ALDA Fall 2018
HW 1
8/29/2018

HW1 contains 6 questions. Please read and follow the instructions.

- **DUE DATE FOR SUBMISSION: 9/13/2018 11: 45 PM**
- **TOTAL NUMBER OF POINTS: 100+5** (5 bonus points if you follow all the instructions and 0 otherwise)
- Make sure you clearly list each team member's **names and Unity IDs** at the top of your submission.
- Your submission should be a **single zip file** containing a PDF of your answers, your code, and a readme file with running instructions. Please follow the naming convention for your zip file: H(homework group number)_HW(homework number), e.g. H1_HW1.

---

1. (13 points) [**Song Ju**] Classify the following attributes as binary, discrete, or continuous. Also classify them as nominal, ordinal, interval, or ratio. Some cases may have more than one interpretation, so briefly justify your answer if you think there may be some ambiguity.
   (a) (1 point) Hair color (Black, Blonde, Red)
   (b) (1 point) Level of agreement (yes, maybe, no)
   (c) (1 point) Income earned in a week
   (d) (1 point) Celsius temperature
   (e) (1 point) Genotype (Bb, bb, BB, bB)
   (f) (1 point) ISBN numbers for books.
   (g) (1 point) Time in terms of AM or PM
   (h) (1 point) Waiting number for restaurant
   (i) (1 point) Years of work experience
   (j) (1 point) Categorization of clothing (hat, shirt, pants, shoes)
   (k) (1 point) Angles as measured in degrees between 0 and 360
   (l) (1 point) Ratings of movies (G, PG, R)
   (m) (1 point) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a numb that you can use to claim your coat when you leave.)

   SOLUTIONS:
   (a) Discrete, Nominal

(b) Discrete, Ordinal

(c) Continuous, Ratio

(d) Continuous, Interval

(e) Discrete, Nominal

(f) Discrete, Nominal (or ordinal because ISBN numbers do have some order information)

(g) Binary, Nominal

(h) Discrete, interval

(i) Discrete, Ratio

(j) Discrete, Nominal

(k) Continuous, Ratio

(l) Discrete, Ordinal

(m) Discrete, Nominal (or ordinal if you are using the order information)

2. (10 points) [**Ruth Okoilu**] Data Transformation.
In natural language processing, we often use **term frequency** and **inverse document frequency** transformation (tf'_{ij}), defined by the following equation:

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i} \tag{1}$$

where tf_{ij} is the term frequency of the i^{th} word (term) in the j^{th} document, m is the number of documents, and df_i is the number of documents in which the i^{th} term appears.

Alternatively, we can define (tf''_{ij}) as:

$$tf''_{ij} = tf_{ij} * \log \frac{\sum_{k=1}^{m} d_k}{\sum_{k=1}^{df_i} d_k} \tag{2}$$

where d_k is the length of a document k.

Assume the max term frequency tf_{ij} is p and answer the following questions.

(a) (6 points) What are the maximum and minimum values of tf'_{ij} and tf''_{ij} respectively? Please specify what cases the max and min value achieves.

(b) (4 points) Briefly explain the purpose for using tf'_{ij} and tf''_{ij} respectively in the context of natural language processing and also explain what is the main difference between tf'_{ij} and tf''_{ij}.

SOLUTIONS:

(a) i. max: $p * log(m)$ in case that a term occurs in only one document.
min: $tf * log(m/m) = tf * 0 = 0$ in case that a term occurs in every document.

ii. max: $p * log((\sum_{k=1}^{m} d_k) * p)$ in case that a term occurs once in only one document.
min: $tf * log(m/m) = tf * 0 = 0$ in case that a term occurs in every document.
$m = df_i$

(b) i. This transformation is used for the purpose of extracting subjects or themes of documents, or clustering documents by similarity, because tf'_{ij} lets us find out relatively unique and important word sets which are dominantly used in each document. Note that as a term more frequently appears in less documents, the transformed value gets bigger. Thus this transformation makes the frequency of common terms diminished (e.g. article, auxiliary verb, conjunction, etc.), while it makes the frequency of relatively unique terms intensified. tf''_{ij} takes the total length of all documents into consideration when finding these unique word sets. The smaller the $\sum_{k=1}^{df_i} d_k$ value for an i^{th} term, the more important that term is.

ii. The main difference between these two weights is that tf'_{ij} finds the log of the value resulting from dividing m by the number document in which i^{th} term appears while tf''_{ij} finds the log of the total length of all documents divided by the total length of documents in which i^{th} term appears. The later is describes the effect of document length in determining unique words set while the former uses document count to determine this uniqueness.

3. (8 points) [**Xi Yang**] Answer the following questions:

(a) (4 points) A healthcare dataset contains 523,000 patients. Among these patients, 26,150 patients have albinism and the remaining 496,850 patients have normal skin. Suppose we will sample 1,000 patients from the dataset to conduct albinotic analysis, which sampling method should be selected to apply in this situation: simple random sampling or stratified sampling, and why? With the selected sampling method, how many albinotic and normal skin patients will be sampled, respectively?

(b) (4 points) Consider the following scenario, a patient's systolic blood pressure (SBP) is *recorded* to be 250. When SBP is higher than 180, a patient is considered to have hypertensive crisis and need to seek the emergency care. For this given scenario, is the recorded data noise or outlier? And why? (no point will be given if you do not give a justification).

SOLUTIONS:

(a) (4 points) Stratified sampling is supposed to be utilized. The given dataset is highly imbalanced. If we use the simple random sampling, we might get a biased set with even higher ratio of normal skin patients to albinotic patients. With the stratified sampling, there will be 50 albinotic patients and 950 normal skin patients sampled from the original dataset.

(b) (4 points) Definition of noise and outlier: a) Noise: Data point has been measured incorrectly due to faulty equipment. Noise usually randomly overlaps with true signal in the observed data; b) Outliers: Data point has been measured correctly, but its behavior is significantly different compared to the rest of data points. They are legitimate values and studying them could be very important for certain tasks.

For the given scenario, without assumption for the distribution of dataset, we cannot directly make the conclusion that whether the recorded data 250 is noise or outlier.

If other patients' SBP are mostly distributed lower than 180, i.e. most patients are normal cases, since 250 is a legitimate value for SBP, then the recorded 250 can be considered as an outlier, which can be utilized for detecting the hypertensive crisis. While if other patients' SBP are also mostly distributed higher than 180, i.e. most of the patients in dataset have hypertensive crisis, then the recorded 250 might be caused by equipment error and can be considered as a noise.

4. (15 points) [**Song Ju**] Write your code in Matlab, R or Python to perform the following tasks, please *report your outputs and key codes* in the document file, and also include your code file (end with .m, .r or .py) in the .zip file.
   (a) (1 point) Generate a 5*5 identity matrix A.
   (b) (1 point) Change all elements in the 2nd column of A to 3.
   (c) (1 point) Sum of all elements in the matrix (use a "for/while loop").
   (d) (1 point) Transpose the matrix A ($A = A^T$)
   (e) (2 points) Calculate sum of the 3rd row, and the diagonal in the matrix A.
   (f) (1 point) Generate a 5*5 matrix B following Gaussian Distribution with mean 5 and variance 3.
   (g) (2 points) From B, using matrix operations to get a new matrix C such that, the first row of C is equal to the first row of B times the second row of B, the second row of C is equal to the sum of the 3rd and 4th row of B minus the 5th row of B.
   (h) (2 points) From C, using one matrix operation to get a new matrix D such that,the first column of D is equal to the first column of C times 2, the second column of D is equal to the second column of C times 3 and so on.
   (i) (2 points) $X = [2, 4, 6, 8]^T$, $Y = [6, 5, 4, 3]^T$, $Z = [1, 3, 5, 7]^T$. Compute the covariance matrix of X, Y and Z.
   (j) (2 points) Verify the equation: $\bar{x^2} = (\bar{x}^2 + \sigma^2(x))$,
   using $x = [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]^T$. $\sigma(x)$ is the standard deviation.

SOLUTIONS: The solution is written in Python. Please refer to the 'hw1q4.py'. The outputs are listed as follows (some of them may not be the same as yours) same with yours:

(a) $\mathrm{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

(b) $\mathrm{A} = \begin{bmatrix} 1 & 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 0 & 3 & 0 & 1 & 0 \\ 0 & 3 & 0 & 0 & 1 \end{bmatrix}$

(c) 19

(d) $A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 3 & 3 & 3 & 3 & 3 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

(e) sum of the 3rd row = 1,
the sum of diagonal = 7

(f) $B = \begin{bmatrix} 6.93378889 & 9.50039536 & 8.40646264 & 11.62771148 & 8.6005849 \\ 10.09253773 & 5.60740753 & 5.33635357 & 3.3332325 & 6.98752526 \\ 3.35575623 & 0.91208348 & 6.24220467 & 7.57714434 & 2.15516488 \\ 2.73964882 & 9.96481621 & 2.20060513 & 2.82164384 & 5.25631463 \\ 2.22536547 & 4.95561462 & 3.53721855 & 7.69931492 & 5.57625638 \end{bmatrix}$

(g) $C = \begin{bmatrix} 69.97952593 & 53.27258848 & 44.85985687 & 38.75786582 & 60.09680425 \\ 3.87003958 & 5.92128507 & 4.90559125 & 2.69947326 & 1.83522314 \end{bmatrix}$

(h) $D = \begin{bmatrix} 139.95905186 & 159.81776545 & 179.43942749 & 193.78932908 & 360.58082551 \\ 7.74007917 & 17.7638552 & 19.62236502 & 13.49736632 & 11.01133882 \end{bmatrix}$

(i) Covariance matrix: $\begin{bmatrix} 6.66666667 & -3.33333333 & 6.66666667 \\ -3.33333333 & 1.66666667 & -3.33333333 \\ 6.66666667 & -3.33333333 & 6.66666667 \end{bmatrix}$

(j) Both side of the equation equals to 154

5. (33 points) [**Ruth Okoilu**] For this exercise, use the provided 'seeds.csv' file, which contains a list of 210 data instances. The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. (Source: https://archive.ics.uci.edu/ml/datasets/seeds) There are 8 columns representing: 1) area A, 2)perimeter P, 3) compactness, 4) length of kernel, 5) width of kernel, 6) asymmetry coefficient, 7) length of kernel, and 8) groove Class (Type of wheat). For the purpose of this exercise, you consider two features, 'area_A' and 'kernel_width' (columns 1 & 5) of the provided 'seeds.csv' dataset. Write your codes in Matlab, R or Python to perform the following tasks, please *report your outputs and key codes* in the document file, and also include your code file (end with .m, .r or .py) in the .zip file.

   (a) (3 points) Load the file and read 'area_A' and 'kernel_width' columns and save them as the original *raw dataset*. Apply normalization (transformed data $z \in [0, 1]$) to the raw dataset to get the *normalized dataset* and apply the standardization to the raw dataset to get the *standardized dataset*. Show the range of the two features in each dataset.

   (b) (30 points) Perform the following operations on the raw, normalized and standardized datasets respectively.

      i. (3 points) Make a 2D plot of the values and label the axes (area_A should be x-axis and kernel_width should be y-axis). Compare the three plots.

      ii. (3 points) Compute the mean of area_A and kernel_width values. Consider this point as P.

iii. (9 points) Compute the distance between P and the 210 data points using the following distance measures: 1) Euclidean distance, 2) Mahalanobis distance, 3) City block metric, 4) Minkowski metric (for r=3), 5) Chebyshev distance, 6) Cosine distance and 7) Canberra distance.

iv. (3 points) For each distance measure, identify the 10 points from the dataset that are the closest to the point P from (ii). (You are allowed to use any package functions to calculate the distances.)

v. (6 points) Create plots, one for each distance measure. Place an 'X' for P and mark the 10 closest points. To mark them, you could place a circle or draw the line between these closest neighbors and the points 'X'. Make sure the points can be uniquely identified.

vi. (3 points) Verify if the set of points is the same across all the distance measures. If there is any big difference, briefly explain why it is.

vii. (3 points) Reason about your results and state the importance of data transformation in the dataset.

