
- χ^2 test determines whether we should reject the null hypothesis and at what significance level (p-value) we should reject the null hypothesis.
- For the example in the previous slide,
  - The null hypothesis is that gender is unrelated with footwear preference
  - But the χ^2 test shows that we should reject this hypothesis at the significance level of 0.01, which means that we are 99% sure that gender and footwear preferences are related.
  - Usually, p-value should be at most 0.05 in order to reject the null hypothesis.

58

How to Calculate χ^2

- Given the contingency table:

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

- Compute the expected frequency for each cell
  - The expected frequency of $\text{cell}_{i,j}$ is

$$E_{ij} = \frac{\text{the total of row i} \times \text{the total of column j}}{\text{sample size}}$$

  - For example, the expected frequency of the upper left cell is $\dfrac{50 \times 19}{100}$

How to Calculate χ^2 (Cont'd)

- Compute the chi-square value for the table

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

  - Let O_{ij} denote the observed value in $\text{cell}_{i,j}$

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  - For example, the chi-square value of the above table is 14.026

How to Calculate χ^2 (Cont'd)

- Calculate the degrees of freedom for the table

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

$$df = (r-1)(c-1)$$

- where *r* is the number of rows and *c* is the number of columns
- For example, the degrees of freedom for the above table is 4.
- This is because, given row or column totals, all but one of the values in a given row or column are free to vary.

61

How to Calculate χ^2 (Cont'd)

- Using the chi-square table to determine the p-value for rejecting the null hypothesis

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
…	…	…	…

- The table lists the critical values (i.e., thresholds)
- The calculated chi-square value for a contingency table must be greater than the critical value corresponding to the df of the table and a p-value (e.g., 0.05) in order to reject the null hypothesis at the significance level (p-value).

62

ChiMerge: a Bottom-up Supervised Method

- Sort all examples according to the values of the attribute to be discretized.
- Place each value in its own interval.
- Merge intervals repeatedly in the following manner:
  - For each pair of adjacent intervals:
    - Calculate the χ^2 value: $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

      where k = # of classes, O_{ij}= # of examples in the ith interval and jth class, E_{ij}= expected frequency of O_{ij}= $\frac{R_i \times C_j}{N}$, in which N is # of examples, R_i=# of examples in the *ith* interval, and C_j= # of examples in the *jth* class.
  - If the lowest χ^2 value is smaller than a threshold, merge the two adjacent intervals with the lowest χ^2 value.
  - This process is repeated until all χ^2 values exceeds this threshold.
- The threshold can be obtained from the standard χ^2 table

63

Entropy-Based Discretization

- Supervised, top-down discretization
- Employs an entropy minimization heuristic for splitting the range of a continuous attribute.
- Given a set *S* of examples and *k* classes, the *entropy* of *S* with respect to the *k* classes is defined as:

$$Ent(S) = -\sum_{i=1}^{k} P(C_i)\log_2(P(C_i))$$

  where $P(C_i)$ is the probability of examples in S that belong to C_i.
- The bigger *Ent(S)* is, the more impure *S* is.

64

Entropy-Based Discretization

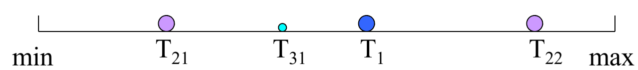
Given an attribute *A* and a set *S* of training examples:

- Sort the examples in a set *S* by increasing values of the attribute *A*: $\{v_1, v_2, \ldots, v_n\}$.
- A potential cut-point *T*: midpoint between v_i and v_{i+1} dividing *S* into S_1: $\{v_1, v_2, \ldots, v_i\}$ and S_2: $\{v_{i+1}, \ldots v_n\}$.
- A total of *n-1* potential cut-points.
- Suppose a cut-point *T* partitions *S* into S_1 and S_2. Entropy (with respect to the class attribute) after the partition induced by cutpoint *T*:

$$Ent(T, S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

where $|S|$, $|S_1|$ and $|S_2|$ = # of examples in S, S_1 and S_2

- Select T_A for which $E(T_A, S)$ is minimal to split the range into two subranges
- The process is recursively applied to partitions obtained until some stopping criterion is met.



65

Stopping Criteria for Entropy-Based Discretization

- Stopping criteria in D-2 (Catlett, 1991):

  Recursive partitioning stops if any of the following is satisfied:
  - all the examples in the interval belong to the same class.
  - number of examples in an interval is below a given level;
  - maximum number of cut-points for an attribute is reached;
  - the entropy reduction on all possible cut-points is equal;
- Stopping criterion based on Minimum Description Length Principle (MDLP) (Fayyad and Irani, 1993):

  Recursive partitioning stops iff

$$Ent(S) - Ent(T, S) \le \frac{\log_2(N-1)}{N} + \frac{\Delta(T;S)}{N}$$

$$\Delta(T;S) = \log_2(3^k - 2) - [kEnt(S) - k_1 Ent(S_1) - k_2 Ent(S_2)]$$

where k, k_1 and k_2 are the number of classes in S, S_1 and S_2, respectively, and N is the number of examples in S.

66

Summary

- Data preparation is a big issue for data mining
- Data preparation includes
  - Data integration
  - Data cleaning
    - Handle missing values
    - Detect and remove noise
  - Data transformation
  - Data reduction
    - feature selection, case reduction and value reduction
  - Discretization
- A lot of methods have been developed but still an active area of research

Readings

- Chapter 3 in Jiawei Han's book
- Chapters 3 and 4 in "Predictive Data Mining, a Practical Guide" by Sholom M. Weiss and Nitin Indurkhya.
- U. M. Fayyad and K. B. Irani, "*Multi-interval discretization of continuousvalued attributes for classification learning*," Proc. of the 13th Int. Joint Conf. on Artificial Intelligence, pp. 1022--1027, Morgan Kaufmann, 1993.