UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 3

Information Theory & Preparing Data

Prof. Vibs Abhishek

The Paul Merage School of Business

University of California, Irvine

Agenda

- Information Theory
- Reminders
  - Assignment 1 posted on Canvas
  - Form groups for project
- Install Weka
- Working with Data

From Probability to Information
Information Theory

- Makes use of the probabilistic relationship between attributes to identify how much information one attribute provides on the other
  - Useful to understand relative importance of attributes
  - Can also be used to measure redundancy across attributes
- Information = surprise
  - How much surprise is created by a news
  - Information = expectation – realization

Logarithm

- $\log_b(X)$ is read as “log of *X* with base *b*”
  - Microsoft Excel : “=log(*X*,*b*)”
  - What does it mean
    - If $Y = \log_b(X)$, then $X = b^Y$
  - Base 10: log(*X*) = log10(*X*)
    - Microsoft Excel : “=log(*X*)”
    - If $Y = \log_{10}(X)$, then $X = 10^Y$
      - If $\log_{10}(1000) = 3$, and $1000 = 10^3$
  - Natural logarithm = ln(*X*) = $\log_e(X)$, where e=2.7183
    - Microsoft Excel : “=ln(*X*)”
  - Logarithm with base 2
    - Microsoft Excel : “=log(*X*,2)”
    - $\mathbf{\log_2(X)} = \log_{10}(X)/\log_{10}(2) = \mathbf{3.3219\ \log_{10}(X)}$

Information Theory

- Entropy of a distribution
  - Let *X* be a random variable with the probability distribution

Pr[X=x_i] = p_i, i=1,2,…,n where

- Entropy of *X* (level of disorder):

$$H(X) = H(p_1,p_2, \ldots, p_n) = \; -$$

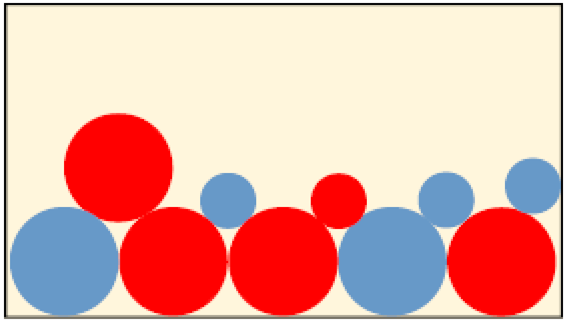
- Let *Y* be another random variable (jointly distributed)
  - Knowledge of *Y* reduces the uncertainty and hence entropy of *X*.
  - Therefore, *Y* provides the following information about *X*:
- $I\,(X;Y) = H(X\,) - H(X\,|Y)$.
  - Thus *I(X;Y)* is called Mutual Information

Properties of Information Measure

- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X)$

- If X and Y are independent
  - $H(Y|X) = H(Y)$
  - $I(X;Y) = 0$

Example



- Consider 10 balls in a basket
  - 4 large and red, 1 small and red, 2 large and blue, and 3 small and blue
  - You are to pick one randomly and predict its color without looking
- Strategy
  - Check the size of the ball and predict
    - Red if it is large (67% accurate — 4 out of 6)
    - Blue if it is small (75% accurate — 3 out of 4)
- Without the size information, you can only be 50% accurate
- Clearly, size provides information about the color
  - We know that since size and color are not independent
  - Color provides information about the size, as well

I(Color; Size)

- *I*(Color; Size) = *H*(Color) – *H*(Color | Size)
- Without size information:

  H(Color) = H() =1
- With size information:

  H(Color | Size = large) = *H*(,) = 0.918.

  H(Color | Size = small) = *H*(,) = 0.811.