SOLUTIONS: The solution is written in Python. Please refer to files 'hw1q5a.py', 'hw1q5b.py', 'hw1q5c.py'

(a) Raw: Area_A: [10.59, 21.18], kernel_width: [2.63,4.033]
Normalization: Area_A: [0,1] , kernel_width: [0,1]
Standardization: Area_A: [-1.4667, 2.1815], kernel_width:[-1.6682, 2.0551]

(b) i. See figures in hw1q5.pdf

ii. Raw P: area_A: 14.847524, kernel_width: 3.258605
Normalized P: area_A: 0.402031 kernel_width: 0.448045
Standardized: P: area_A : -1.895309e-15 kernel_width: -2.913014e-16

iii. See solution in python file hw1q5a.py, hw1q5b.py, hw1q5c.py

iv. See distance results in hw1q5.pdf

v. See figures in hw1q5.pdf

vi. While most distances have similar points or very close points (there are a few differences), cosine distance measure is different from the others, because cosine distance reflects the direction of distribution.

vii. We transform attributes of different units so that larger values will not dominate the data analysis and make our result bias. 'area_A' and 'kernel_width' are of different units. It can be seen that the resulting scale of area_A and kernel_width are now comparable.

6. (21 points) [**Xi Yang**] In this question, please summarize and explore data in the provided file "hw1q6_data.csv", which comes from the Pima Indians Diabetes Database (https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/). In this data file, each row indicates the data for a patient. The first 6 columns are features for patients, and the last column "*Class*" indicates if a patient has diabetes: 1 (diabetic) or 0 (nondiabetic). The specific meaning for each feature is as follows:

1. *Glucose*: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
2. *BloodPressure*: Diastolic blood pressure (mm Hg).

3. *SkinThickness*: Triceps skinfold thickness (mm).
4. *BMI*: Body mass index (weight in kg/(height in m)^2).
5. *DiabetesPedigreeFunction*: Diabetes pedigree function.
6. *Age*: (years).

Write code in Matlab, R or Python to perform the following tasks. Please *report your outputs and key codes* in the document file, and also include your code file (end with .m, .r or .py) in the .zip file.

(a) (1 point) How many diabetic and nondiabetic patients are in the dataset?

(b) (2 points) There are missing values in the features which are marked as 0. What is the missing rate (%) for each feature?

(c) (4 points) Specify two methods for missing data handling and discuss their respective advantages and disadvantages.

Remove the patients (rows) in dataset with missing values, then answer the following questions based on the **remaining data**:

(d) (1 point) How many diabetic and nondiabetic patients are in the remaining data?

(e) (3 points) Compute the mean, median, standard deviation, range, 25$^{\text{th}}$ percentiles, 50$^{\text{th}}$ percentiles, 75$^{\text{th}}$ percentiles for each feature.

(f) (4 points) Create histogram plot using 10 bins for the two features *BloodPressure* and *DiabetesPedigreeFunction*, respectively.

(g) (6 points) Quantile-quantile plot can be used for comparing the distribution of data against the normal distribution. Create quantile-quantile plot for the two features *BloodPressure* and *DiabetesPedigreeFunction*, respectively. Give a brief analysis for the two plots.

SOLUTIONS:

Python code is provided in the file 'hw1q6.py'.

(a) (1 point)
Number of diabetic patients: 268
Number of nondiabetic patients: 500

(b) (3 points)
Missing rates for features are as follows:
Glucose : 0.006510416666666667
BloodPressure : 0.045572916666666664
SkinThickness : 0.2955729166666667
BMI : 0.014322916666666666
DiabetesPedigreeFunction : 0.0
Age : 0.0

(c) (4 points) Missing data handling methods:

Complete-case analysis: omits records with missing entries, and then proceeds analysis only based on remaining data. It is straightforward to implement, but significantly limits the scope and power of study, and introduces bias when data is not missing at random.

Data Imputation: missing entries are estimated from the present values. Common approaches include mean- or median-filling, carrying forward, hot-deck, resampling, multiple imputation, knn, and so on. Comparing to complete-case analysis, it can utilize all present data for analysis. However, some data imputation methods are only suitable when missing rate is low, e.g. mean- or median-filling and knn.

