

BNU-HKBU United International College
DS4013: Data Mining (For DS students)
Fall 2019 Assignment1

Q1. Discuss whether or not each of the following activities is a data mining task:

- 1) Dividing the students in a university based on their grades.
- 2) Predicting the outcomes of tossing a fair coin.
- 3) Extracting the frequencies of a sound wave.
- 4) Sorting a student database based on student identification numbers.
- 5) Monitoring the heart rate of a patient for abnormalities.

Q2. Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) appears in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by:

$$w_{ij} = tf_{ij} * \log \frac{m}{df_i}$$

where df_i is the number of documents in which the i^{th} term appears and is known as the document frequency of the term. The transformation ($\log \frac{m}{df_i}$) is known as the inverse

document frequency transformation.

- 1) What is the effect of this transformation if a term occurs in one document? In every document?
- 2) What might be the purpose of this transformation?

Q3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) [13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 40, 45].

- 1) Indicate the median, mode, 25th percentile and 75th percentile values.
- 2) Use *smoothing by bin median* to smooth these data (in this method, each bin value is replaced by its bin median value), with bin size set to 5. Show the steps for your calculation.

Q4. You are given the following data set, where

- attributes “name”, “eye-color” and “hair-color” are categorical;
- attributes “height” and “weight” are numeric values;
- the class label “region” has three possible values: Asia, Europe and America.

| <i>name</i> | <i>height(cm)</i> | <i>weight(kg)</i> | <i>eye-color</i> | <i>hair-color</i> | <i>region</i> |
|---------------|-------------------|-------------------|------------------|-------------------|---------------|
| <i>Abel</i> | 180 | 73 | hazel | black | Europe |
| <i>Bob</i> | 183 | 75 | brown | brown | Europe |
| <i>Carl</i> | 176 | 70 | hazel | blond | Europe |
| <i>Dale</i> | 168 | 60 | brown | black | Asia |
| <i>Eric</i> | 174 | 65 | brown | brown | Asia |
| <i>Felix</i> | 170 | 59 | brown | blond | Asia |
| <i>George</i> | 180 | 78 | blue | black | America |
| <i>Howard</i> | 179 | 80 | hazel | brown | America |
| <i>Igor</i> | 175 | 75 | blue | blond | America |

- 1) Use Python (or Matlab/R) to draw boxplots for the numeric attributes “height” and “weight”. Attach your result.
- 2) We expect “height” and “weight” to be strongly related. Show some evidence for this.
- 3) Since “height” and “weight” are strongly related, we will use PCA to reduce these to one-dimensional attribute. Show:
 - a) the covariance matrix of the height and weight (unbiased estimation);
 - b) the required eigenvector;
 - c) the reduced 1-dimensional attribute value for Abel, Dale and Igor. (Note: Remember to center each attribute of the data set before the calculation/transformation).
- 4) Check whether the attributes “eye-color” and “hair-color” are independent of each other, by performing the χ^2 -test at a significance level of 0.1 (you can refer to the following Chi-Square distribution table or use Chi-Square Distribution Calculator to assess the probability associated with test statistics). Show your steps clearly.

| Degrees of Freedom | Chi-Square (χ^2) Distribution Area to the Right of Critical Value | | | | | | | | | |
|--------------------|---|--------|--------|--------|--------|---------|---------|---------|---------|---------|
| | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
| 1 | — | — | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.071 | 12.833 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| 11 | 2.603 | 3.053 | 3.816 | 4.575 | 5.578 | 17.275 | 19.675 | 21.920 | 24.725 | 26.757 |
| 12 | 3.074 | 3.571 | 4.404 | 5.226 | 6.304 | 18.549 | 21.026 | 23.337 | 26.217 | 28.299 |
| 13 | 3.565 | 4.107 | 5.009 | 5.892 | 7.042 | 19.812 | 22.362 | 24.736 | 27.688 | 29.819 |
| 14 | 4.075 | 4.660 | 5.629 | 6.571 | 7.790 | 21.064 | 23.685 | 26.119 | 29.141 | 31.319 |
| 15 | 4.601 | 5.229 | 6.262 | 7.261 | 8.547 | 22.307 | 24.996 | 27.488 | 30.578 | 32.801 |
| 16 | 5.142 | 5.812 | 6.908 | 7.962 | 9.312 | 23.542 | 26.296 | 28.845 | 32.000 | 34.267 |
| 17 | 5.697 | 6.408 | 7.564 | 8.672 | 10.085 | 24.769 | 27.587 | 30.191 | 33.409 | 35.718 |
| 18 | 6.265 | 7.015 | 8.231 | 9.390 | 10.865 | 25.989 | 28.869 | 31.526 | 34.805 | 37.156 |
| 19 | 6.844 | 7.633 | 8.907 | 10.117 | 11.651 | 27.204 | 30.144 | 32.852 | 36.191 | 38.582 |
| 20 | 7.434 | 8.260 | 9.591 | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 | 39.997 |
| 21 | 8.034 | 8.897 | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 | 41.401 |
| 22 | 8.643 | 9.542 | 10.982 | 12.338 | 14.042 | 30.813 | 33.924 | 36.781 | 40.289 | 42.796 |
| 23 | 9.260 | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 | 44.181 |
| 24 | 9.886 | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 | 45.559 |
| 25 | 10.520 | 11.524 | 13.120 | 14.611 | 16.473 | 34.382 | 37.652 | 40.646 | 44.314 | 46.928 |
| 26 | 11.160 | 12.198 | 13.844 | 15.379 | 17.292 | 35.563 | 38.885 | 41.923 | 45.642 | 48.290 |
| 27 | 11.808 | 12.879 | 14.573 | 16.151 | 18.114 | 36.741 | 40.113 | 43.194 | 46.963 | 49.645 |
| 28 | 12.461 | 13.565 | 15.308 | 16.928 | 18.939 | 37.916 | 41.337 | 44.461 | 48.278 | 50.993 |
| 29 | 13.121 | 14.257 | 16.047 | 17.708 | 19.768 | 39.087 | 42.557 | 45.722 | 49.588 | 52.336 |
| 30 | 13.787 | 14.954 | 16.791 | 18.493 | 20.599 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |
| 40 | 20.707 | 22.164 | 24.433 | 26.509 | 29.051 | 51.805 | 55.758 | 59.342 | 63.691 | 66.766 |
| 50 | 27.991 | 29.707 | 32.357 | 34.764 | 37.689 | 63.167 | 67.505 | 71.420 | 76.154 | 79.490 |
| 60 | 35.534 | 37.485 | 40.482 | 43.188 | 46.459 | 74.397 | 79.082 | 83.298 | 88.379 | 91.952 |
| 70 | 43.275 | 45.442 | 48.758 | 51.739 | 55.329 | 85.527 | 90.531 | 95.023 | 100.425 | 104.215 |
| 80 | 51.172 | 53.540 | 57.153 | 60.391 | 64.278 | 96.578 | 101.879 | 106.629 | 112.329 | 116.321 |
| 90 | 59.196 | 61.754 | 65.647 | 69.126 | 73.291 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| 100 | 67.328 | 70.065 | 74.222 | 77.929 | 82.358 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |