# CSC 240/440 - 2017 Spring: Homework 1

Hand in the hardcopy before the class on Feb. 14

## Requirement

Due to the request from some students, the homework is posted online right now, but would be updated probably every week until it is formally released. (A) or (G) indicates questions for all or just graduate students. Undergraduate students are not required to do (G) questions, but they can get bonus points from that. **Please hand in the hardcopy of your homework before the class.**

The homework must be completed individually. However, you are encouraged to discuss the general algorithms and ideas with classmates in order to help you answer the questions. If you work with one or more other people on the general discussion of the assignment questions, please record their names over every question they participated.

However, the following behaviors will receive heavy penalties (lose all points and apply the honest policy explained in syllabus)

- explicitly tell somebody else the answers;
- explicitly copy answers or code fragments from anyone or anywhere;
- allow your answers to be copied;
- get code from Web.

Please also indicate how many late days you want to apply to your submission (Check the late policy in the Syllabus). All late submission without indicating late days or running out the late days cannot be accepted. For medical reasons, if the homework in time cannot be submitted on time, you have to submit the certificate with your homework.

## 1 (A) 2 points

Classify the attributes as binary/discrete/continuous. Also classify them as qualitative (nominal/ordinal) or quantitative (interval/ratio). Also briefly indicate your reasoning for your answer. Example: Age in years. Answer: Discrete, quantitative, ratio.

1. Student ID number;
2. Body weight;
3. Students' scores of homework 1;
4. Number of students in school.

## 2 (A) 3 points

Read the textbook and give a practical example for each task in data mining:

- Classification
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery
- Regression
- Deviation Discovery

(Do not use the examples included in the slides or mentioned in our class.)

## 3 (A) 2 points

Answer the following questions about cosine measure:

1. What is the range of values that are possible for the cosine measure?
2. If two objects have a cosine measure of 1, are they identical? Why or why not?

## 4 (A) 2 points

Calculate the indicated similarity or distance measures for vectors $\mathbf{x}, \mathbf{y}$:

1. $\mathbf{x} = (1, 0, 1, 0), \mathbf{y} = (0, 1, 0, 1)$ cosine, correlation, Euclidean, Jaccard;
2. $\mathbf{x} = (0, -1, 0, 1), \mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean;
3. $\mathbf{x} = (1, 1, 0, 1, 0, 1), \mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard;
4. $\mathbf{x} = (2, -1, 0, 2, 0, -3), \mathbf{y} = (-1, 1, -1, 0, 0, -1)$ cosine, correlation.

## 5 (A) 2 points

Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, provide and prove the mathematical relationship between cosine similarity $cos(\mathbf{x}, \mathbf{y})$ and Euclidean distance $d(\mathbf{x}, \mathbf{y})$ when each data object has an L2 length of 1 i.e. $\sum_{i=1}^{p} x_i^2 = 1, \sum_{i=1}^{p} y_i^2 = 1$.

## 6 (A) 2 points

Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.

## 7 (A) 2 points

Different data visualization methods have their own advantages on different applications.

1. Please give a specific example which is more suitable to use histogram than stem and leaf plot;
2. Please give a specific example which is more suitable to use scatter plots than histogram (both are two-dimensional).

## 8 (A) 2 points

Please answer the following about box plot:

1. Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed;
2. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11 of the text book (Introduction to data mining)?

## 9 (A) 4 points

Read Appendix A in the textbook about the introduction to linear algebra and answer the following questions

1. What are eigenvalues and eigenvectors? (Give the mathematical description)
2. What is SVD? (Give the mathematical description)
3. If you know the SVD of matrix $A = U\Sigma V^\top$ and how to get the SVD of matrix $A^\top A$ and $AA^\top$?
4. Prove that if $A$ is a symmetric matrix with two distinct eigenvalues $\lambda_1$ and $\lambda_2$, then the eigenvector corresponding to $\lambda_1$ is orthogonal to the eigenvector corresponding to $\lambda_2$.

## 10 (A) 4 points

This is a programming question, here we use the Iris dataset (which is used in chapter 3 of the book). Please write program to implement the following:

1. Load the data from the text file, generate three matrices $(50 \times 4)$ to contain the flowers with 4 attributes from each class respectively. Save the three matrices to 3 text files and use the class names as file names.
2. Plot a histogram for petal length of all 150 flowers. You should use 10 bins and the range should be set as [min-petal-length, max-petal-length].
3. Make a scatter plot with attributes sepal length and sepal width. Use different markers for different classes (provide a legend to show marker-class relation).

This programming assignment is to help you to be familiar with libraries/tools to do matrix operation and data visualization. So we recommend you to use libraries rather than implementing everything from scratch (e.g., numpy[1] and matplotlib[2] packages if you are using Python). If you are not familiar with Python or numpy, http://cs231n.github.io/python-numpy-tutorial/ is a very good tutorial for beginners.

You should upload the code to Blackboard and print out the figures (histogram and scatter plot) as hard copy (hand in it together with your answers for other questions).

## 11 (A) 2 points

Question 21 in Chapter 2.

## 12 (G) 2 points

Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, provide and prove the mathematical relationship between correlation similarity $corr(\mathbf{x}, \mathbf{y})$ and Euclidean distance $d(\mathbf{x}, \mathbf{y})$ when each data point has been been standardized by subtracting its mean and dividing by its standard deviation.

## 13 (G) 2 points

Recall the three properties in metric axioms, define $d(\mathbf{x}, \mathbf{y}) = 1 - SMC(\mathbf{x}, \mathbf{y})$, where $SMC$ is Simple Matching Coefficient. Does $d(\cdot, \cdot)$ satisfy these properties? Prove your answer.

[1] http://www.numpy.org/
[2] http://matplotlib.org/