# 1

# Introduction

1. [Fall 2008]

   For each data set given below, give specific examples of classification, clustering, association rule mining, and anomaly detection tasks that can be performed on the data. For each task, state how the data matrix should be constructed (i.e., specify the rows and columns of the matrix).

   (a) Ambulatory Medical Care data[1], which contains the demographic and medical visit information for each patient (e.g., gender, age, duration of visit, physician's diagnosis, symptoms, medication, etc).

   **Answer:**

| Classification |
|---|
| Task: Diagnose whether a patient has a disease.<br>Row: Patient<br>Column: Patient's demographic and hospital visit information (e.g., symptoms), along with a class attribute that indicates whether the patient has the disease. |
| **Clustering** |
| Task: Find groups of patients with similar medical conditions<br>Row: A patient visit<br>Column: List of medical conditions of each patient |
| **Association rule mining** |
| Task: Identify the symptoms and medical conditions that co-occur together frequently<br>Row: A patient visit<br>Column: List of symptoms and diagnosed medical conditions of the patient |
| **Anomaly detection** |
| Task: Identify healthy looking patients with rare medical disorders<br>Row: A patient visit<br>Column: List of demographic attributes, symptoms, and medical test results of the patient |

[1] See for example, the National Hospital Ambulatory Medical Care Survey `http://www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm`

(b) Stock market data, which include the prices and volumes of various stocks on different trading days.

**Answer:**

| Classification |
|---|
| Task: Predict whether the stock price will go up or down the next trading day<br>Row: A trading day<br>Column: Trading volume and closing price of the stock the previous 5 days and a class attribute that indicates whether the stock went up or down |
| **Clustering** |
| Task: Identify groups of stocks with similar price fluctuations<br>Row: A company's stock<br>Column: Changes in the daily closing price of the stock over the past ten years |
| **Association rule mining** |
| Task: Identify stocks with similar fluctuation patterns(e.g., {Google-Up, Yahoo-Up})<br>Row: A trading day<br>Column: List of all stock-up and stock-down events on the given day. |
| **Anomaly detection** |
| Task: Identify unusual trading days for a given stock (e.g., unusually high volume)<br>Row: A trading day<br>Column: Trading volume, change in daily stock price (daily high − low prices), and average price change of its competitor stocks |

(c) Database of Major League Baseball (MLB).

| Classification |
|---|
| Task: Predict the winner of a game between two MLB teams.<br>Row: A game.<br>Column: Statistics of the home and visiting teams over their past 10 games they had played (e.g., average winning percentage and hitting percentage of their players) |
| **Clustering** |
| Task: Identify groups of players with similar statistics<br>Row: A player<br>Column: Statistics of the player |
| **Association rule mining** |
| Task: Identify interesting player statistics (e.g., 40% of right-handed players have a batting percentage below 20% when facing left-handed pitchers)<br>Row: A player<br>Column: Discretized statistics of the player |
| **Anomaly detection** |
| Task: Identify players who performed considerably better than expected in a given season<br>Row: A (player,season) pair e.g, (player1 in 2007)<br>Column: Ratio statistics of a player (e.g., ratio of average batting percentage in 2007 to career average batting percentage) |

# 2

# Data

## 2.1 Types of Attributes

1. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

   (a) Number of courses registered by a student in a given semester.
   **Answer:** Discrete, quantitative, ratio.

   (b) Speed of a car (in miles per hour).
   **Answer:** Discrete, quantitative, ratio.

   (c) Decibel as a measure of sound intensity.
   **Answer:** Continuous, quantitative, interval or ratio. It is actually a logratio type (which is somewhere between interval and ratio).

   (d) Hurricane intensity according to the Saffir-Simpson Hurricane Scale.
   **Answer:** Discrete, qualitative, ordinal.

   (e) Social security number.
   **Answer:** Discrete, qualitative, nominal.

2. Classify the following attributes as:

   - discrete or continuous.
   - qualitative or quantitative
   - nominal, ordinal, interval, or ratio

Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

(a) Julian Date, which is the number of days elapsed since 12 noon Greenwich Mean Time of January 1, 4713 BC.
**Answer:** Continuous, quantitative, interval

(b) Movie ratings provided by users (1-star, 2-star, 3-star, or 4-star).
**Answer:** Discrete, qualitative, ordinal

(c) Mood level of a blogger (cheerful, calm, relaxed, bored, sad, angry or frustrated).
**Answer:** Discrete, qualitative, nominal

(d) Average number of hours a user spent on the Internet in a week.
**Answer:** Continuous, quantitative, ratio

(e) IP address of a machine.
**Answer:** Discrete, qualitative, nominal

(f) Richter scale (in terms of energy release during an earthquake).
**Answer:** Continuous, qualitative, ordinal
In terms of energy release, the difference between 0.0 and 1.0 is not the same as between 1.0 and 2.0. Ordinal attributes are qualitative; yet, can be continuous.

(g) Salary above the median salary of all employees in an organization.
**Answer:** Continuous, quantitative, interval

(h) Undergraduate level (freshman, sophomore, junior, and senior) for measuring years in college.
**Answer:** Discrete, qualitative, ordinal

3. For each attribute given, classify its type as:

- discrete or continuous AND
- qualitative or quantitative AND
- nominal, ordinal, interval, or ratio

Indicate your reasoning if you think there may be some ambiguity in some cases.

**Example:** Age in years.

**Answer:** Discrete, quantitative, ratio.

(a) Daily user traffic volume at YouTube.com (i.e., number of daily visitors who visited the Web site).
**Answer:** Discrete, quantitative, ratio.

(b) Air pressure of a car/bicycle tire (in psi).
**Answer:** Continuous, quantitative, ratio.

(c) Homeland Security Advisory System ratings - code red/orange/etc.
**Answer:** Discrete, qualitative, ordinal.

(d) Amount of seismic energy release, measured in Richter scale.
**Answer:** Continuous, qualitative, ordinal.

(e) Credit card number.
**Answer:** Discrete, qualitative, nominal.

(f) The wealth of a nation measured in terms of gross domestic product (GDP) per capita above the world's average of $10,500.
**Answer:** Continuous, quantitative, interval.