  H(Color | Size) = *H*(Color | Size = large)× P(Size = large)

  + *H*(Color | Size = small) P(Size = small)

  =0.918 × 0.6 + 0.811 × 0.4 = 0.875
- Information gain = 1 – 0.875 = 0.125 bit
  - Size, on average, provides 0.125 bit of information on color

I(Size;Color)

- $I(\text{Size; Color}) = H(\text{Size}) - H(\text{Size} \mid \text{Color})$
- Without color information:

  H(Size) (0.6,0.4) =

- With color information:

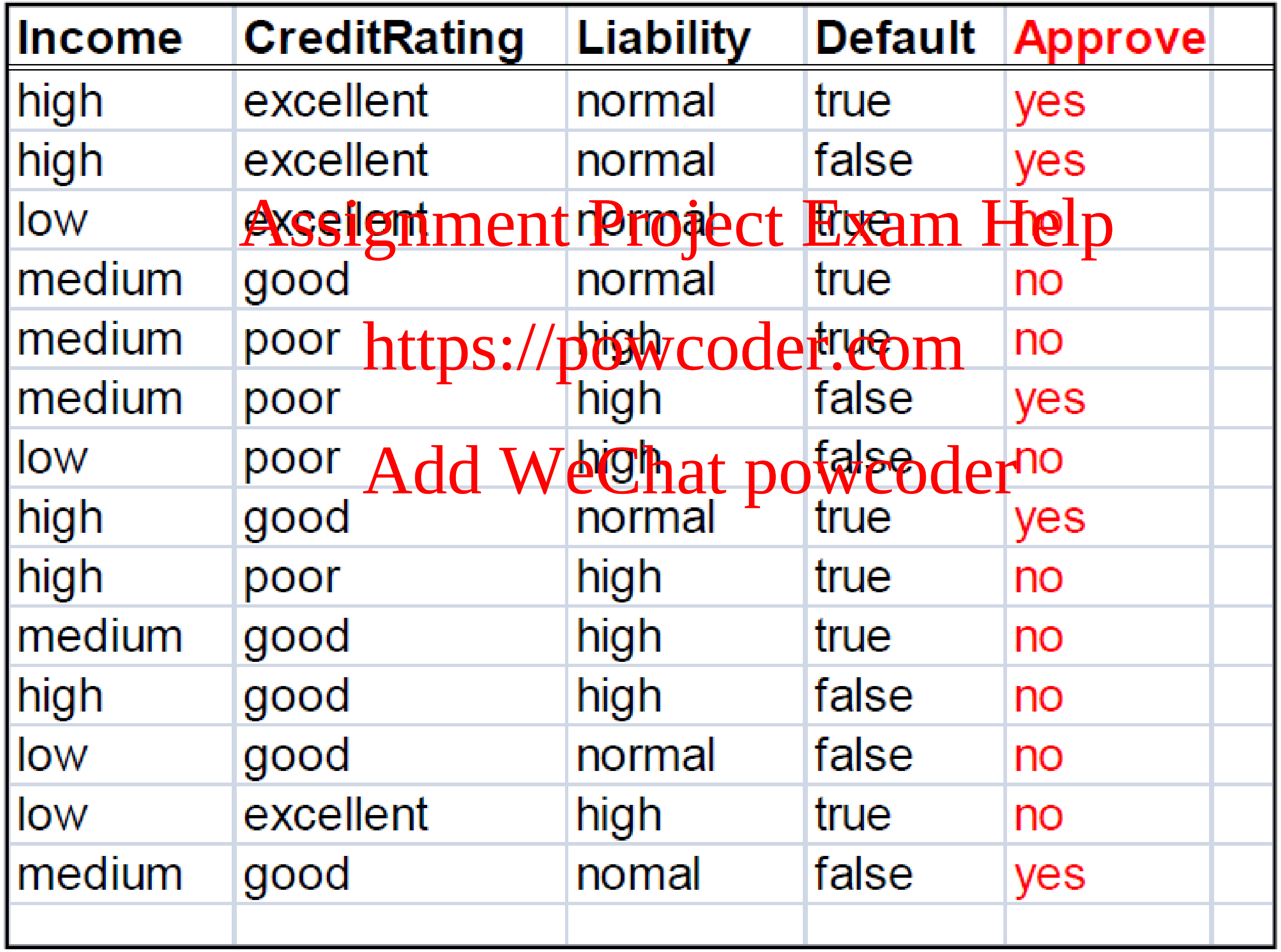
  $H(\text{Size} \mid \text{Color} = \text{red})$