After missing rows have been removed:

(d) (1 point)

Number of diabetic patients: 177
Number of nondiabetic patients: 355

(e) (3 points) The results are as follows:

Glucose :
Mean: 121.03007518796993
Median: 115.0
Standard Deviation: 30.970077688382087
Range: 143
25th percentile: 98.75
50th percentile: 115.0
75th percentile: 141.25

BloodPressure :
Mean: 71.50563909774436
Median: 72.0
Standard Deviation: 12.298678262218631
Range: 86
25th percentile: 64.0
50th percentile: 72.0
75th percentile: 80.0

SkinThickness :
Mean: 29.18233082706767
Median: 29.0
Standard Deviation: 10.51398226832601
Range: 92
25th percentile: 22.0
50th percentile: 29.0
75th percentile: 36.0

BMI :
Mean: 32.89022556390977

Median: 32.8
Standard Deviation: 6.874638633419096
Range: 48.89999999999999
25th percentile: 27.874999999999996
50th percentile: 32.8
75th percentile: 36.9

DiabetesPedigreeFunction :
Mean: 0.5029661654135338
Median: 0.41600000000000004
Standard Deviation: 0.34422227745866446
Range: 2.335
25th percentile: 0.25875000000000004
50th percentile: 0.41600000000000004
75th percentile: 0.6585

Age :
Mean: 31.61466165413534
Median: 28.0
Standard Deviation: 10.751464810071871
Range: 60
25th percentile: 23.0
50th percentile: 28.0
75th percentile: 38.0

(f) (4 points) For the two features *BloodPressure* and *DiabetesPedigreeFunction*, their histograms are as shown in Figure 1:

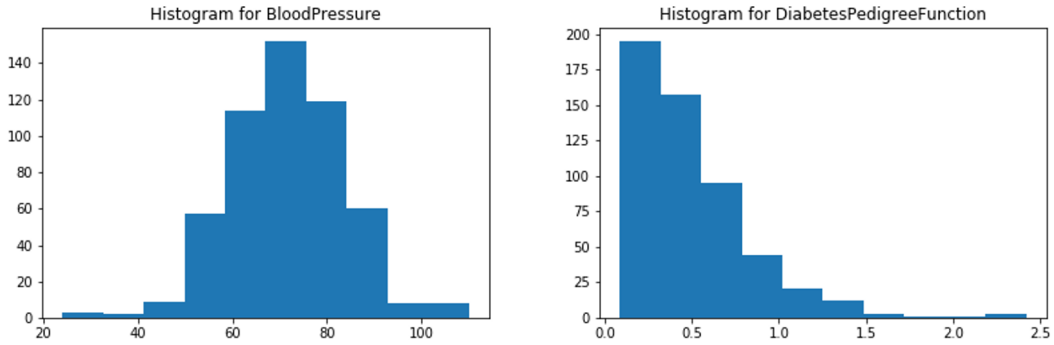

Figure 1: Histograms for the two features

(g) (6 points) For the two features *BloodPressure* and *DiabetesPedigreeFunction*, their quantile-quantile plots are as shown in Figure 2:

Quantile-quantile plots indicates the quantiles of our data against the quantiles of a normal distribution. If the data follows a normal distribution, it should be close to the straight red line which indicates the normal distribution.

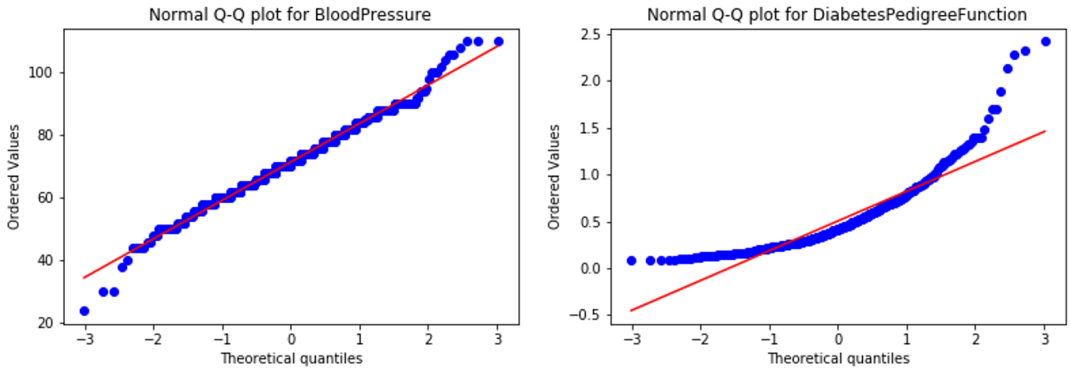

Figure 2: Quantile-quantile plots for the two features

According to the quantile-quantile plots, BloodPressure follows a normal distribution, while DiabetesPedigreeFunction does not follow the normal distribution. This can also be justified by the histograms in Figure 1.