4. For each attribute given, classify its type as:

- discrete or continuous AND
- qualitative or quantitative AND
- nominal, ordinal, interval, or ratio

Indicate your reasoning if you think there may be some ambiguity in some cases.

**Example:** Age in years.

**Answer:** Discrete, quantitative, ratio.

(a) Favorite movie of each person.
**Answer:** Discrete, qualitative, nominal

(b) Number of days since Jan 1, 2011.
**Answer:** Discrete, quantitative, interval.

(c) Category of a hurricane (The Saffir-Simpson Hurricane Wind Scale ranges from category 1 to category 5).
**Answer:** Discrete, qualitative, ordinal.

(d) Number of students enrolled in a class.
**Answer:** Discrete, quantitative, ratio

5. For each attribute given, classify its type as:

   - discrete or continuous AND
   - qualitative or quantitative AND
   - nominal, ordinal, interval, or ratio

   Indicate your reasoning if you think there may be some ambiguity in some cases.

   **Example:** Temperature in Kelvin

   **Answer:** Continuous, quantitative, ratio.

   (a) Number of years since 1 BC. For example, 2 BC is year -1, 1 BC is year 0, 1 AD is year 1, and 2013 AD is year 2013 (note, there is no 0 AD in Gregorian calendar).

   **Answer:** Discrete/Continuous, quantitative, interval.

   (b) GPA of a student.

   **Answer:** Continuous, qualitative, ordinal.

   (c) Mood level of a blogger (cheerful, calm, relaxed, bored, sad, angry or frustrated).

   **Answer:** Discrete, qualitative, nominal.

   (d) Sound intensity in decibel scale.

   **Answer:** Continuous, qualitative, ordinal. In terms of sound intensity, the difference between 0dB and 1dB is not the same as the difference between 10 dB and 11 dB (decibels are in log scale); thus, it is not an interval attribute.

6. State the type of each attribute given below before and after we have performed the following transformation.

   (a) Hair color of a person is mapped to the following values: black = 0, brown = 1, red = 2, blonde = 3, grey = 4, white = 5.

   **Answer:** Nominal (both before and after transformation).

   (b) Grade of a student (from 0 to 100) is mapped to the following scale: A = 4.0, A- = 3.5, B = 3.0, B- = 2.5, C = 2.0, C- = 1.5, D = 1.0, D- = 0.5, E = 0.0

   **Answer:** Ratio (before transformation) to ordinal (after transformation).

(c) Age of a person is discretized to the following scale: Age $<$ 12, 12 $\leq$ Age $<$ 21, 21 $\leq$ Age $<$ 45, 45 $\leq$ Age $<$ 65, Age $>$ 65.

**Answer:** Ratio (before transformation) to ordinal (after transformation)

(d) Annual income of a person is discretized to the following scale: Income $<$ \$20K, \$20K $\leq$ Income $<$ \$60K, \$60K $\leq$ Income $<$ \$120K, \$120K $\leq$ Age $<$ \$250K, Age $\geq$ \$250K.

**Answer:** Ratio (before transformation) to ordinal (after transformation).

(e) Height of a person is changed from meters to feet.

**Answer:** Ratio (both before and after transformation)

(f) Height of a person is changed from meters to {Short, Medium, Tall}.

**Answer:** Ratio (before transformation) to ordinal (after transformation).

(g) Height of a person is changed from feet to number of inches above 4 feet.

**Answer:** Ratio (before transformation) to interval (after transformation).

(h) Weight of a person is standardized by subtracting it with the mean of the weight for all people and dividing by its standard deviation.

**Answer:** Ratio (before transformation) to interval (after transformation)

7. State whether it is meaningful (based on the properties of the attribute values) to apply the following operations to the data given below

(a) Average amplitude of seismic waves (in Richter scale) for the 10 deadliest earthquakes in Asia.

**Answer:** No because Richter scale is ordinal.

(b) Average number of characters in a collection of spam messages.

**Answer:** Yes because number of characters is a ratio attribute.

(c) Pearson's correlation between shirt size and height of an individual.

**Answer:** No because shirt size is ordinal.

(d) Median zipcode of households in the United States.

**Answer:** No because zipcode is nominal.

(e) Entropy of students (based on the GPA they obtained for a given course).

**Answer:** Yes because entropy is applicable to nominal attributes.

(f) Geometric mean of temperature (in Fahrenheit) for a given city.

**Answer:** No because temperature (in Fahrenheit) is not a ratio attribute.

## 2.2 Data Preprocessing

1. Consider the following dataset that contains the age and gender information for 9 users who visited a given website.

| UserID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Age | 17 | 24 | 25 | 28 | 32 | 38 | 39 | 49 | 68 |
| Gender | Female | Male | Male | Male | Female | Female | Female | Male | Male |

(a) Suppose you apply equal interval width approach to discretize the Age attribute into 3 bins. Show the userIDs assigned to each of the 3 bins.

**Answer:** Bin width $= \frac{68-17}{3} = \frac{51}{3} = 17$.

Bin 1: 1, 2, 3, 4, 5

Bin 2: 6, 7, 8

Bin 3: 9

(b) Repeat the previous question using the equal frequency approach.

**Answer:** Since there are 9 users and 3 bins, every bin must contain 3 users.

Bin 1: 1, 2, 3

Bin 2: 4, 5, 6

Bin 3: 7, 8, 9

(c) Repeat question (a) using a supervised discretization approach (with Gender as class attribute). Specifically, choose the bins in such a way that their members are as "pure" as possible (i.e., belonging to the same class).

**Answer:**

Bin 1: 1, 2, 3, 4

Bin 2: 5, 6, 7

Bin 3: 8, 9

2. Consider an attribute X of a data set that takes the values $\{x_1, x_2, \cdots, x_9\}$ (sorted in increasing order of magnitude). We apply two methods (equal interval width and equal frequency) to discretize the attribute into 3 bins. The bins obtained are shown below:

   Equal Width: $\{x_1, x_2, x_3\}$, $\{x_4, x_5, x_6, x_7, x_8\}$, $\{x_9\}$

   Equal Frequency: $\{x_1, x_2, x_3\}$, $\{x_4, x_5, x_6\}$, $\{x_7, x_8, x_9\}$

   Explain what will be the effect of applying the following transformations on each discretization method, i.e., whether the elements assigned to each bin can change if you discretize the attribute **after** applying the transformation function below. Note that $\bar{X}$ denotes the average value and $\sigma_x$ denotes standard deviation of attribute X.