  $H(\text{Size} \mid \text{Color} = \text{blue})$

  $H(\text{Size} \mid \text{Color}) = H(\text{Size} \mid \text{Color} = \text{red}) \times P(\text{Color} = \text{red}) +$
  $H(\text{Size} \mid \text{Color} = \text{blue}) \times P(\text{Color} = \text{blue})$

- Information gain =

Loan Application Data



Income	CreditRating	Liability	Default	Approve
high	excellent	normal	true	yes
high	excellent	normal	false	yes
low	excellent	normal	true	no
medium	good	normal	true	no
medium	poor	high	true	no
medium	poor	high	false	yes
low	poor	high	false	no
high	good	normal	true	yes
high	poor	high	true	no
medium	good	high	true	no
high	good	high	false	no
low	good	normal	false	no
low	excellent	high	true	no
medium	good	nomal	false	yes

Contingency Table

(Expressing relationship between two attributes)

Compute H(Liability)
& *H*(Liability | CR)
& *I* (Liability; CR)

		Liability		
		normal	high	**Total**
CreditRating	excellent	3	1	**4**
	good	4	2	**6**
	poor	0	4	**4**
	Total	**7**	**7**	**14**

$H(\text{Liability}) = H() = 1$

$H(\text{Liability} \mid \text{CR} = \text{excellent}) = H(\; 0.811$

$H(\text{Liability} \mid \text{CR} = \text{good}) = H(\; 0.918$

$H(\text{Liability} \mid \text{CR} = \text{poor}) = H(\; 0$

$H(\text{Liability} \mid \text{CR}) = 0.811 \times () + 0.918 \times () + 0 \times () = \; 0.625$

$I\,(\text{Liability}; \text{CR}) = 1 - 0.625 = 0.375$

$\Rightarrow I\,(\text{CR}; \text{Liability}) = 0.375.$

Entropy and Gain Ratio

- Even though the mutual information between two random variables is always symmetric, observing or recording one variable may be more difficult than the other
  - The more uncertainty about a variable, higher is the level of this difficulty
  - Entropy measures this difficulty

Gain Ratio

- Gain ratio (*G*) measures the information gain relative to the level of difficulty of coding the attribute
- $G(X; Y) = I(X; Y) / H(Y)$
- $G(Y; X) = I\,(Y; X) / H(X)$
- $G(X; Y) \neq G(Y; X)$

G(CR;Liability) & G(Liability;CR)

		Liability		
		normal	high	**Total**
CreditRating	excellent	3	1	**4**
	good	4	2	**6**
	poor	0	4	**4**
	Total	**7**	**7**	**14**

$I(CR; Liability) = I(Liability; CR) = 0.375$

$H(Liability) = H(\frac{1}{2}, \frac{1}{2}) = -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}) = 1$

$H(CR) = H(4/14, 6/14, 4/14)$

$G(CR; Liability) = I(Liability; CR) / H(Liability) = 0.375$

$G(Liability; CR) = I(Liability; CR) / H(CR) = 0.241$

$G(CR; Liability) \neq G(Liability; CR)$

2022.11.24

Preparing Data

Prof. Vibs Abhishek

The Paul Merage School of Business

University of California, Irvine

Steps in Data Mining

1. Develop an understanding of the purpose of the data mining project
2. Obtain the data set to be used in the analysis
   - random sampling from a large database to capture records
   - While data mining deals with very large databases
     - usually the analysis to be done requires only thousands or tens of thousands of records

