CSE 881: Data Mining (Fall 2016 Homework 1)

Due date: Sept 13, 2016 (before midnight).

A soft copy of your homework must be submitted via handin. All submitted homework must be your own work. Students who plagiarize the solution of others or allow others to plagiarize their work will receive zero for the homework.

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

   - discrete or continuous AND
   - qualitative or quantitative AND
   - 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.

   Example: Temperature in Kelvin

   Answer: Continuous, quantitative, ratio.

   (a) IP address of a device.

   (b) Week of a year, whose value is in the range between 1 to 52.

   (c) Speed of a vehicle.

   (d) Sound intensity in decibel (dB) scale.

   (e) Total household income, measured in terms of dollars above median US household income ($51,939). For example, if the total household income is $101,939, the value recorded is $50,000.

   (f) U.S. military rank (e.g., private, corporal, sergeant, and major).

2. State whether it is valid to apply the following operations to the attributes given below (based on the properties of the attribute values). If not valid, state your reason clearly.

   (a) Calculating the median salary of computer engineers.

   (b) Calculating the correlation between weight and height of individuals.

   (c) Calculating the average magnitude of seismic energy released during earthquakes in a particular region for the past 50 years, where the magnitude of seismic energy released is measured in Richter scale. See `https://en.wikipedia.org/wiki/Richter_magnitude_scale` for more explanation about the Richter scale.

   (d) Calculating the total entropy of a population based on the gender of the individual members of the population.

   (e) Calculating the geometric mean of daily temperature (in degree Celsius) for a given location based on its 50 year historical records.

(f) Calculating the standard deviation of a student's GPA based on the courses taken by the student.

3. Consider the diagram shown in Figure 1. Let

$$\mathbf{x}_1 = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.2 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 0.5 \\ 0.6 \\ 0.4 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 0.1 \\ 0.4 \\ 0.8 \end{pmatrix}$$

and Ω is the subspace spanned by the pair of vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$.

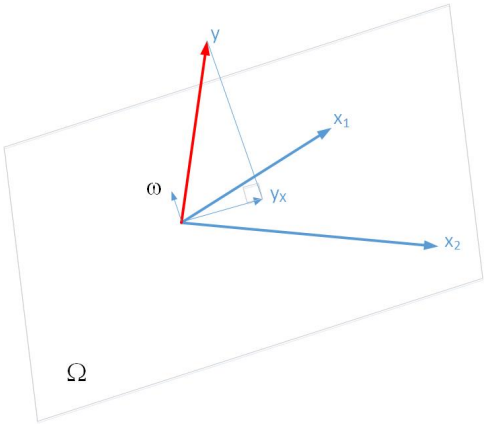

Figure 1: Orthogonal projection of vector $\mathbf{y}$ onto a subspace spanned by vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.

Write a Matlab or python program to compute $\mathbf{y}_x$, which is the projection of y onto the subspace Ω, and ω, which is the unit vector orthogonal to the hyperplane defined by Ω. You should submit your source code and the sample output produced by your program. For Matlab users, you can use the `diary` command to spool the display buffer to a file. See the `Introduction to Matlab` handout provided as one of the reading materials for lecture 3 on the class web page.

4. Suppose you are given a database of patient's demographic information from a healthcare provider. The covariance matrix obtained for three attributes: age, weight, and systolic blood pressure (bp) is shown below:

$$\begin{array}{r} age \rightarrow \\ weight \rightarrow \\ bp \rightarrow \end{array} \begin{pmatrix} 389.75 & 199.37 & 135.12 \\ 199.37 & 610.52 & 426.30 \\ 135.12 & 426.30 & 359.36 \end{pmatrix}$$

(a) Does this imply that user's age is more correlated with his/her weight than systolic blood pressure? Answer yes or no and explain your reasons clearly.

(b) Suppose the weight attribute is centered by subtracting it with the average weight of all patients in the database. For example, a 200-pound patient has a weight recorded as -50 (if the average weight of the patients is 150 pounds). Would the covariance between the centered weight attribute and age be greater than, smaller than, or equal to 199.37? To obtain full credit, you must prove your answer by showing the computations clearly.

(c) If the measurement unit for weight is converted from pounds to kilograms (where 1 kg = 2.2 pounds), will the covariance between weight (in kilogram) and age be greater than, smaller than, or equal to 199.37? To obtain full credit, you must prove your answer by showing the computations clearly.

(d) Suppose you standardize both the age and weight attributes (by subtracting their respective means and dividing by their corresponding standard deviations). Would their covariance value be greater than, smaller than, or equal to 199.37? To obtain full credit, you must prove your answer by showing the computations clearly.

5. You developed a new method for predicting human activities from a wearable sensor device. To evaluate its effectiveness, you applied the method to a data set containing 150 labeled examples and found its accuracy to be 95%. You then applied a baseline method on the same data and found its accuracy to be 85%. Is it safe to conclude (say, with 95% confidence) that your method outperforms the baseline method? Answer yes or no and explain your reason. To obtain full credit, you must and show your calculations clearly.