   (a) $\mathrm{X} \rightarrow \mathrm{X} - \bar{X}$ (i.e., if the attribute values are centered).

   **Answer:** No change for equal width because the distance between $x_i$ and $x_{i+1}$ is unchanged. No change for equal frequency because the relative ordering of data points remain the same (i.e., if $x_i < x_{i+1}$ then $x_i - \bar{X} < x_{i+1} - \bar{X}$).

   (b) $\mathrm{X} \rightarrow \frac{X-\bar{X}}{\sigma_x}$ (i.e., if the attribute values are standardized).

   **Answer:** Since the distances between every pair of points $(x_i, x_{i+1})$ change uniformly (by a constant factor of $\sigma_x$, the elements in the bins are unchanged for equal width discretization. No change for equal frequency because the relative ordering of data points remain the same.

   (c) $\mathrm{X} \rightarrow \exp\left[\frac{X-\bar{X}}{\sigma_x}\right]$ (i.e., if the values are standardized and exponentiated).

   **Answer:** The bin elements may change for equal width because the distances between $x_i$ and $x_{i+1}$ may not change uniformly. No change for equal frequency because the relative ordering of data points remain the same.

3. Consider a dataset that has 3 attributes ($x_1$, $x_2$, and $x_3$). The distribution of each attribute is as follows and shown in Figure

   - $x_1$ has a uniform distribution in the range between 0 and 1.
   - $x_2$ is generated from a mixture of 3 Gaussian distributions centered at 0.1, 0.5, and 0.9, respectively. The standard deviation of the

distributions are 0.02, 0.1, and 0.02, respectively. Assume each point is generated from one of the 3 distributions and the number of points associated with each distribution is different.

- $x_3$ is generated from an exponential distribution with mean 0.1.

(a) Which attribute(s) is likely to produce the same bins regardless of whether you use equal width or equal frequency approaches (assuming the number of bins is not too large).

**Answer:** $x_1$.

(b) Which attribute(s) is more suitable for equal frequency than equal width discretization approaches.

**Answer:** $x_3$.

(c) Which attribute(s) is not appropriate for both equal width and equal frequency discretization approaches.

**Answer:** $x_2$.

(d) If all 3 are initially ratio attributes, what are their attribute types after discretization?

**Answer:** Ordinal.

4. An e-commerce company is interested in identifying the highest spending customers at its online store using association rule mining. One of the rules identified is:

$$21 \leq \text{Age} < 45 \text{ AND NumberOfVisits} > 50 \rightarrow \text{AmountSpent} > \$500,$$

where the Age attribute was discretized into 5 bins, NumberOfVisits was discretized into 8 bins, and AmountSpent was discretized into 8 bins. The confidence of an association rule $A, B \rightarrow C$ is defined as

$$\text{Confidence}(A, B \rightarrow C) = P(C|A, B) = \frac{P(A, B, C)}{P(A, B)} \qquad (2.1)$$

where $P(C|A, B)$ is the conditional probability of $C$ given $A$ and $B$, $P(A, B, C)$ is the joint probability of $A$, $B$, and $C$, and $P(A, B)$ is the joint probability of $A$ and $B$. The probabilities are empirically estimated based on their relative frequencies in the data. For example, $P(\text{AmountSpent} > \$500)$ is given by the proportion of online users who visited the store and spent more than \$500.

(a) Suppose we increase the number of bins for the Age attribute from 5 to 6 so that the discretized Age in the rule becomes $21 \leq$ Age $< 30$ instead of $21 \leq$ Age $< 45$, will the confidence of the rule be non-increasing, non-decreasing, stays the same, or could go either way (increase/decrease)?

**Answer:** Can increase/decrease.

(b) Suppose we increase the number of bins for the AmountSpent attribute from 8 to 10, so that the right hand side of the rule becomes \$500 $<$ AmountSpent $<$ \$1000, will the confidence of the rule be non-increasing, non-decreasing, stays the same, or could go either way (increase/decrease)?

**Answer:** Non-increasing.

(c) Suppose the values for NumberOfVisits attribute are distributed according to a Poisson distribution with a mean value equals to 4. If we discretize the attribute into 4 bins using the equal frequency approach, what are the bin values after discretization? Hint: you need to refer to the cumulative distribution table for Poisson distribution to answer the question.

**Answer:** Choose the bin values such that the cumulative distribution is close to 0.25, 0.5, and 0.75. This corresponds to bin values: 0 to 2, 3, 4 to 5, and greater than 5.

5. Null values in data records may refer to missing or inapplicable values. Consider the following table of employees for a hypothetical organization:

| Name | Sales commission | Occupation |
|---|---|---|
| John | 5000 | Sales |
| Mary | 1000 | Sales |
| Bob | null | Non-sales |
| Lisa | null | Non-sales |

The null values in the table refer to inapplicable values since sales commission are calculated for sales employees only. Suppose we are interested to calculate the similarity between users based on their sales commission.

(a) Explain what is the limitation of the approach to compute similarity if we replace the null values in sales commission by 0.

**Answer:** Mary will be more similar to Bob and Lisa than to John.

(b) Explain what is the limitation of the approach to compute similarity if we replace the null values in sales commission by the average value of sales commission (i.e., 3000).

**Answer:** Both Mary and John are less similar to each other than to Bob and Lisa.

(c) Propose a method that can handle null values in the sales commission so that employees that have the same occupation are closer to each other than to employees that have different occupations.

**Answer:** One way is to change the similarity function as follows:

$$\text{Similarity}(a,b) = \begin{cases} \infty, & \text{if both a and b are null;} \\ 0, & \text{if one of a or b is null;} \\ s(a,b), & \text{otherwise.} \end{cases}$$

where $s(a,b)$ is the original similarity measure used for the sales commission.

6. Consider a data set from an online social media Web site that contains information about the age and number of friends for 5,000 users.

(a) Suppose the number of friends for each user is known. However, only 4000 out of 5000 users provide their age information. The average age of the 4,000 users is 30 years old. If you replace the missing values for age with the value 30, will the average age computed for the 5,000 users increases, decreases, or stays the same (as 30)?

**Answer:** Average age does not change.

$$\begin{aligned} \overline{x}_{\text{old}} &= \frac{1}{4000}\sum_{i=1}^{4000} x_i \\ \overline{x}_{\text{new}} &= \frac{1}{5000}\sum_{i=1}^{5000} x_i = \frac{1}{5000}\left[\sum_{i=1}^{4000} x_i + \sum_{i=4001}^{5000} x_i\right] \end{aligned}$$

Since $x_i = \overline{x}_{\text{old}}$ for $i = 4001, 4002, \cdots, 5000$ and $\sum_{i=1}^{4000} x_i = 4000\overline{x}_{\text{old}}$, we have

$$\overline{x}_{\text{new}} = \frac{1}{5000}\left[4000\overline{x}_{\text{old}} + 1000\overline{x}_{\text{old}}\right] = \overline{x}_{\text{old}}$$

(b) Suppose the covariance between age and number of friends calculated using the 4,000 users (with no missing values) is 20. If you replace the missing values for age with the average age of the 4,000 users, would the covariance between age and number of friends increases, decreases, or stays the same (as 20)? Assume that the average number of followers for all 5,000 users is the same as the average for 4,000 users.

**Answer:** Covariance will decrease. Let $C_1 = \sum_{i=1}^{4000}(x_i - \overline{x})(y_i - \overline{y})/3999$ be the covariance computed using the 4,000 users without missing values. If we impute the missing values for age with average age, $\overline{x}$ remains unchanged according to part (a). Furthermore, $\overline{y}$ is assumed to be unchanged. Thus, the new covariance is

$$\begin{aligned} C_2 &= \frac{1}{4999}\sum_{i=1}^{5000}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{1}{4999}\left[\sum_{i=1}^{4000}(x_i - \overline{x})(y_i - \overline{y}) + \sum_{i=4001}^{5000}(x_i - \overline{x})(y_i - \overline{y})\right] \\ &= \frac{1}{4999}\left[\sum_{i=1}^{4000}(x_i - \overline{x})(y_i - \overline{y}) + \sum_{i=4001}^{5000}(\overline{x} - \overline{x})(y_i - \overline{y})\right] \\ &= \frac{1}{4999}\sum_{i=1}^{4000}(x_i - \overline{x})(y_i - \overline{y}) < C_1 \end{aligned} \tag{2.2}$$

7. Consider the following data matrix on the right, in which two of its values are missing (the matrix on the left shows its true values).

$$\begin{bmatrix} -0.2326 & 0.2270 \\ -0.0847 & 0.7125 \\ 0.1275 & 0.3902 \\ 0.1329 & -0.1461 \\ 0.3724 & 0.1756 \\ 0.4975 & 0.8536 \\ 0.6926 & 0.7834 \\ 0.7933 & 0.7375 \\ 0.8229 & 0.2147 \\ 0.8497 & 0.4980 \\ 1.0592 & 0.7600 \\ 1.5028 & 1.0122 \end{bmatrix} \longrightarrow \begin{bmatrix} -0.2326 & 0.2270 \\ -0.0847 & 0.7125 \\ 0.1275 & 0.3902 \\ ? & -0.1461 \\ 0.3724 & 0.1756 \\ 0.4975 & 0.8536 \\ 0.6926 & 0.7834 \\ 0.7933 & 0.7375 \\ 0.8229 & 0.2147 \\ 0.8497 & 0.4980 \\ 1.0592 & ? \\ 1.5028 & 1.0122 \end{bmatrix}$$

(a) Impute the missing values for the matrix on the right by their respective column averages. Show the imputed values and calculate their root-mean-square-error (RMSE).

$$\text{RMSE} = \sqrt{\frac{(\mathbf{A}_{4,1} - \tilde{\mathbf{A}}_{4,1})^2 + (\mathbf{A}_{11,2} - \tilde{\mathbf{A}}_{11,2})^2}{2}}$$

where $\mathbf{A}_{i,j}$ denotes the true value of the $(i, j)$-th element of the data matrix and $\tilde{\mathbf{A}}_{i,j}$ denotes its corresponding imputed value.

**Answer:** The column averages are [0.5819 0.4962]. The imputed values are

$$\begin{bmatrix} -0.2326 & 0.2270 \\ -0.0847 & 0.7125 \\ 0.1275 & 0.3902 \\ 0.5819 & -0.1461 \\ 0.3724 & 0.1756 \\ 0.4975 & 0.8536 \\ 0.6926 & 0.7834 \\ 0.7933 & 0.7375 \\ 0.8229 & 0.2147 \\ 0.8497 & 0.4980 \\ 1.0592 & 0.4962 \\ 1.5028 & 1.0122 \end{bmatrix}$$

and the RMSE value is

$$\text{RMSE} = \sqrt{\frac{(0.1329 - 0.5819)^2 + (0.7600 - 0.4962)^2}{2}} = 0.3683$$

(b) The Expectation-Maximization (E-M) algorithm is a well-known approach for imputing missing values. Assuming the data is generated from a multivariate Gaussian distribution, E-M iteratively computes the following conditional mean for each attribute and uses it to impute the missing values:

$$\mu_{i|j} = \hat{\mu}_i + \mathbf{\Sigma}_{ij}\mathbf{\Sigma}_{jj}^{-1}(\mathbf{x}_j - \hat{\mu}_j)$$

where the indices $i, j \in \{1, 2\}$ refer to one of the two attributes of the data and $\Sigma^{-1}$ denote inverse of the covariance matrix. Repeat the previous question by applying the E-M algorithm iteratively for

5 times. Assume the covariance matrix of the data is known and given by

$$\mathbf{\Sigma} = \begin{bmatrix} 0.25 & 0.1 \\ 0.1 & 0.15 \end{bmatrix}$$

In the first iteration, compute the mean value for each column using only the non-missing values. In subsequent iterations, compute the mean value for each column using both the non-missing and imputed values. Show the imputed values after each iteration and compute the root-mean-square-error. Compare the error against the answer in part (a).