3. Explore, clean, and preprocess the data
   - This involves verifying that the data are in reasonable condition.
   - How should missing data be handled?
   - Are the values in a reasonable range, given what you would expect for each variable?
   - Are there obvious “outliers?”
   - The data are reviewed graphically - for example, a matrix of scatterplots showing the relationship of each variable with each other variable

4. Reduce the data, if necessary
   - eliminate unneeded variables
   - transforming variables
   - creating new variables

Steps in Data Mining

5. Determine the data mining task
   - classification, prediction, clustering, etc.
6. Choose the data mining techniques to be used
   - regression, neural nets, hierarchical clustering, etc.
7. Use algorithms to perform the task
   - This is typically an iterative process
   - Choosing different variables or settings within the algorithm
8. Interpret the results of the algorithms
   - Recall that each algorithm may also be tested on the validation data for tuning purposes
     - validation data becomes a part of the fitting process!
     - likely to underestimate the error in the deployment of the model that is finally chosen
9. Deploy the model in real world
   - For example, the model might be applied to a purchased list of possible customers
   - action might be “include in the mailing if the predicted amount of purchase is > $10”

Data Types

- Variable Measures
  - Categorical variables (e.g., CA, AZ, UT…)
  - Ordered variables (e.g., course grades)
  - Numeric variables (e.g., money)
- Dates & Times
- Fixed-Length Character Strings (e.g., Zip Codes)
- IDs and Keys – used for linkage to other data in other tables
- Names (e.g., Company Names)
- Addresses
- Free Text (e.g., annotations, comments, memos, email)
- Unstructured Data (e.g., audio, images)

Nominal quantities

- Values are distinct symbols
  - Values themselves serve only as labels or names

- Example: attribute “outlook” from weather data

  - Values: “sunny”,”overcast”, and “rainy”

- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

Ordinal quantities

- Impose order on values
- But: no distance between values defined
- Example: attribute “temperature” in weather data
  - Values: “hot” > “mild” > “cool”
- Note: addition and subtraction don’t make sense
- Distinction between nominal and ordinal not always clear (e.g. attribute “outlook”)

The ARFF format

```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

Additional attribute types

- ARFF supports *string* attributes:

```
@attribute description string
```

  - Similar to nominal attributes but list of values is not pre-specified

- It also supports *date* attributes:

```
@attribute today date
```

  - Uses the ISO-8601 combined date and time format *yyyy-MM-dd-THH:mm:ss*

Sparse data

- In some applications most attribute values in a dataset are zero
    - E.g.: word counts in a text categorization problem
- ARFF supports sparse data

```
0, 26, 0,  0, 0 ,0, 63, 0, 0, 0, “class A”
0,  0, 0, 42, 0, 0,  0, 0, 0, 0, “class B”
```