**Answer:**

The inverse of the covariance matrix is

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 5.4545 & -3.6364 \\ -3.6364 & 9.0909 \end{bmatrix}$$

The results after each iteration are shown below:

| Iteration | $\hat{\mu}_1$ | $\hat{\mu}_2$ | Imputed $x_{4,1}$ | Imputed $x_{11,2}$ | RMSE |
|---|---|---|---|---|---|
| 1 | 0.5819 | 0.4962 | 0.2315 | 0.9301 | 0.1390 |
| 2 | 0.5527 | 0.5324 | 0.1826 | 0.9928 | 0.1683 |
| 3 | 0.5486 | 0.5376 | 0.1756 | 1.0018 | 0.1736 |
| 4 | 0.5480 | 0.5384 | 0.1746 | 1.0030 | 0.1743 |
| 5 | 0.5479 | 0.5385 | 0.1745 | 1.0032 | 0.1745 |

The root-mean-square-error for EM algorithm is considerably lower than that using mean imputation.

8. The purpose of this exercise is to illustrate the relationship between PCA and SVD. Let $\mathbf{A}$ be an $N \times d$ rectangular data matrix and $\mathbf{C}$ be its $d \times d$ covariance matrix.

   (a) Suppose $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an $N \times N$ matrix whose elements are equal to 1, i.e., $\forall i, j : (\mathbf{1})_{ij} = 1$. Show that the covariance matrix $\mathbf{C}$ can be expressed into the following form:

$$\mathbf{C} = \frac{1}{N-1}\mathbf{A}^T\left[\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right]\mathbf{A}$$

.

**Answer:** The covariance between columns $i$ and $j$ in matrix $\mathbf{A}$ is given by

$$C_{ij} = \frac{\sum_k (A_{ki} - \overline{A}_i)(A_{kj} - \overline{A}_j)}{N-1}, \tag{2.3}$$

where $\overline{A}_i$ and $\overline{A}_j$ are their corresponding column averages. A matrix of column averages for $\mathbf{A}$ can be computed as follows:

$$\begin{aligned}
\frac{1}{N}\mathbf{1}_N\mathbf{A} &= \frac{1}{N}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1d} \\ A_{21} & A_{22} & \cdots & A_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ A_{N1} & A_{N2} & \cdots & A_{Nd} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{N}\sum_j A_{j1} & \frac{1}{N}\sum_j A_{j2} & \cdots & \frac{1}{N}\sum_j A_{jd} \\ \frac{1}{N}\sum_j A_{j1} & \frac{1}{N}\sum_j A_{j2} & \cdots & \frac{1}{N}\sum_j A_{jd} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{N}\sum_j A_{j1} & \frac{1}{N}\sum_j A_{j2} & \cdots & \frac{1}{N}\sum_j A_{jd} \end{pmatrix}
\end{aligned} \tag{2.4}$$

Thus, each term $(A_{ki} - \overline{A}_i)$ in Equation (2.3) can be expressed in matrix notation as $A_{ki} - \frac{1}{N}\sum_j A_{ji} = [\mathbf{A} - \frac{1}{N}\mathbf{1}_N\mathbf{A}]_{ki}$. The covariance matrix $\mathbf{C}$ can therefore be computed as follows:

$$\begin{aligned}
\mathbf{C} &= \frac{1}{N-1}(\mathbf{A} - \frac{1}{N}\mathbf{1}_N\mathbf{A})^T(\mathbf{A} - \frac{1}{N}\mathbf{1}_N\mathbf{A}) \\
&= \frac{1}{N-1}\left[(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N)\mathbf{A}\right]^T\left[(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N)\mathbf{A}\right] \\
&= \frac{1}{N-1}\mathbf{A}^T\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right)\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right)\mathbf{A}
\end{aligned} \tag{2.5}$$

where we have use the following property of matrix transpose $(\mathbf{XY})^T = \mathbf{Y}^T\mathbf{X}^T$ on the last line. Furthermore, since the identity matrix and the matrix of all ones are symmetric, i.e., $\mathbf{I}_N^T = \mathbf{I}_N$ and $\mathbf{1}_N^T = \mathbf{1}_N$, therefore $(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N)^T = (\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N)$. Finally, it can be shown that the matrix $(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N)$ is idempotent, which means it is the

same as the square of the matrix:

$$\begin{aligned}(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N)(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N) &= \mathbf{I}_N - \frac{2}{N}\mathbf{1}_N + \frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N \\ &= \mathbf{I}_N - \frac{2}{N}\mathbf{1}_N + \frac{1}{N^2}N\mathbf{1}_N \\ &= \mathbf{I}_N - \frac{2}{N}\mathbf{1}_N + \frac{1}{N}\mathbf{1}_N \\ &= \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N, \end{aligned} \tag{2.6}$$

where $\mathbf{1}_N\mathbf{1}_N = N\mathbf{1}_N$ is an $N \times N$ matrix whose elements are equal to $N$. Substituting (2.6) into (2.5), we obtain:

$$\mathbf{C} = \frac{1}{N-1}\mathbf{A}^T\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right)\mathbf{A} \tag{2.7}$$

(b) Using singular value decomposition, the matrix $\mathbf{A}$ can be factorized as follows: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is the $N \times N$ left singular matrix, $\mathbf{\Sigma}$ is the $N \times d$ matrix containing the singular values, and $\mathbf{V}$ is the $d \times d$ right singular matrix. Similarly, using eigenvalue decomposition, the covariance matrix can be factorized as follows: $\mathbf{C} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T$. Show the relationship between SVD and PCA is given by the following equation:

$$\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T - \frac{1}{N}\mathbf{A}^T\mathbf{1}_N\mathbf{A} = (N-1)\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T.$$

**Answer:** From the previous question, we can write:

$$\mathbf{C} = \frac{1}{N-1}\mathbf{A}^T\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right)\mathbf{A} = \frac{1}{N-1}\left(\mathbf{A}^T\mathbf{A} - \frac{1}{N}\mathbf{A}^T\mathbf{1}_N\mathbf{A}\right) \tag{2.8}$$

Since $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ and $\mathbf{U}$ is an orthogonal matrix,

$$\mathbf{A}^T\mathbf{A} = [\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T]^T[\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T] = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T.$$

If $N > d$, then $\Sigma$ has $N-d$ rows of all zeros. If we remove such rows, $\Sigma$ becomes a $d \times d$ square matrix and $\Sigma^T\Sigma = \Sigma^2$. By substituting $\mathbf{C} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T$ and $\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}$ into Equation (2.8), we have:

$$\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T = \frac{1}{N-1}\left[\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T - \frac{1}{N}\mathbf{A}\mathbf{1}_N\mathbf{A}\right].$$

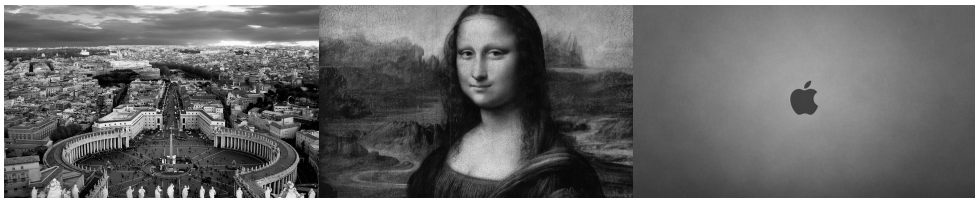
(c) Find the relationship between the right singular matrix $\mathbf{V}$ and the matrix of principal components $\mathbf{X}$ if the data matrix $\mathbf{A}$ has been column-centered (i.e., every column of $\mathbf{A}$ has been subtracted by the column mean) before applying SVD.

**Answer:** If the matrix $\mathbf{A}$ has been column-centered, then its column mean is zero, which means $\mathbf{A}^T\mathbf{1}_N$ is a matrix of all zeros. Thus, the last equation in the previous question reduces to:

$$\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T = \frac{1}{N-1}\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T.$$

This suggests that the right singular matrix $\mathbf{V}$ corresponds to the principal components $\mathbf{X}$, while the square root of the singular values are the same as $N-1$ times the eigenvalues.

9. Principal component analysis (PCA) can be used for image compression by transforming a high-resolution image into its lower rank approximation. In this exercise, you will be provided with the following three images of size 1080 $\times$ 1920 pixels each (the filenames are `img1.jpg`, `img2.jpg`, and `img3.jpg`).



(a) img1 (b) img2 (c) img3

**Figure 2.1.** Image data set.

You will use Matlab to apply PCA to each of the following images.

(a) Load each image using the `imread` command. For example:

```
matlab> A = imread('img1.jpg');
```

(b) Plot the image in gray scale.

```
matlab> imagesc(A);
matlab> colormap(gray);
```

**Answer:** See Figure 2.1.

(c) Apply principal component analysis to obtain a reduced rank approximation of the image.

For example, to obtain a rank-10 approximation (i.e., using the first 10 principal components), use the following commands:

```
matlab> A = double(A);          % convert A from uint8 to double format
matlab> [U,V] = princomp(A);    % apply principal component analysis
matlab> rank = 10;              % set rank to be 10
matlab> B = V(:,1:rank)*U(:,1:rank)';   % B is the compressed image of A
matlab> figure;
matlab> imagesc(B);
matlab> colormap(gray);
```

For each image, vary the rank (i.e., number of principal components) as follows: 10, 30, 50, and 100. Save each image as follows:

```
matlab> saveas(gcf, 'filename.jpg', 'jpeg');
```

Insert the compressed (reduced rank) images to the solution file of your homework (don't submit the jpg files individually).

**Answer:** See Figure 2.2.

(d) Compare the size of matrix A (in bytes) to the total sizes of matrices U and V (in bytes). Compute the compression ratio:

$$\text{Compression ratio} = \frac{\text{Size of matrix A}}{\text{Size of matrix U} + \text{Size of matrix V}}$$

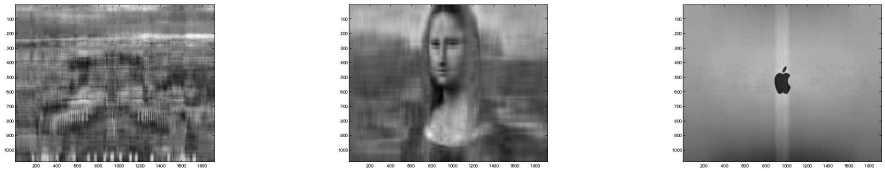
for each reduced rank (10, 30, 50, 100) of the images. You can use the `whos` command to determine the size of the matrices:

```
matlab> whos A U V
```

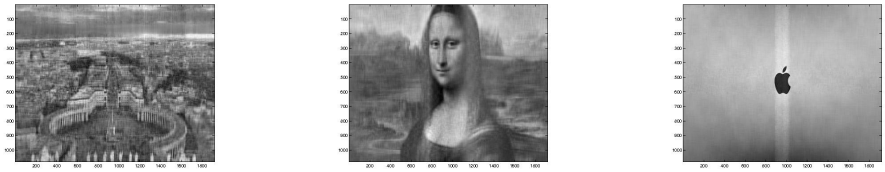
**Answer:** See Table 2.1.

| rank | size of A | size of U | size of V | compression rate |
|---|---|---|---|---|
| 10 | 16588800 | 153600 | 86400 | 69.12 |
| 30 | 16588800 | 460800 | 259200 | 23.04 |
| 50 | 16588800 | 768000 | 432000 | 13.824 |
| 100 | 16588800 | 1536000 | 864000 | 6.912 |

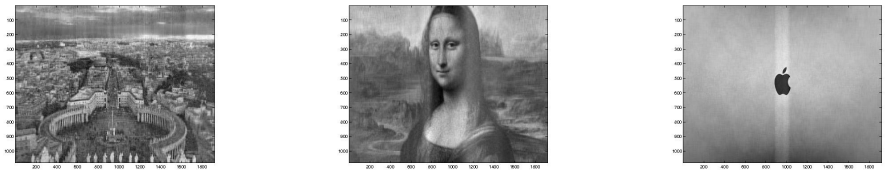
**Table 2.1.** Compression ratio for various images