```
{1 26, 6 63, 10 “class A”}
{3 42, 10 “class B”}
```

- More details about ARFF:
    - http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Sampling Data

- Sampling can be used to create better data sets (training or testing) to build better models.
- Random sampling techniques:
  - **Simple random sampling**
  - **Proportionate stratified sampling**: Select a representative sample. Divide data set into strata: e.g. ‘Business’ travelers and ‘Private’ travelers. Assuming the proportions were 90% Business and 10% Private, and we needed at least 100 Private travelers for our model, we would randomly select 900 Business travelers.
  - **Disproportionate stratified sampling**: Select a weighted sample. Also called ‘oversampling’: used if a particular group of examples is important but not well represented in the data set.
    e.g. In direct mail response prediction you might select 10 responders in the dataset for every non-responder you select. For claims analysis, you might weigh the fraudulent claims (which are often naturally rare).

Missing Value Treatment

- **Reason for missing?**
  - Not recorded
  - Not applicable
  - Customer refused to provide

- **Dealing with missing values:**
  - Delete the records with missing values
  - Add flag fields (‘address_missing’=true) to indicate missing values, or
  - Estimate missing value:
    - Use average over entire data set
    - Use average over similar records
    - Use an advanced prediction technique

Noise in Data

- The biggest challenge with noisy data sets is that it is difficult to identify noise
- In some special cases, noise can be identified
  - Value out of range (e.g., negative age)
  - Meaningless value (e.g., License# for someone without a license)
  - Mismatched value (e.g., City, State, and PIN not matching against the postal database)

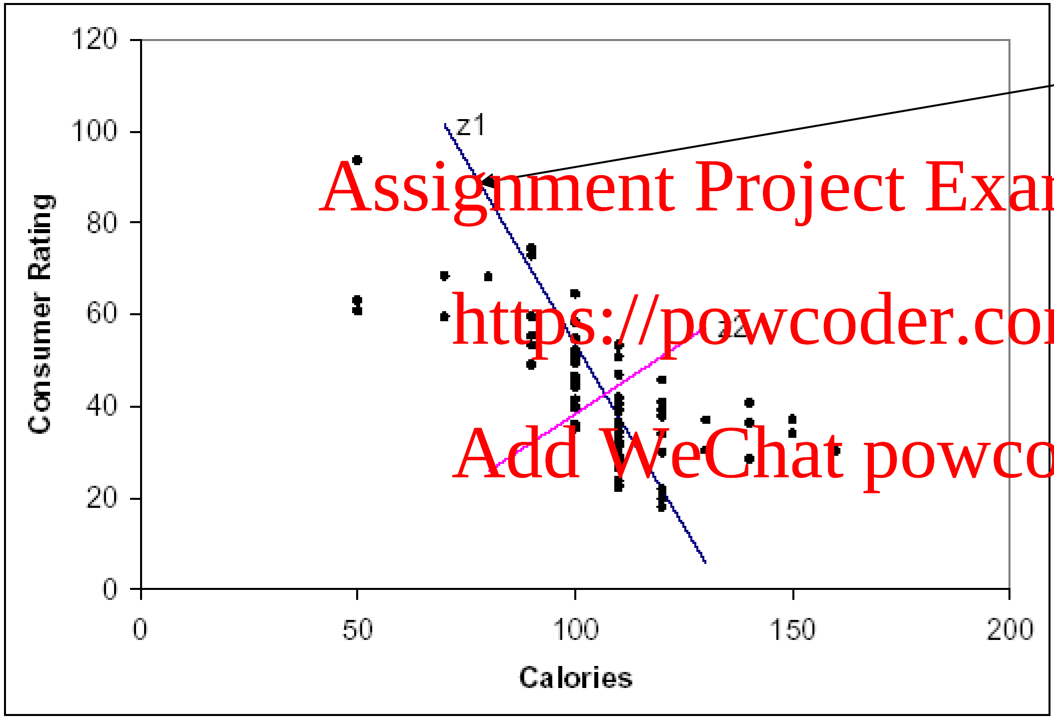
Attribute Selection

- Smaller attribute sets are simpler to understand, but may produce an overly simplistic model
- Larger attribute sets may lead to overfitting
- Eliminate useless attributes
  - Related to redundancy and feature selection
- Attribute consolidation
  - Combine a set of binary attributes into one nominal attribute
- Attribute expansion
  - Expand a nominal attribute into a set of binary ones
- Attribute conversion
  - Change the data type of an attribute

Formal Dimension Reduction

- If you have multiple highly correlated columns, then reduce number of columns
  - e.g. height in inches and cm
- Principal components analysis (PCA)
  - Subsets of numeric (not categorical) variables
    - measured on the same scale
    - highly correlated
  - Come up with few variables (one or two or three)
    - that are weighted linear combinations of original variables
    - retain the explanatory power of the original data

Principal Components

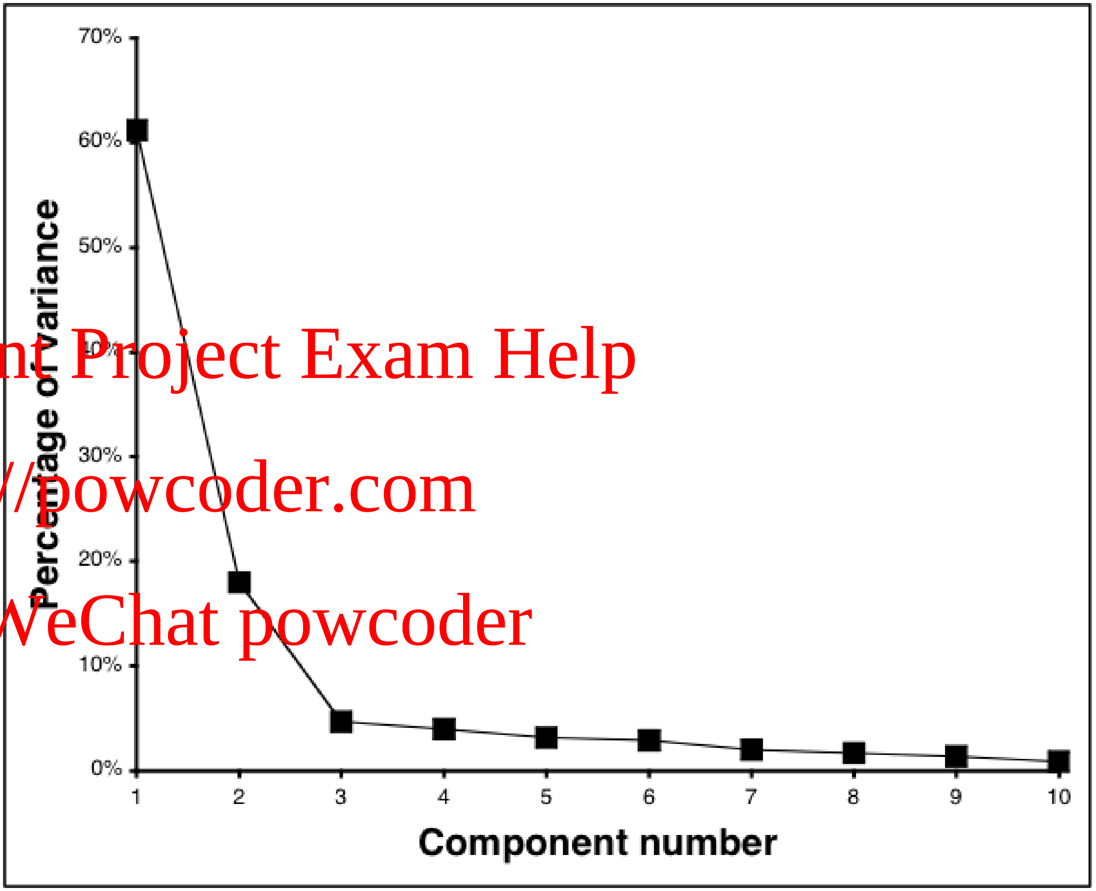


The line z1 is the direction in which the variability of the points is largest.

1st principal component

Example: 10-dimensional data

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%



Attribute Consolidation

- Example 1: Suppose you have two 0/1 attributes: “Male” and “Female”
  - A row of data cannot have 1 for both the attributes
- At the same time, both cannot be 0
  - Create a new nominal attribute: “Gender” with two possible values — *male and female*

Attribute Expansion

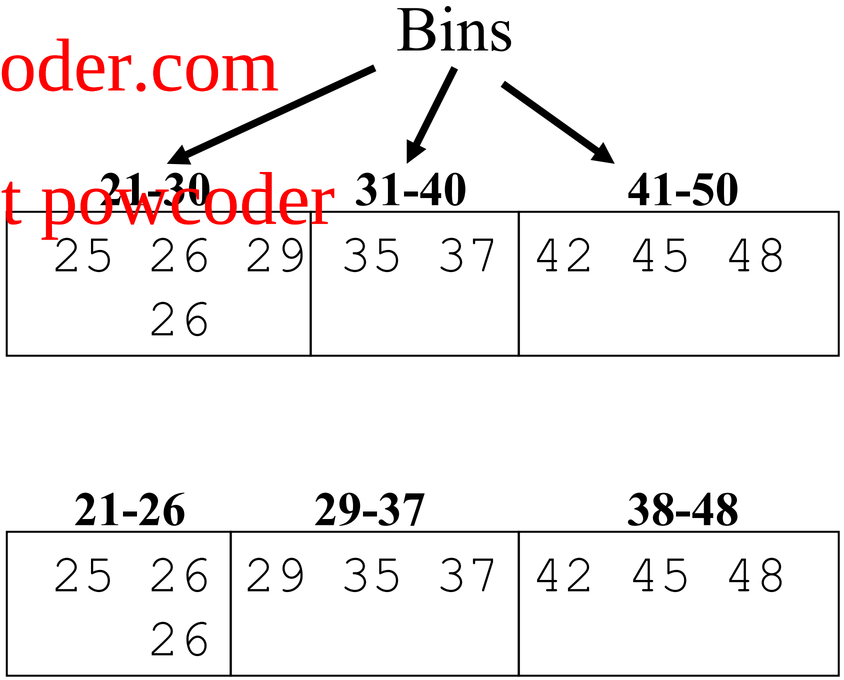
- Attribute expansion is the opposite of attribute consolidation
  - A nominal attribute is converted to a set of 0/1 attribute
- Set-values attributes
  - Example: Hobby, Genre, Interest
- It can be replaced by a set of binary attributes

Attribute Conversion

- Ratios
  - e.g. Try income divided by number of employees, to get a measure of productivity per employee
- Derived Values
  - e.g. derive customer (or product) *age* from *birthdate* (or *production-date*), as age may be more predictive.
- Changing the data type of attributes
  - Nominal to numeric or vice versa

Binning

- Binning (discretization) converts numeric values to discrete categories. e.g. low-income is <= 30, high-income is > 30