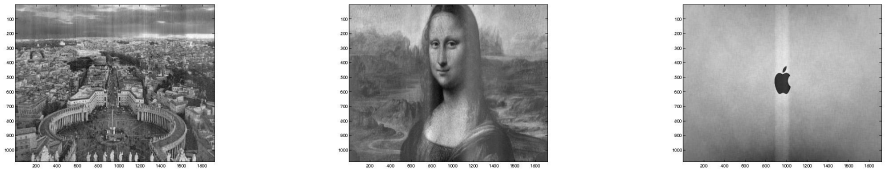
(a) rank 10 img1 (b) rank 10 img2 (c) rank 10 img3



(d) rank 30 img1 (e) rank 30 img2 (f) rank 30 img3



(g) rank 50 img1 (h) rank 50 img2 (i) rank 50 img3



(j) rank 100 img1 (k) rank 100 img2 (l) rank 100 img3

**Figure 2.2.** Reduced-rank images using PCA

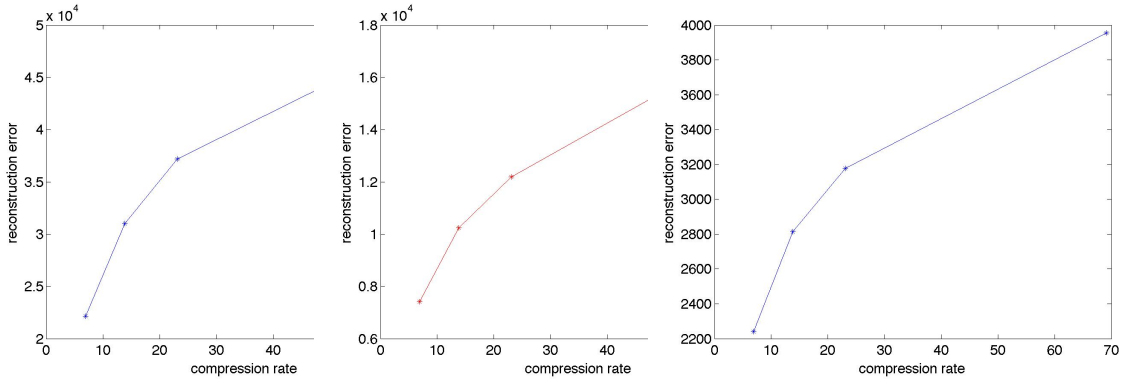
(e) Compute the reconstruction error $\|A - B\|_F$ of each reduced rank image, where $\|\cdot\|_F$ denote the Frobenius norm of a matrix. Note that the higher the reconstruction error, the lower the quality of the compressed image. Plot a graph of reconstruction error (y-axis) versus compression ratio (x-axis) for each image.

**Answer:** See Table 2.2 and Figure 2.3.

(f) State the minimum number of principal components (10, 30, 50, 100) needed to (visually) retain most of the salient features of each

| image | rank | reconstruction error |
|---|---|---|
| img1 | 10 | $4.9565 \times 10^4$ |
| img1 | 30 | $3.7198 \times 10^4$ |
| img1 | 50 | $3.0998 \times 10^4$ |
| img1 | 100 | $2.2135 \times 10^4$ |
| img2 | 10 | $1.7798 \times 10^4$ |
| img2 | 30 | $1.2190 \times 10^4$ |
| img2 | 50 | $1.0236 \times 10^4$ |
| img2 | 100 | $7.4063 \times 10^3$ |
| img3 | 10 | $3.9544 \times 10^3$ |
| img3 | 30 | $3.1775 \times 10^3$ |
| img3 | 50 | $2.8146 \times 10^3$ |
| img3 | 100 | $2.2397 \times 10^3$ |

**Table 2.2.** Reconstruction error for various images




(a) img1 (b) img2 (c) img3

**Figure 2.3.** Reconstruction error versus compression ratio

image (i.e., the city square in `img1.jpg`, shape of the face in `img2.jpg`, and shape of the apple in `img3.jpg`). Which image requires the least number of principal components? Which image requires the most number of principal components?

**Answer:**

img1.jpg: 50 components
img2.jpg: 30 components
img3.jpg: 10 components

## 2.3 Measures of Similarity and Dissimilarity

1. Consider the following binary vectors:

   $\mathbf{x_1} = (1, 1, 1, 1, 1)$

   $\mathbf{x_2} = (1, 1, 1, 0, 0)$

   $\mathbf{y_1} = (0, 0, 0, 0, 0)$

   $\mathbf{y_2} = (0, 0, 0, 1, 1)$

   (a) According to Jaccard coefficient, which pair of vectors—$(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{y}_1, \mathbf{y}_2)$—are more similar to each other?

   **Answer:**

   $\text{Jaccard}(\mathbf{x}_1, \mathbf{x}_2) = \frac{3}{5} = 0.6$.

   $\text{Jaccard}(\mathbf{y}_1, \mathbf{y}_2) = \frac{0}{5} = 0$.

   Therefore, according to Jaccard coefficient, $(\mathbf{x}_1, \mathbf{x}_2)$ are more similar.

   (b) According to simple matching coefficient, which pair of vectors—$(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{y}_1, \mathbf{y}_2)$—are more similar to each other?

   **Answer:**

   $\text{SMC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{3}{5} = 0.6$.

   $\text{SMC}(\mathbf{y}_1, \mathbf{y}_2) = \frac{3}{5} = 0.6$.

   Therefore, according to simple matching coefficient, they are both equally similar.

   (c) According to Euclidean distance, which pair of vectors—$(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{y}_1, \mathbf{y}_2)$—are more similar to each other?

   **Answer:**

   $\text{Euclidean}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2} = 1.4142$.

   $\text{Euclidean}(\mathbf{y}_1, \mathbf{y}_2) = \sqrt{2} = 1.4142$.

   Therefore, according to Euclidean distance, they are both equally similar.

2. Consider a weighted, undirected, graph $G$ (see Figure 2.4 as an example). Let $e(u, v)$ be the weight of the edge between nodes $u$ and $v$, where $e(u, u) = 0$ and $e(u, v) = \infty$ if $u$ and $v$ is disconnected. Assume the

graph is a connected component, i.e., there exists a path between every two nodes. Suppose the path length, $d(u, v)$, is defined as follows:

$$d(u,v) = \begin{cases} 0 & \text{if } u = v; \\ e(u,v), & \text{if there is an edge between } u \text{ and } v; \\ \min_{w \neq u \neq v} d(u,w) + d(w,v), & \text{otherwise.} \end{cases}$$

Is $d(u, v)$ a metric? State your reasons clearly. (Check whether the positivity, symmetry, and triangle inequality properties are preserved.).

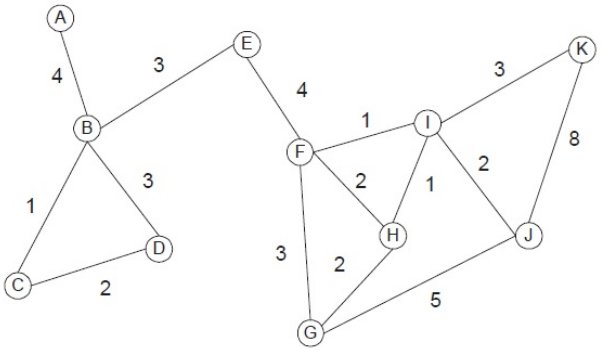



**Figure 2.4.** Weighted undirected graph.

**Answer:**

(a) Positivity property is preserved by definition since $d(u, u) = 0$ and $d(u, v) > 0$ if $u \neq v$.

(b) Symmetry property is preserved since the graph is undirected.

(c) Triangle inequality is not preserved. A counter-example is $d(K, J) \geq d(K, I) + d(I, J)$.

Therefore $d(u, v)$ is not a metric.

3. For document analysis, numerous measures have been proposed to determine the *semantic similarity* between two words using a domain ontology such as WordNet. For example, words such as `dog` and `cat` have higher semantic similarity than `dog` and `money` (since the former refers to two types of carnivores). Figure 2.5 below shows an example for computing the Wu-Palmer similarity between `dog` and `cat` based on their path

length in the WordNet hypernym hierarchy. The depth $h$ refers to the length of the shortest path from the root to their lowest common hypernym (e.g., `carnivore` for the word pair `dog` and `cat`), whereas $k$ is the minimum path length between the two words.

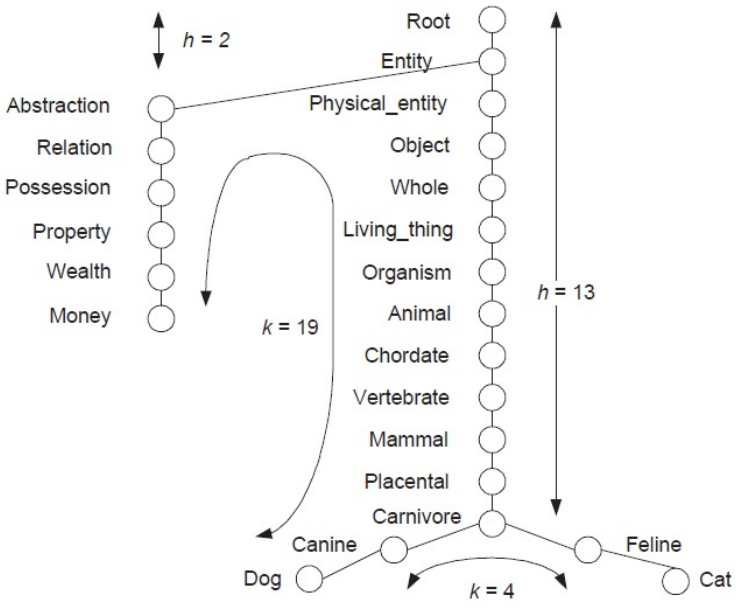



**Figure 2.5.** Sample of the hypernym hierarchy in WordNet.

The Wu-Palmer similarity measure is defined as follows:

$$W = \frac{2h}{k + 2h}$$

For example[1], for `dog` and `cat`, $W = 26/(4 + 26) = 0.867$, whereas for `dog` and `money`, $W = 4/(19 + 4) = 0.174$.

(a) What is the maximum and minimum possible value for Wu-Palmer similarity?

[1]In this simplified example, we assume each word has exactly 1 sense. In general, a word can have multiple senses. As a result, the Wu-Palmer measure is given by the highest similarity that can be achieved using one of its possible senses.

**Answer:** Maximum value is 1; minimum value approaches 0.

(b) Let $1 - W$ be the Wu-Palmer distance measure.

- Does $1 - W$ satisfy the positivity property?
  **Answer:** Yes. Since $1 - W = \frac{k}{2h} = 0$ when $k = 0$, this implies that $d(u, v) = 0$ if and only if $u = v$.
- Does $1 - W$ satisfy the symmetry property?
  **Answer:** Yes because $W$ is a symmetric measure.
- Does $1 - W$ satisfy the triangle inequality property?
  **Answer:** No because each node can have more than one path to the root, some maybe shorter than others. For example, the words (money, statute) are very dissimilar to each other. But (money, bill) and (bill, statute) are very similar, thus violating triangle inequality. The actual path for these words in the WordNet ontology are shown in Figure 2.6.

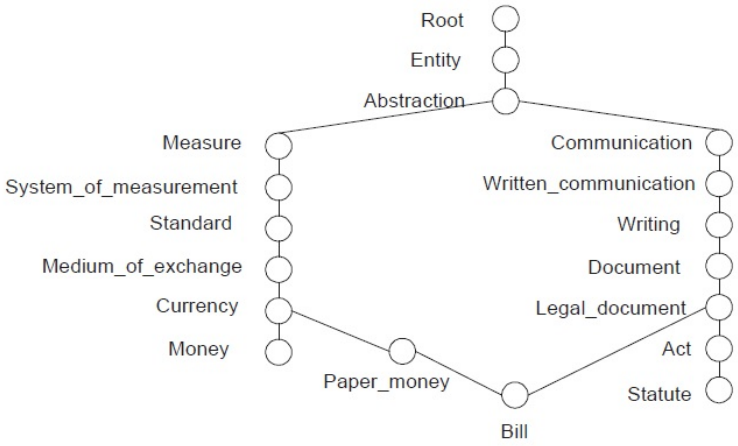



**Figure 2.6.** Sample of the hypernym hierarchy in WordNet.

4. Suppose you are given a census data, where every data object corresponds to a household and the following continuous attributes are used to characterize each household: total household income, number of household residents, property value, number of bedrooms, and number of vehicles owned. Suppose we are interested in clustering the households based on these attributes.