- For example:
  - □ Equal-Interval binning
    - Bin intervals of equal width, irrespective of number of items per bin
  - □ Equal-Frequency binning
    - Equal number of items per bin, irrespective of bin width



21-30	31-40	41-50
25 26 29 26	35 37	42 45 48

21-26	29-37	38-48
25 26 26	29 35 37	42 45 48

The entropy of a random variable is higher when

A: It has many different states, each of which has low likelihood

B: It has very few states, each of which has high likelihood

C: It has many different states, where only a few states have very high likelihood

D: It has very few states

E: None of the above

The file format used by WEKA is called

A. DOCX
B. XCL
C. WEK
D. ARFF
E. TXT

When to Normalize Data?

- Rescale attributes to the range of 0 to 1
  - Subtract the min, and divide by (max – min)
- Results in all variables getting equal importance
- Not advisable
  - When the units of measurement are common for the variables (e.g. dollars), and when their scale reflects their importance
    - e.g. sales of jet fuel, sales of heating oil
- Advisable
  - if the variables are measured in quite differing units
    - unclear how to compare the variability of different variables
    - e.g. dollars for some, parts per million for others
  - or if variables measured in the same units, but scale does not reflect importance
    - e.g. earnings per share, gross revenues

Data Preprocessing using Weka

- Download file 4bank-data.csv from Canvas
- Follow steps on the following page:
- http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/preprocess.html

RFM, Pivot Tables and London Jets Data

- http://www.dbmarketing.com/articles/Art149.htm
- London Jets Data in Excel format posted on Canvas for RFM analysis and Pivot tables.
  - Do RFM analysis on this data
  - Think about strategies that London Jets could use to revive their fortunes
- Go to http://office.microsoft.com/en-us/
  - Search for "Pivot Table" and read up on creating and using them

Next Session

- Classification using Exact Bayes & Naïve Bayes